ChinaScope 數庫

Chinese Language NLP

White Paper on the methodology and processes of ChinaScope's NLP

May 15, 2019

First Edition

Legal Disclaimer

This white paper is provided by ChinaScope Limited and its affiliates (hereinafter referred to as "ChinaScope"). The company or individual who receives this white paper (hereinafter referred to as "you") should read this disclaimer carefully when receiving this white paper. If you do not agree to this legal disclaimer, please do not continue reading it and promptly delete it. If you continue to read, distribute or use this white paper, you then agree to the conditions outlined herein.

All rights to this white paper (including but not limited to its textual expressions, trademarks, logos, images, layouts, data, and technologies covered in this white paper) are owned by ChinaScope, protected by the laws and regulations of the People's Republic of China, the intellectual property laws of jurisdiction you are domiciled in and other relevant international conventions. The issuance of this white paper by ChinaScope to you is not to be construed as granting you exclusive rights, title, license or interest in any part of this white paper. You may not modify any of the content in this white paper, and you may not use this white paper for any unlawful application.

The information contained in this white paper is designed to assist you in understanding and evaluating the natural language processing methodologies and processes employed by ChinaScope. ChinaScope does not make any representations or warranties regarding the accuracy or completeness of this white paper or the information contained therein, or regarding any omissions or any other written or oral communication that you may receive during your evaluation of this white paper. In the event specific representations and warranties (if any) are made to you in the final written agreement relating to any transaction with ChinaScope, they are subject to the terms set forth in those agreements separate from this white paper.

Under no circumstance shall this white paper constitute or form a part of an offer to sell or to purchase any product. This white paper and any part of it does not form the basis of any contract or commitment. The dissemination of this white paper in or from a specific jurisdiction may be restricted or prohibited by law of that jurisdiction, it is your responsibility to promptly understand and comply with such restrictions and prohibitions.

In the event that you make a commercial decision based on the contents of this white paper, such a decision constitutes an independent conduct for which you are solely responsible, and ChinaScope is held harmless from any consequences that shall arise from such a decision.

Any product or technical performance and features contained in this white paper may be subject to change due to design modifications or technological innovations, and ChinaScope reserves the right to make changes or modifications to any of the products or technologies discussed in this white paper.

Table of Contents

1. ChinaScope NLP Overview

Since our founding in 2009, ChinaScope has committed itself to the odyssey of capturing knowledge in the financial services sector. In pursuit of this endeavor, we have adopted a philosophy of "data connectedness is king". Over the past 10 years, we have constructed a robust data schema and an ensemble of machine-assisted protocols to extract, synthesize and normalize information. A significant and inalienable part of the information construct that feeds into the decision-making paradigm in Finance is descriptive information expressed in natural language. This paper looks at the scope and methodologies of ChinaScope's Natural Language Processing (NLP) system in deciphering unstructured text and how that fits into knowledge construction in the financial services vertical.

ChinaScope, as the name suggests, takes China as the epicenter of our information universe and spirals outwards into the rest of the world. As of the authoring of this paper, ChinaScope's data coverage is still largely focused on Mainland China, with the exception of certain core data sets such as supply chain, which have emanated out to cover other markets such as Hong Kong and the United States. When it comes to NLP, our expertise thus far resides in Mandarin Chinese as used in the People's Republic of China. Although our algorithms can be applied to Chinese used in Hong Kong and Taiwan, the margin of error of understanding is larger due to the difference in vernacular in these two regions.

The focus of our NLP situates in the extraction of core information from written text and structures them in a way that synchronizes with the ChinaScope data schema. Although the origins of content ChinaScope processes can vary widely from publicly available sources (e.g. news, exchange filings, social media, etc.) to client content (e.g. credit reports, meeting reports, contracts), this paper will focus exclusively on the digestion of news which form the foundation of our standardized product suite.

This paper serves as a guide to the current and prospective consumers of ChinaScope iNews and SmarTag data feed products by giving a look under the hood of the engine and help them navigate the approaches we have taken in constructing these NLP-based product lines.

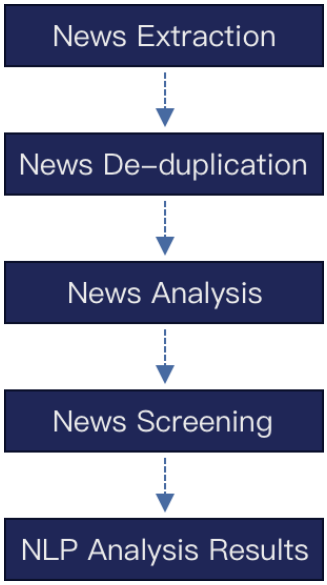
Contact

Tom Liu
CEO

Email: tom.liu@chinascope.com
LinkedIn: **https://www.linkedin.com/in/tom-liu-31337114/**
WeChat: Tommah2000

2. News Source Management

Journalistic news is the primary source of unstructured textural information that financial professionals access in their daily work. In China today, almost all news are consumed entirely via the internet, through web portals, apps, and social media. Given the pervasive and yet scattered nature of news on the internet, it is important to establish a system that methodically tracks and robustly evolves with the myriad of news sources, so there can be sufficient coverage for the purpose of business and financial analytics.



ChinaScope manages and controls in a string of processes that focus on news extraction, analysis, and screening. The filtering process involves a de-duplication step that removes identical or close to identical news from platform redistributions, as well as a noise removal step that removes extraneous information like ads.

2.1. News Extraction

ChinaScope captures news through our proprietary web crawler system, which is designed to meet the flexibility and robustness requirements that come with a constantly expanding universe of content, the mercurial state of webpages as they change unpredictably, and the noise that accompany genuine content. As of the authoring of this paper, ChinaScope covers 250+ news media outlets and 1,000+ website pages.

2.2. Information Processing

2.2.1. Data History

We began our research into NLP in late 2013. We released our first NLP engine in 2015, and we first

systematically captured news via the internet through our current web crawler system was in November 2016. Initially, the content coverage was poor, but over the years, we have steadily built up our news coverage. For the purpose of quantitative back testing, we have backfilled news data to 2008 with a third-party news provider. Currently, on average we process circa 15,000 articles daily.

2.2.2. Validity Screening

ChinaScope employs a news source filtering mechanism for vetting valid news sources. Due to the wide range of news sources, diversified layout of content, and the varied reliability of the raw captured news content. ChinaScope keeps an internal standard of filtering logic for effective news content based on attributes such as content classification, site stability and extraneous content noise.

1. Content Classification: ChinaScope classifies content into company, economic, industry, etc. We exclude content that have little relevance to financial and economic research, such as entertainment and literary articles.

2. Site Stability: Certain sites are poorly managed and often become "ghost" sites.

3. Extraneous Content Noise: This pertains to information that has little to do with the core content but is nevertheless interlaced in with the website content. Such noise is mostly advertisements. Some advertisements are hard commercials, others are soft placements.

ChinaScope continuously optimizes the news source screening mechanism to ensure the reliability of news delivery and their analytical results.

2.2.3. Scope of Content Coverage

ChinaScope's news sources are selected from mainstream financial media, covering popular financial news websites, and their sub-sites. News sites often have different foci, some are macro focused, others are sector focused. It is important to track the predilection of different news sites, as this understanding yields valuable information when it comes to content filtering and algorithm training.

Table 1. Content distribution based on news source types

News Site Type	Number of Sites	Percent Makeup	Number of Articles (1)	Percent Makeup
Macro	79	7.00%	46,039	5.79%
Sector	438	38.83%	374,194	47.06%
Company	115	10.20%	91,422	11.50%
Comprehensive	215	19.06%	45,796	5.76%
Market Quotations	176	15.60%	163,998	20.62%
Other Financial Markets	105	9.31%	73,726	9.27%
Total	1,128	100.00%	795,175	100.00%

(1) Based on news aggregated from 1 October 2018 to 31 March 2019.

Taking a view from a sector perspective, we see that close to 30% of the news sites we cover pertain to information a single sector.

Table 2. News site distribution based on sector

SAM Level 1 (1)	Number of Sites (2)	Percent Makeup
Energy	20	1.77%
Materials	8	0.71%
Industrials	83	7.36%
Consumer Discretionary	13	1.15%
Consumer Staples	46	4.08%
Health Care	9	0.80%
Financials	49	4.34%
Information Technology	71	6.29%
Communication Services	1	0.09%
Utilities	14	1.24%
Real Estate	17	1.51%
Trade	1	0.09%
Comprehensive (3)	796	70.57%
Total	1,128	100.00%

(1) SAM Level 1 is the equivalent of GICS Level 1 classification with Trade added to accommodate the commercial reality in China that many companies engage in trading of goods that span across a wide selection of sectors.

(2) New sites covered as of 31 March 2019.

(3) Comprehensive represents sites covering more than one sector as measured by SAM Level 1 classification.

News websites undergo rigorous screening before being included in the information source to ensure that the quality of news capture is guaranteed at the source. Aside from mainstream media portals, we select the more boutique news sites based on indicators such as website rankings, website traffic, and website influence. For sector specific websites, we also look to references by industry associations and key industrial publications.

2.2.4. Timeliness

As a source of high-frequency information, news has keen requirements on timeliness of delivery. Based on the aggregation of news captured by ChinaScope for the month of March in 2019, more than 70% of pairs of articles have less than 30 minutes of gap between them. Typically speaking, the peak times for news publications are before market opens and after market closes. The highest frequency of news delivery ChinaScope can provide is every five minutes.

Table 3. News publication frequency distribution

Frequency (1)	Article Pairs (2)	Percent Makeup
[0,2]	21,509	15.66%
[2,5]	38,754	28.21%
[5,10]	14,068	10.24%
[10,30]	24,159	17.59%
[30,60]	11,722	8.53%
[60+]	27,146	19.76%
Total	137,358	100.00%

(1) Frequency represents the time differential (in minutes) between the publication of two adjacent articles from the same news site.

(2) Based on news aggregated from 1 March 2019 to 31 March 2019.

2.2.5. Noise Reduction

In addition to the controls placed on news source screening, ChinaScope also performs quality control on the content of the information source. We currently screen and filter for content deduplication, news on stock performance (often times robot-generated), and advertisement related content.

(1) De-duplication of News Content

In life, the amount of market moving events on daily basis is quite finite. As such, news sites generate much of their content by regurgitating the same events or retransmitting content from other sources with minimal alterations. Looking at the aggregate news articles collected by ChinaScope for the six months ending 31 March 2019, we see that content repetition rate is 55.5%. ChinaScope uses Simhash algorithm to mark and filter out duplicate news.

(2) News on Stock Performance

There is a certain type of news content that can severely skew meta-tagging of articles. These are daily news on market performance of stocks, which typically take the form of a list of stocks, their percentage growth or decline for a said period of time and some superficial commentary on the performance. These are usually machine generated and devoid of meaningful substance. ChinaScope identifies such news articles by algorithm and tags them separately. Based on aggregated news for the

six months ending 31 March 2019, such type of news represents approximately 14% of all articles.

(3) Advertisement Related Content

One of the most distracting elements of crawling news from websites is that they often have advertising interlaced into the content of the articles. This typically affects 10% of the articles ChinaScope attains, and it is systematically removed.

2.2.6. Monitoring for Source Stability

An indelible reality associated with content sourced from the internet is that there is significant uncertainty and unpredictability associated with the whims of the source sites. Whether it's a structural change of a site, a new design of a specific webpage, a site failure, or just simply relegating a webpage into dereliction, it all affects the quality of information extraction. ChinaScope monitors the quantity and quality of news from source sites, such as tracking the number of site crawls, systematically checking the quality of news crawling, and repair failed sites immediately upon discovery.

3. SmarTag System

SmarTag system is ChinaScope's flagship NLP tagging system. It identifies key information within documents and lifts it from the text and transforms it into standardized labels that map to ChinaScope's data topology. This not only allows unstructured text to be machine readable, but it also seamlessly integrates with numerical data in the ChinaScope database.

3.1. Name Entity Recognition

3.1.1. ChinaScope's Definition of An Entity

An entity, under ChinaScope's definition, refers to an individual who is capable of autonomously generating activities. As such, entities mainly refer to companies and people. ChinaScope covers a wide range of companies, including China A-shares, Hong Kong stocks and listed companies on the NEEQ market, listed companies on NYSE and NASDAQ, China-debt issuers and unlisted Chinese companies. For entities that are people, ChinaScope mainly covers senior executives, shareholders and directors of domestic listed companies.

3.1.2. Company and People Tags

Algorithm on Extraction of Company Tags

Background

Identifying which companies are mentioned in the news is important foundational work for analyzing financial news. By extracting companies in the news, the news can be associated with a specific company or multiple companies, and then the other content mentioned in the news, such as sector, products, events, public opinion, etc. are linked to relevant companies to provide data support for further analysis.

Input / Output

Algorithm Input：

Headline and content of news articles.

Algorithm Output：

```
{
        "stat": 0,
        "version": version,
```

```
"message":[
        {
            "idx": [["abs_cut_id": int, "abs_sen_id": int], ],    #keyword position in the news
            "abbr": company abbreviate ( stock name ) ",
            "comcode": company unique code,
            "chiname": company full name,
            'sim': relevancy between the company and the news,
            'kw':[company keyword, ]
            "mkt": stock market (optional)
            "code": ticker (optional)
        }
    ]
}
```

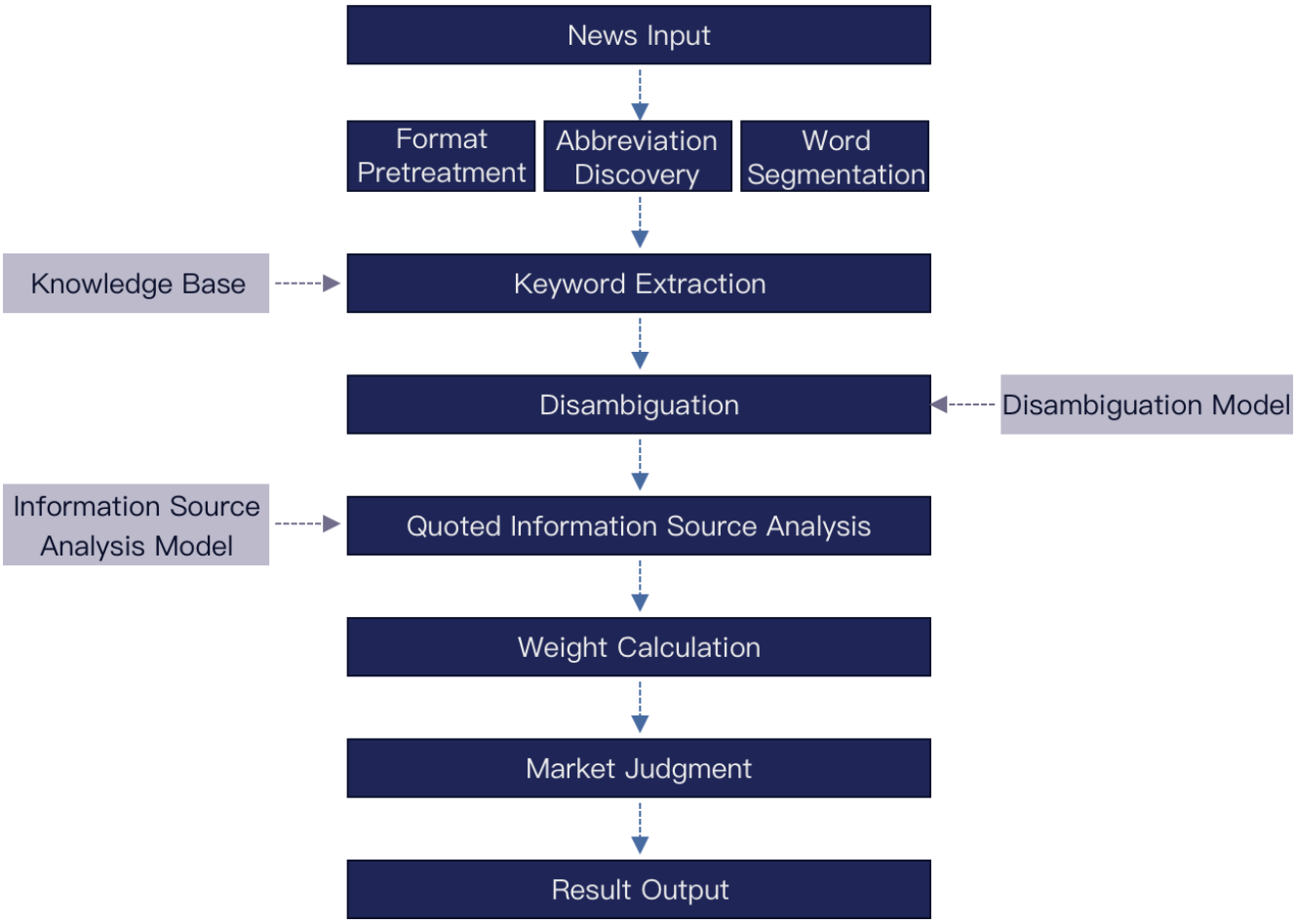
Technical Principle

Technological Background：

There are usually two ways to extract a company from the news: the first is to build a company knowledge base, collect the company's name and various short names, and then look up in the knowledge base the company mentioned in the news. The advantage of this method is that it is simple to develop, fast in calculation and easy to maintain. The disadvantage is that it is difficult to deal with semantic ambiguity, and when the number of companies is large, it needs to occupy significant storage resources.

The second way is to think of the extraction company as an entity recognition task, using a machine learning model to identify the company from the sentence. The current mainstream approach is to use the neural network model of BiLSTM+CRF, or use various amended versions of this model, and use Lattice LSTM in the LSTM part to introduce word and word order information to improve the model result. The advantage of this method is that it can identify semantic information to some extent, partially solve ambiguity problems, and it does not rely on the thesaurus. This approach has three shortcomings. One is that the model is slower, especially when the news article is long. Second, it requires a lot of training corpora. Third, it is difficult to maintain and difficult to optimize and correct specific error samples.

System Architecture：



Algorithm Architecture：

- Extraction algorithm based on knowledge base;
- ChinaScope's corporate knowledge base includes the company's basic information database, company synonym database, company ambiguous thesaurus, publishing company name library, and company registry library;
- The company's basic information database includes companies listed on the A-share, NEEQ, Hong Kong Stock Exchange, NYSE and NASDAQ, as well as non-listed companies directly associated with these listed companies. ChinaScope records the company's full name, abbreviated name, market and stock code (if any), company identification ID (CSFID) and other relevant identification information;
- The synonyms database contains the alternative names of companies in the company's basic information base, including Chinese synonyms and English synonyms;
- The ambiguous thesaurus database contains names of companies that are also nouns or verbs used in common vernacular or in specialized fields;
- Publishing company name library incorporates publishers that put out reports and articles. It is important to differentiate the issuer of a report to the subject company of a report. Such

publishing companies include ratings agencies, broker dealers, research houses, etc.;

- Company registry database includes all companies registered with the State Administration for Market Regulation of the People's Republic of China. This includes more than 60 million companies, of which approximately 40 million are operating.

The extraction algorithm first parses words from inputted news, then finds the company's related terms in the company knowledge base on word segmentation results, and then performs distribution analysis of the company's related words. Based on the distribution of related words in the news, a relevance weighting score of the company in the text is calculated. A higher weighting is given to companies that appear in the title.

Disambiguation:

For knowledge-based name entity extraction, ambiguity discrimination is crucial. There is a large number of name abbreviations of companies that share meaning in common vernacular. For instance, the term "Tesla" is both the name of Tesla Inc. and the family name of Nicola Tesla. This situation is more common in small to medium sized companies and in relatively young companies, as new age companies in China founded by millennials have a predilection for "cute" names. For NEEQ listed companies, the abbreviation of more than 500 companies have common meaning elsewhere.

As mentioned before, the ambiguity phenomenon can be mitigated to a certain extent by means of name entity identification. But entity recognition for all sentences can greatly affect the speed of operation. This is because the current mainstream entity recognition technology is based on a cyclic neural network model, usually BiLSTM. This kind of neural network model works on serial calculation. For an input sentence, it needs to be computed word by word, causing the calculation speed to be slow, and hence is not suitable for large-scale use under the condition that the computing resources are limited and speedy production is key.

Therefore, ChinaScope's extraction algorithm divides the disambiguation process into two steps. First, it determines whether there may be ambiguous words in the sentence, and then the disambiguation is performed on the sentences that may have ambiguous words.

Judging the existence of ambiguity relies on a lexicon base of ambiguous terms, which is maintained by the ChinaScope team on a continuous basis. Filtering based on this lexicon base can cover 95% of the situations as of the authoring of this paper.

The company's extraction algorithm uses entity recognition techniques to eliminate ambiguity. The model used is the neural network model of Lattice LSTM and CRF, where Lattice LSTM is an improved version of LSTM. Usually, LSTM is used to perform embedding on a single word for entity recognition. This avoids errors caused by word segmentation, but it cannot utilize existing vocabulary information. Lattice LSTM adds vocabulary information to the calculations of the LSTM unit, enabling the LSTM unit

to use vocabulary-level information to obtain better results when calculating the current state.

Abbreviation Discovery:

For listed companies, the abbreviated name is usually fixed in common vernacular, usually with the commonly adopted name of the listed securities as the abbreviation. However, the abbreviated names of private companies are often myriad fold. Except for a few well-known companies, the abbreviated names of most unlisted companies is produced at whim by the author of the article, so the same company often has different abbreviated names in different news articles. In order to better identify private companies, ChinaScope's extraction algorithm analyzes the news before word parsing, extracts potential company abbreviated names, and interconnect them with the appropriate companies in subsequent analysis.

Market Recognition:

Another challenge that comes with dealing with abbreviated names is that many companies share the same name abbreviations. This is rather common when it comes to companies listed across different exchanges. The more markets we cover, the more of such instances occur. In order to discriminate between the real entities behind each name abbreviation that occurs in news articles, we apply market recognition to the algorithm. By identifying which exchange the abbreviated name belongs to throughout the article, we are able to ascertain the true identity of the company. Of course, there are also cases where there is no market-related information in the context of the article. In this situation, we would try to determine the company's identity by triangulating other information in the article. If it is really impossible to pinpoint the specific company's identity, all possible results corresponding to the abbreviation would be returned.

Quoted Information Source Analysis:

There is a situation in analyzing news in which a company is named in the news but not as the subject of a particular topic, but as the third-party opinion provider or publisher of the subject. These companies are often times themselves listed entities, and are frequently the subjects of news topics. An example of this would be broker dealers, which releases research reports that are frequently quoted in the press. In the situation where a company appears in an article as a quoted information source rather than the subject of the topic, we need to apply a different treatment.

A module that analyzes quoted information sources is added to the extraction algorithm, which effectively lowers the relevance score of these type of companies in the context of the articles.

The Quoted Information Source Analysis module converts the problem of identifying the company in this special case into a classification task for processing, and it uses a convolutional neural network model for classification. Based on a TextCNN network, the Quoted Information Source Analysis module

optimizes the model structure, adds an expansion convolution technique, and uses the Dropout layer to perform over-fitting control.

Performance Optimization:

When company entity extraction only pertains to listed companies and their associated companies, memory usage is not a problem, because the number of companies is only a few hundred thousand. When company entity extraction applies to private companies, the number jumps dramatically to more than tens of millions, and this places constraints on memory resource consumption.

To address the aforementioned issue, ChinaScope takes the following measures:

1. Data Structure Optimization

The company entity extraction algorithm involves a variety of data types, data volume is immense and there is a multitude of relationships between the data. The algorithm is developed using Python. Due to the inherent characteristics of the Python language, if you use its default dict, list and other data structures for storage, it will consume a lot of RAM, far beyond the memory capacity of a standard server. ChinaScope resolves this issue by optimizing the data structure inside the company extraction algorithm.

Numericize Textual Strings：

Company names can contain long textual strings. When it comes to processing tens of millions of such texts, it can quickly eat into memory storage. By transforming textural strings into integer IDs, we are able to greatly reduce memory uptake.

Storage and Query：

Python's dict structure is a data structure that sacrifices memory space for fast lookup. Therefore, when it comes to private company data, we do not use dict storage, but use sorting + sequential storage + index positioning + binary search for storage and query. Sequential storage can greatly reduce memory footprint. Sorting the data enables the data to be queried by means of binary search. The time complexity of the query is $O(\log_2 N)$, and N is the total amount of data points (e.g. N=66,804,810, based on private company data as of March 2019). By indexing the data extraction features, you can quickly locate the data before the binary search, thus reducing the number of binary searches. The query time complexity after joining the index is $O(\log_2 n)$, where n is the number of data points under the same key in the index, (e.g. the average of n is 103.5, based on private company data as of March 2019). It can be seen that adding indexes has a marked improvement on query performance.

2. Cache Relating to Private Companies

In optimizing the data structure, the query speed of private company data is to a certain extend compromised. In order to mitigate this impact, a cache mechanism is set up for the data of private company searches. This can greatly reduce the speed impact caused by data structure optimization.

3. Word Segmentation Optimization

The company entity extraction algorithm uses a segmentation protocol based on directed acyclic graphs using indexed and separated dictionaries. This is due to the need to ensure that company names in the knowledge base are segmented and that the lexicon base of 60 million can be used. The algorithm ensures that the words in the knowledge base are segmented, and memory uptake is minimized and processing speed is optimized.

Mapping Securities Codes to Unique Company IDs:

ChinaScope identifies companies as operating organizations in texts, and then map securities codes to these organizations via a separate mapping layer. This replaced the original method of identifying securities directly in tests, which works well if the focus of coverage is within a single securities market. However, when coverage expands to companies are listed on multiple exchanges with varying securities instruments, a bifurcated approach eliminates potential for confusion. Also, it provides flexibility in back testing where analysts may or may not want to identify companies based on their historical listing status.

Algorithm on Extraction of People Tags

Background

When analyzing name entities involved in news, aside from companies, people are also important. As of the authoring of this paper, ChinaScope tracks people who are related to companies in such capacities as directors, supervisors, senior executives, and shareholders. We also track some celebrity business persons.

Input / Output

Algorithm Input:

Headline and content of news articles.

Algorithm Output:

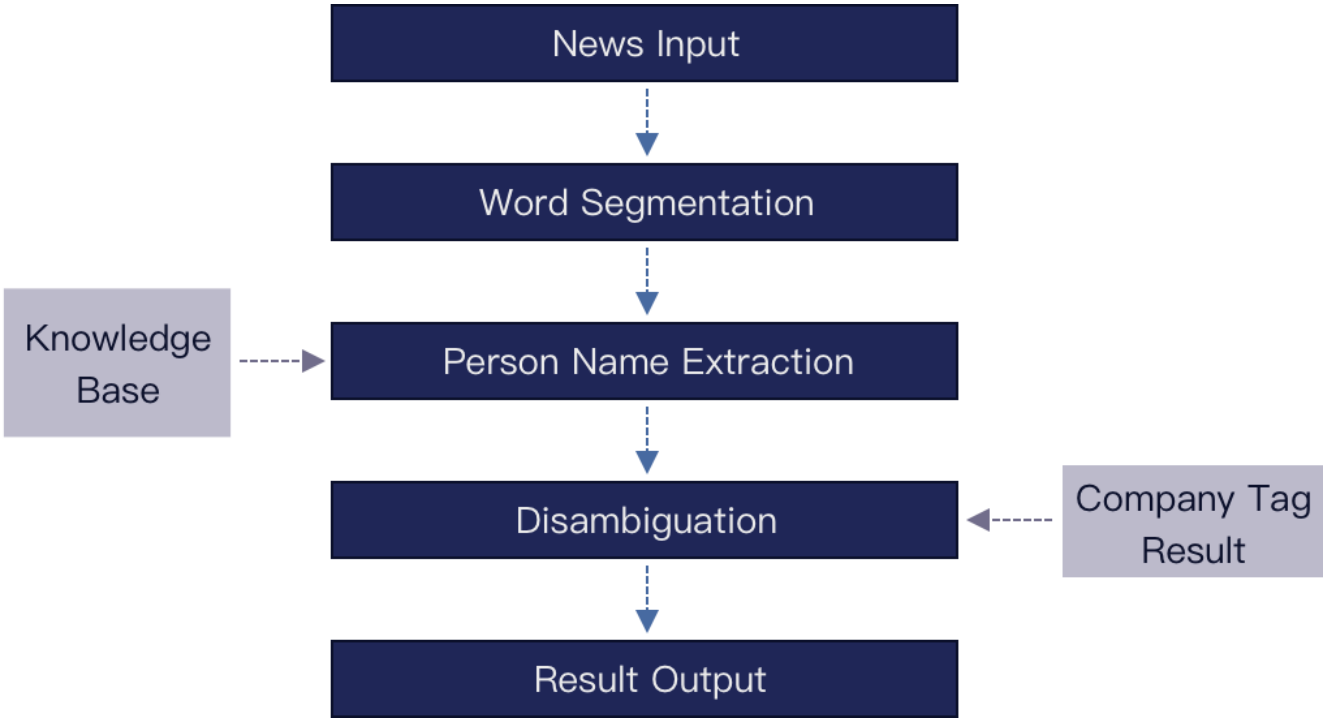
```
        {
    "stat": 0,
    "version": version,
    "message": [
        {
            'name': person name,
            'pcode': person unique code,
            'speak': 0,          #whether person talk in the news, 1:Yes,0:No
            'comcode': company (person service) unique code,
            'secu': company abbreviate,
                'position': person position,
            'code': ticker,
            'speak_con': person speech content,
            'idx': [{'abs_cut_id': m, 'abs_sen_id':n}, ],    #keyword position In the news
            'kw':[ person keyword, ]
        }
    ]
}
```

Technical Principle

Technical Background:

Similar to extracting company entities, extracting personnel entities can also be achieved via a knowledge base approach or by name entity recognition. Since the people to be extracted belong to a known universe of individuals, hence a limited data set, ChinaScope adopts the knowledge base approach.

System Architecture:



Algorithm Architecture:

- Algorithm based on knowledge base;
- ChinaScope's personnel knowledge base comes from our professionals database and shareholders database, including the actual controlling shareholder database;
- Actual controlling shareholder database, as of the authoring of this paper, only pertains to A-share companies;
- All personnel information includes unique ID, name, and relationship to associated company.

The extraction algorithm first parses characters from the input news, and then finds the person's name in the personnel information database based on the character segmentation results, and then outputs the result after disambiguation.

Disambiguation:

There are two kinds of ambiguities that may exist when it comes to identifying persons in Chinese. The first kind relates to the fact that many Chinese people have exactly the same first name and last name. The second kind relates to names that are also things or concepts in life. A similar example of this in English would be the name "Destiny". However, because there is no capitalization in Chinese characters to denote proper nouns, disambiguating in this situation presents more of a challenge.

To eliminate the aforementioned ambiguities, the people entity extraction algorithm uses company

tags as anchor information. For people who are not celebrities, there is typically mention of his or her connection to a particular company in the context of the article, which allows us to anchor the identity of the individual of interest.

3.1.3. News and Exchange Filing Coverage of Companies

News, given its inherent characteristic of journalistic focus, tends to cover companies that have "news worthy" events of companies that the public cares about. They need to be consumed in combination with stock exchange filings of companies. Taking A-share companies as an example, the distribution of news coverage of companies is different from that of exchange filings. In the table below, which tabulates news and exchange filings for the six months ending 31 March 2019, we see that more information can be gathered on large cap companies from news compared to their stock exchange filings. However, as companies move down the market cap scale, less information can be garnered from news in comparison to their stock exchange filings. ChinaScope provides more than 960 event categorizations for stock exchange filings to complement our news analytics coverage.

Table 4. Distribution of A-share company tags in news and stock exchange filings

Company Market Cap	Market Cap Range (RMB) [3]	No. of Companies	No. of Companies with News	Percent Coverage	No. of News Articles [1]	% Dist. of news	Exch. Filings [2]	% Dist. of Exch. Filings
Large	Above 500 bn	10	10	100.00%	8,994	5.29%	518	0.24%
Large	100 – 500 bn	77	75	97.40%	30,430	17.89%	5,126	2.38%
Mid	30 – 100 bn	270	269	99.63%	27,887	16.39%	19,532	9.05%
Mid	10 – 30 bn	737	736	99.86%	39,848	23.42%	48,274	22.37%
Mid	3 – 10 bn	1,846	1,844	99.89%	50,386	29.62%	109,589	50.79%
Small	Less than 3 bn	661	650	98.34%	12,578	7.39%	32,727	15.17%
Total		3,601	3,584	99.53%	170,123	100.00%	215,766	100.00%

(1) News aggregated from 31 October 2018 to 31 March 2019.

(2) Stock exchange filings aggregated from 31 October 2018 to 31 March 2019.

(3) Market cap calculated based on free float shares as of 4 April 2019.

For many quantitative research analysts, it is important to gauge what the company coverage situation is like in comparison to liquid tradable stocks that are constituents of indexes. Furthermore, now that northbound trading of A-share stocks via the HKEX-Shanghai-Shenzhen Stock Connect program has grown tremendously in prominence, it is also important to look at news coverage of tradable stocks via this program.

Table 5. News coverage of stocks in key indexes and HKEX-Shanghai-Shenzhen Stock Connect program

Markets	Indexes	Constituents (1)	No. of Constituents Mentioned in News (2)	% Proportion
A-Share	MSCI China A Onshore	789	783	99.24%
	CSI 800	800	798	99.75%
	Shanghai & Shenzhen Connect (stocks eligible for both buy and sell)	1,318	1,306	99.09%
HK-Share	Hang Seng Index	50	45	90.00%
	Hang Seng Composite Index	485	448	92.37%
US-Share	S&P 500	505	336	66.53%
	Dow Jones Industrial Average	30	29	96.67%
	Nasdaq 100	103	80	77.67%

(1) Index Constituents as of 30 April 2019.

(2) News aggregated from 1 April 2019 to 30 April 2019.

3.2. SAM and Supply Chain Tags

3.2.1. SAM Classification System

Under the umbrella of a globalized economy, ChinaScope recognizes that a top down industry classification system like GICS severely falls short of the requirement of tracking goods and services being rendered by companies. As such, ChinaScope built a taxonomy of product and services that map to GICS based on the reporting of listed companies in China, Hong Kong and the US. The name SAM stands for Segment Analysis & Mapping, which refers to the origins of its development from the business segment reporting of ~20,000 listed companies. ChinaScope is constantly upkeeping this taxonomy as we expand our coverage to more exchanges and to the private sector.

3.2.2. SAM Data Structure

SAM's 4,000+ product nodes are built onto a data topology that extends from traditional industry classifications. The example below illustrates the Information Technology sector, where GICS industry classification schema drops down to four layers, and ChinaScope's SAM picks up from the 4th layer and further subcategorizes into twelve layers. Different sectors have varying degrees of depth.


01
Information Technology
02
Technology Hardware & Equipment
Software & Services
03
Electronic Equipment, Instruments & Components
Communication Equipment
.....
04
Electronic Manufacturing
Electronic Manufacturing Services
GICS Level 4
= SAM Level 1
05
Electronic Equipment & Instruments
Electronic Components
06
Electronic Cards
Instruments & Meters
.....
07
Applied Electronic Instruments & Equipment
Automation Instruments & Meters
08
Optical Equipment & Instruments
Non-defense Navigation Equipment
.....
09
Optoelectronic Equipment & Instruments
Optical Instruments
10
Optoelectronic Displays
11
Large-screen Splicing & Display Systems
Liquid Crystal Display Systems
12
Touch Screen Display

3.2.3. Sector and Product Tags

Background

The purpose of the sector and product labeling algorithms is to analyze which industries and products are involved in news information, and to give weights to industries and products in the news in a quantitative way. This links news to industry and products to provide data support for further analysis. Sector refers to SAM 4th layer classification, and product refers to anything below 4th layer.

Input / Output

Algorithm Input:

Headline and body text of news articles.

Algorithm Output:

The output includes the sector, products and SAM unique ID. A weighting score of between 0 and 1 is also produced to measuring the relevance level of each of these tags in the context of each article.

Product Tag Output:

```
{
 "message": [
      {
        "name": product name,
        "sim": relevancy between the product and the news,
        'kw': [product keyword, ] ,
        'samcode': product samcode
      },

 ]
 "stat": 0
 "version": 'v0.4.0.9'     #version
}
```

Sector Tag Output:

```
{
  "message": [
          {
                "name": sector name,
                "sim": relevancy between the sector and the news,
```

```
                    'kw': [ sector keyword, ] ,
                    'samcode': sector samcode
                }

    ]
        "stat": 0
    "version": 'v0.4.0.9'     #version
}
```

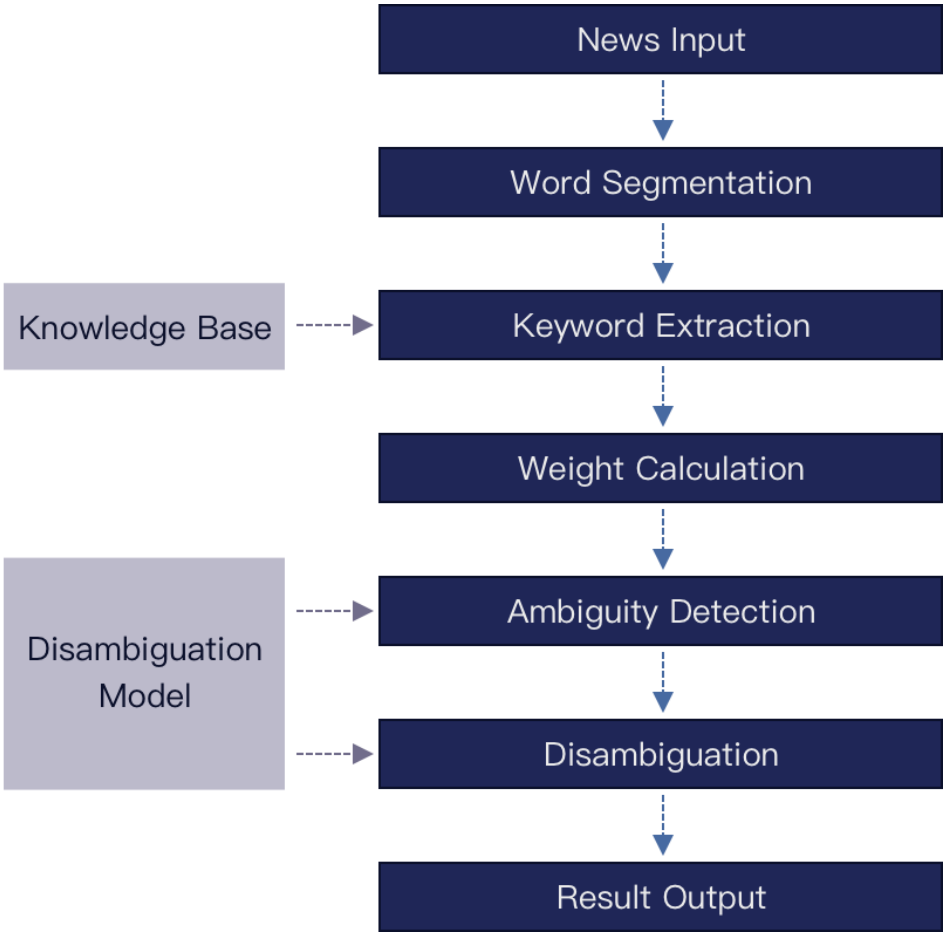
Technical Principle

Technical Background:

There are usually two methods to extract sector and product tags from news. The first method is to treat the extracted tags as a text classification task. Each tag acts as a category and trains the algorithm model to distinguish which category the news belongs to. The advantage of this method is that it has better generalization ability under normal circumstances. The disadvantage is that it is more dependent on labeled data, which can be problematic given that sector distribution in news is highly irregular. It may be that a piece of news talks about multiple sectors, or only one or two paragraphs relate to an industry, which will affect the effect of the classification model. The second method is to obtain sector and product information by extracting sector and product related terms from the news by establishing a sector and product knowledge base. The advantage of this method is that it does not require labeled data. The disadvantage is that the ambiguity caused by the polysemy of the word is not well handled.

ChinaScope integrates the above two methods, combining the advantages of both, based on the sector and product knowledge base, identifying potential sector and product related words from the news, and then identifying the possible ambiguities through the ambiguity discovery algorithm. The ambiguity discriminant model is used to add an additional layer of analysis and judgement to arrive at the final desired result.

System Architecture：



Algorithm Architecture:

- Parsing of text from news;
- Extract key words for different sectors;
- Calculate the distribution of key words;
- Perform first round of filtering based on the statistical result;
- Filter out potential keywords with ambiguity;
- Perform disambiguation on key words, eliminate words that do not reflect sector or product meaning;
- Calculate relevance weighting of sector and product tags;
- Outputs result.

Knowledge Base-Based Extraction Algorithm:

- The sector and product knowledge base of ChinaScope includes the sector and product basic information database and the sector and product ambiguity thesaurus;
- Sector and product basic information database includes more than 100 sector categories and over 4,000 product categories. The database contains information on names, related terms, and the unique ID of each data item;

- The ambiguous term thesaurus contains words that could mean other things aside from sectors and products;
- The extraction algorithm first classifies the input news, then finds related words in the sector and product knowledge base on the word segmentation results, and calculates the distribution of related words. According to the distribution of related words in the news, the relevance weighting of the sectors and products is calculated in the context of the news article. Considering that the title is the most concentrated and important part of the news, the weight of the sectors or products that appear in the title would be higher.

Disambiguation:

For knowledge base-based sector and product extraction, ambiguity discrimination is an important part. There are many industries and products whose names and related words have many meanings in actual use. For example, the Chinese word for "cattle" also refers to people who scalp tickets.

To eliminate the influence of polysemy on sector and product extraction, the algorithm divides the disambiguation process into two steps. First, it judges whether there may be ambiguity in the news, and then disambiguation is performed on those texts where ambiguity is determined to exist.

There are two ways to judge whether ambiguity exists. The first way is to build an ambiguous thesaurus that contains ambiguous words that are common in the industry and product related fields. This is an accumulating effort that takes time, as ambiguous words are added to the thesaurus as we run into them. The second way is to analyze the relationships of different sector and product tags. There is usually a logical pathway between different tags in the same news article. This logical pathway break down quite starkly when the wrong meaning is assigned to a potentially ambiguous term. Calculating the vector distance between the different tags can yield reliable signals of potential ambiguities.

The sector and product extraction algorithm uses a classification model to discriminate ambiguity. Since there are many categories to be distinguished, and the data distribution of each category is severely unbalanced, the gap between the categories with more data and those with few data can reach more than 100 times. Therefore, the classification model adopts an "ovr" (one versus rest) strategy to create a classifier for each sector to determine whether a piece of news belongs to or does not belong to that particular sector. The classification model uses the Naive-Bayes model.

Ambiguity Discovery Update:

As new content enters into our coverage universe and new products and industries are created and released into the world, ChinaScope periodically updates the ambiguity discovery module with more training data.

Knowledge Base Upkeep:

With the changing tides of products and services in the market, we continuously update the sector and product knowledge base. There is a heavy human involvement in this part of the exercise.

3.2.4. Sector and Product Distribution in News

The sector and product tags map to ChinaScope's SAM & Supply Chain data structure. This allows us to track the distribution of mention in news articles. The following table looks at what the sector and product node coverage is like in aggregated news articles for the six months ending 31 March 2019.

Table 6. Sector and product coverage in news articles

SAM Product Levels	No. of Tag Categories Mentioned in News (2)	SAM Total No. of Nodes	Tag Coverage Ratio in News
Product Level 1 (Sector Level 4) (1)	98	114	85.96%
Product Level 2	253	436	58.03%
Product Level 3	713	1,179	60.47%
Product Level 4	794	1,296	61.27%
Product Level 5	446	713	62.55%
Product Level 6	153	253	60.47%
Product Level 7	33	48	68.75%
Product Level 8	3	6	50.00%
Product Level 9	1	1	100.00%
Total	2,494	4,046	61.64%

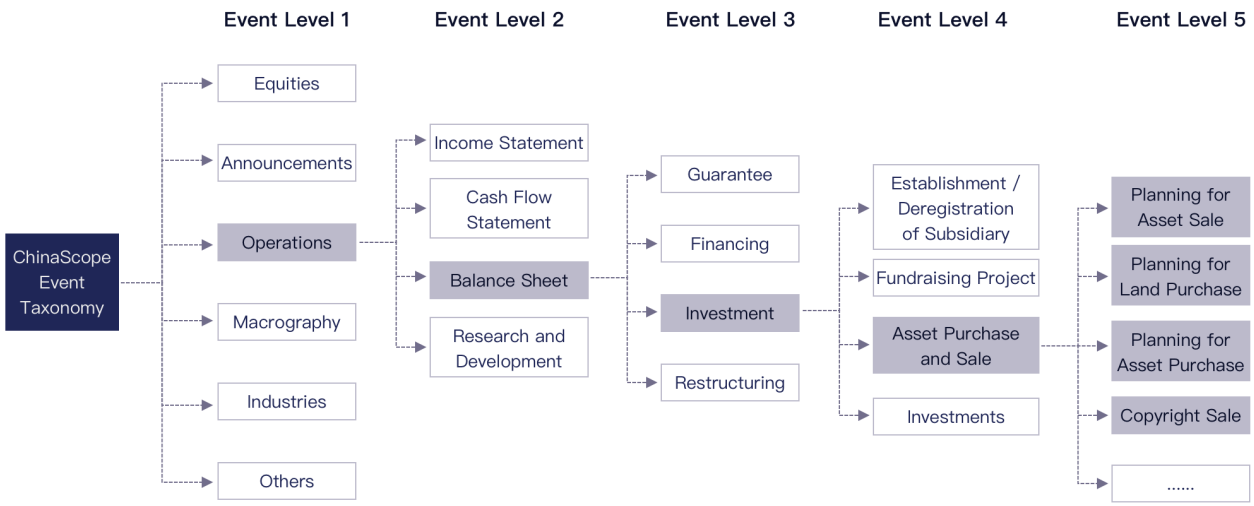
(1) SAM product level 1 is the same as sector level 4. Sector levels map to GICS, altered to better reflect the characteristics of Chinese industry.

(2) Aggregated news covering the six months period from 1 October 2018 to 31 March 2019.

3.3. Event Tags

3.3.1. ChinaScope Event Taxonomy

ChinaScope tracks a total of 1,800 event categories in news. These events are built into an event classification system of 6 broad categories of Equity Events, Company Announcement Events, Operating Events, Macro Economic Events, Sector Events, and Other. Each category breaks down into 5 subsequent layers. Company Announcement Events map to the event categories of exchange filings of A-share companies. For each of the most granular event categories, we have assigned separate predetermined positive, negative or neutral sentiment tags, which are used in the determination of sentiment analysis discussed later in this paper.



3.3.2. Event Tag Extraction

Background

The purpose of the event tag algorithm is to analyze which events are primarily described in news articles, thereby correlating news and events to provide data support for further analysis. The event types come from a pre-set event knowledge base.

Input / Output

Algorithm Input:

Headline of news articles.

Algorithm Output:

Output includes event name, event category, and the polarity tag and the associated unique codes.

Event Tag Output:

```
{
    'message':[
        {
            'name': event name,
            'cat': event category,
            'pos': event polarity tag,
            'code': event unique code
        },
```

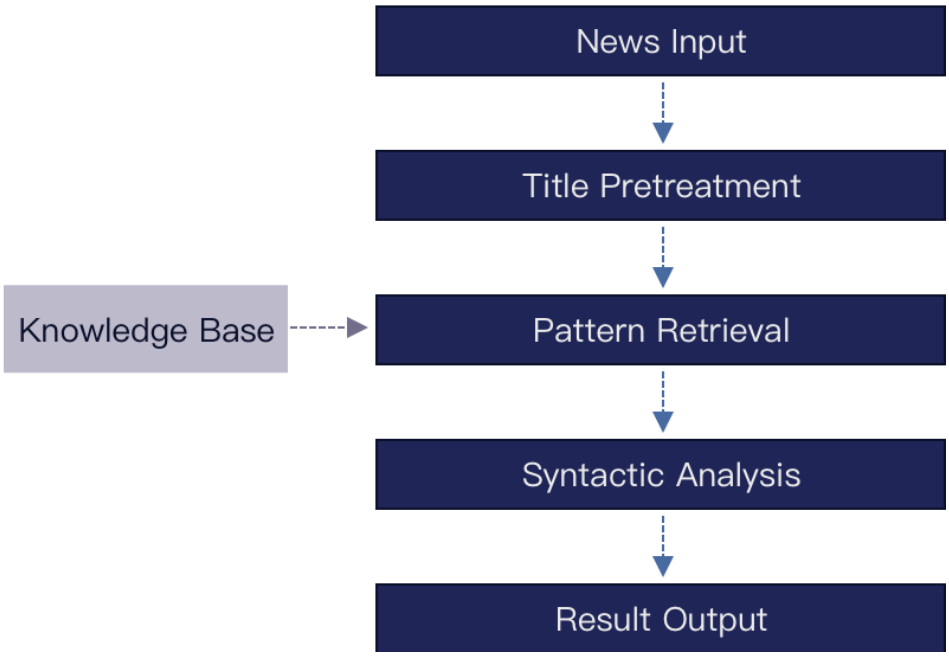
```
            ],
            'stat':0,
            "version": version
}
```

Technical Principle

Technical Background:

Event tag extraction involves more than 1,800 event categories, thus in order to achieve better accuracy, ChinaScope applies a rules-based system for event extraction, while layering a veneer of syntactic analysis on top to correct for errors.

System Architecture:



Algorithm Architecture:

- Extract title from news article;
- Parse and cleanse news title;
- Analyze title and retrieve the corresponding template;
- Syntactic analysis of the title;
- Outputs event tag.

Rule Template Retrieval:

Due to the large number of events we track, and each event corresponds to multiple rules, the efficiency of the system becomes particularly important. In order to improve the computing speed,

ChinaScope uses an information retrieval model to index event rules. Use the news headline to query for available rules in the index, and then apply the rules for extraction attempt.

Syntactical Analysis:

There are certain limitations to a rules-based system, so syntactic analysis is added as an additional correction mechanism. The event tag algorithm uses dependency syntax analysis to determine the grammatical relationship between the words in each sentence. This approach can help cover situations where rules fall short and drive the overall robustness of the extraction algorithm.

Maintenance of the Rule Knowledge Base:

The rule book needs to be updated and amended from time to time to ensure its robustness in precision and recall rates.

3.3.3. Event Tag Distribution

Looking at how event categories are typically distributed in the news, we can garner that operating events represent the highest proportion of close to 30%, based on six months of news ending 31 March 2019.

Table 7. Event tag distribution in news (Only aggregating level 1 and level 2 categories)

Level 1	No. of Tags [1]	% Proportion	Level 2	No. of Tags
Operations	73,717	29.96%	Income Statement	43,348
			Cash Flow Statement	253
			Research and Development	1,774
			Balance Sheet	28,342
Equities	21,660	8.80%	Equities of Companies	14,799
			Equities of Shareholders	6,861
Announcements	52,257	21.24%	IPO	4,953
			Risks	7,806
			Transactions	1,807
			Performance	24,191
			Systems	8,043
			Information of Companies	5,457
Macro	33,155	13.47%	Commodities and Currencies	3,426

			Policies	22,219
			Economy	7,510
Industries	537	0.22%	Industry Economy	537
Transaction Related	35,555	14.45%	Capital Market Performance	31,603
			Institution Rating and Survey	3,952
Others	29,174	11.86%	Negative Events	11,383
			International Politics	2,841
			National Security	289
			Actual Controller	1,849
			Calamity	4,707
			Major Events	5,838
			Others	236
			Funds	2,031
Total	246,055	100%	NA	

(1) Based on news aggregated for the six months period from 1 October 2018 to 31 March 2019.

3.4. Thematic Concept Tags

3.4.1. Concept Definition

Thematic concepts aren't as crisply defined as other tags. They spawn from hot topics in the news, and continue to linger in the collective psyche of the investment public. These themes could be based on certain events, geographies, people, products or sectors. An example of a thematic concept would be "One-Belt-One Road Initiative", another one would be "Artificial Intelligence".

3.4.2. The Emergence of a Concept

A new concept is usually caused by heated topics in the news. The news hotspots that investors pay more attention to may turn into investment-themed concepts. Therefore, to discover new concepts, we first need to extract hot topics in the news. ChinaScope discovers news hotspots by clustering news over time, usually taking the last 24 hours as a time window. The clustering algorithm gathers news into group topics which then become potential thematic concepts based on human selection criteria.

3.4.3. Concept Tags

Background

The purpose of the concept tag algorithm is to analyze which investment concepts are involved in the news and to give their relevance weighting in the context of news in a quantitative way. This links news and concepts to provide data support for further analysis.

Input / Output

Algorithm Input:

Headline and body of news articles.

Algorithm Output:

The output includes the concepts in the news and their unique identifiers, the relevance weighting (a number of between 0 and 1) of each concept in the context of the news, and the related words when each concept is mentioned in the news.

Concept Tag Output:

```
{
        'message':[
          {
                'name': concept name,
                'sim': relevancy between the news and the concept,
                'idx':[['abs_sen_id': int,'abs_cut_id': int ],], #indexes of th concept kewords
                'kw':[concept keyword, ],
                'code':'CP0001' #concept unique code
          },
      ],
      'stat':0
      "version": version
}
```

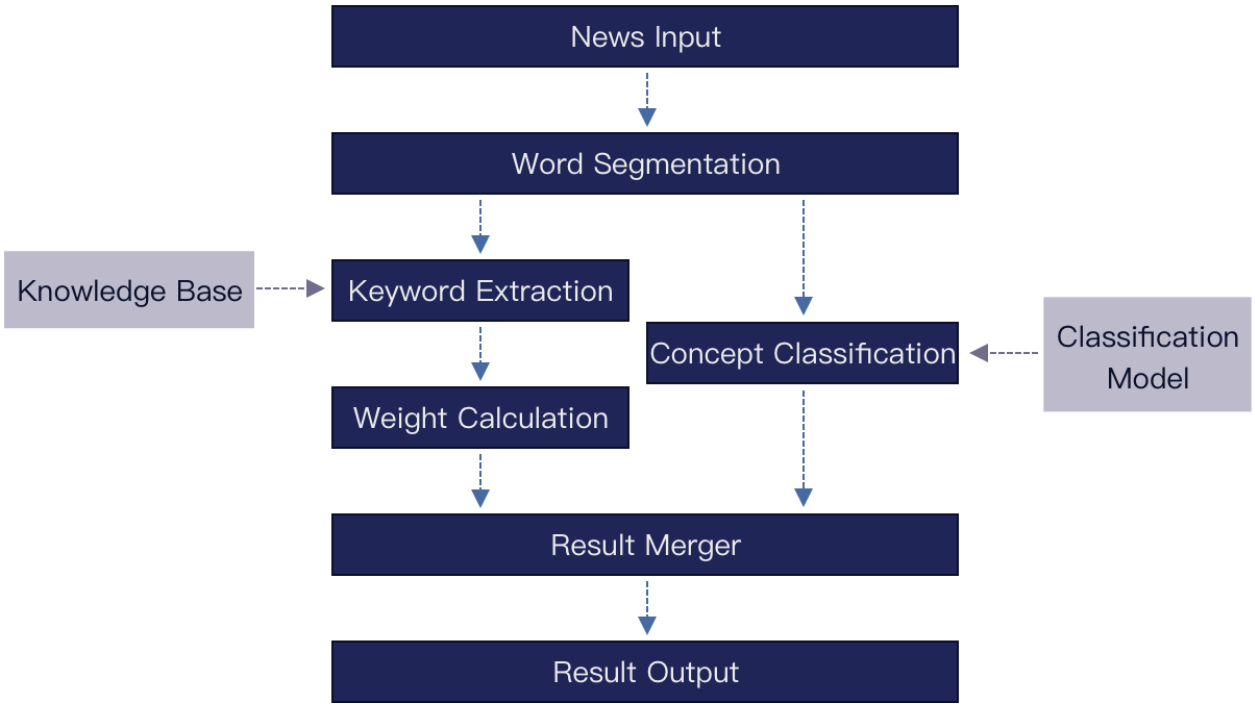
Technical Principle

Technical Background:

Like the sector and product extraction algorithm, the concept tags are extracted from news based on two methods: 1) knowledge base and 2) a classification model. However, unlike sectors and products, there are fewer ambiguities when it comes to concepts. ChinaScope uses a knowledge base-based

extraction approach, which identifies potential concept-related words from the news, and then judges what the main concepts are in the news by calculating the distribution of concept-related terms.

System Architecture:



Algorithm Architecture:

- Parsing news article;
- Extract concept related terms;
- Calculate distribution of key terms;
- Filter based on distribution results;
- Calculate relevance weighting;
- Apply classification model to narrow down on concept tags;
- Combine the results of the above two steps;
- Outputs concept tag results.

Knowledge Base-Based Extraction Algorithm:

- ChinaScope's concept knowledge base comprises of basic information database, concept thesaurus and companies relating to concepts;
- The basic information database of the concept knowledge base includes concept name and unique ID. Currently, there are over 200 thematic concepts;
- Concept thesaurus contains related words to concepts under coverage;
- Concept related company database contains A-share companies that have association to different thematic concepts.

The extraction algorithm first classifies input news, then finds related words by comparing the concept thesaurus and the word segmentation results, and calculates the distribution of related words. According to the distribution of related words in the news, the relevance weighting of the concept in the text is calculated. Information in the title is given more weighting.

Concept Classification Model:

The concept data distribution is similar to the industry's data distribution, insomuch that there is significant imbalance in terms of their coverage in the news. A concept can be an industry, a product, a technology or even an event. The coverage of a large-scale concept may have a large amount of news, while the concept of less popular concept may have a small amount of news reporting on it. As such, we use the "ovr" strategy in our classifier of concepts, which only targets concepts that have enough news to support it. This approach makes sense because concepts that have little historical data to support it are likely to be negligible as an investment theme.

The classification algorithm is based on ChinaScope's proprietary BD-classification algorithm, which is a semi-supervised classification system. The sample category is determined by calculating the distance between the sample and the category center, and the training data can be extended by an iterative method. There are few requirements for the amount of labeled data, and it can handle the issue of imbalance of category samples.

Knowledge Base Maintenance:

As with other knowledge bases maintained by ChinaScope, the concept knowledge base also requires manual upkeep from time to time to ensure new concepts are updated.

3.5. Geographical Tags

Background

Geographical or regional tags are an important element in news content, as they provide additional insight when used in concert with other data tags.

Input / Output

Algorithm Input:

Title and content body of news articles.

Algorithm Output:

{

```
    'stat': 0,
    'version': version,
    'message': [
                    {
                        'txt': '中国-华东-浙江省-金华市-义乌市', #region and upper level regions
                        'code': 'CSF_CN_330782',              #region unique code
                        'name': '义乌市',                      #region name
                        'ename": "Yiwu"                       #region English name
                        'path': 'CSF_000000, CSF_001000, CSF_001001, CSF_CN, CSF_CN_R00003, CSF_CN_330000,
                        CSF_CN_330700,CSF_CN_330782'          # region unique code and upper level regions' code
                }
    ]
}
```

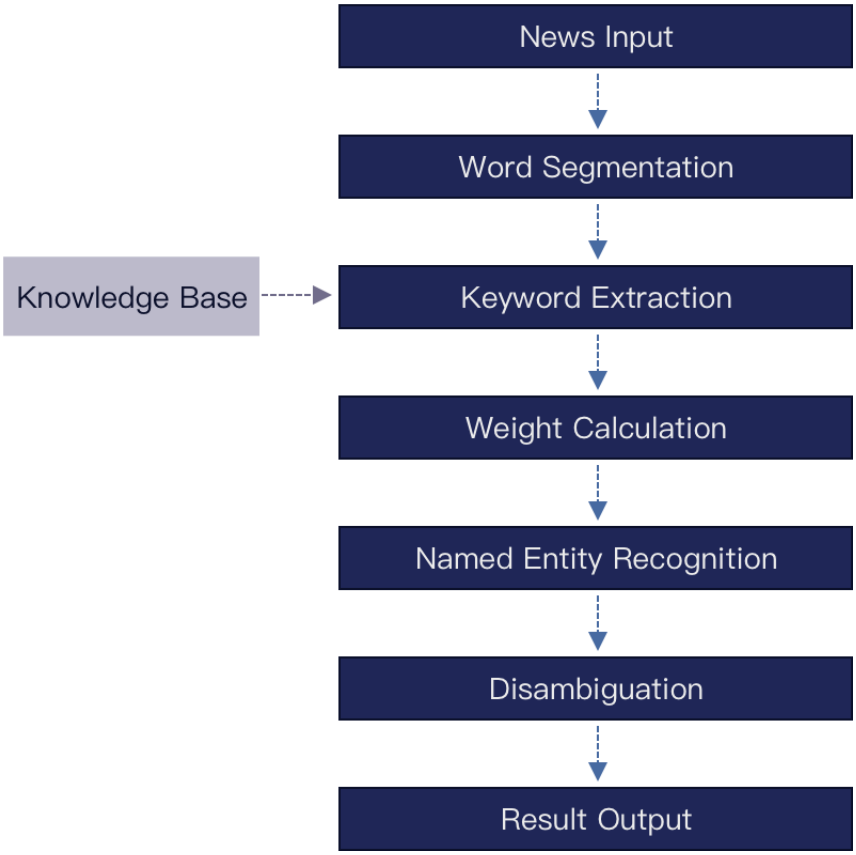
Technical Principle

Technical Background:

To extract regional labels from news, there are usually two methods: The first method is to treat the extracted regions as an entity recognition task, establish an entity recognition model, and directly identify which strings are regions from the news text. The advantage of this method is that it has better generalization ability under normal conditions. The disadvantage is that it is more dependent on the annotated data, and the processing effect on the long tail data is not ideal. Moreover, when the region has a duplicate name, the entity identification method can only identify the name of the region itself, and cannot determine which region the name refers to. The second method is to obtain regional information by extracting regionally related words from the news by establishing a regional knowledge base. The advantage of this method is that it does not need labeled data, and the disadvantage is that it may be affected by ambiguity.

ChinaScope integrates the above two methods, combining the advantages of the two, using the regional knowledge base as the foundation to identify potential region-related words from news, and then eliminates ambiguity through the entity recognition model and disambiguation algorithm.

System Architecture:



Algorithm Architecture:

- Parsing news article;
- Extract region-related key words;
- Calculate the distribution of key words and their weighting;
- Entity recognition on region;
- Perform disambiguation;
- Outputs results.

Knowledge Base-Based Extraction Algorithm:

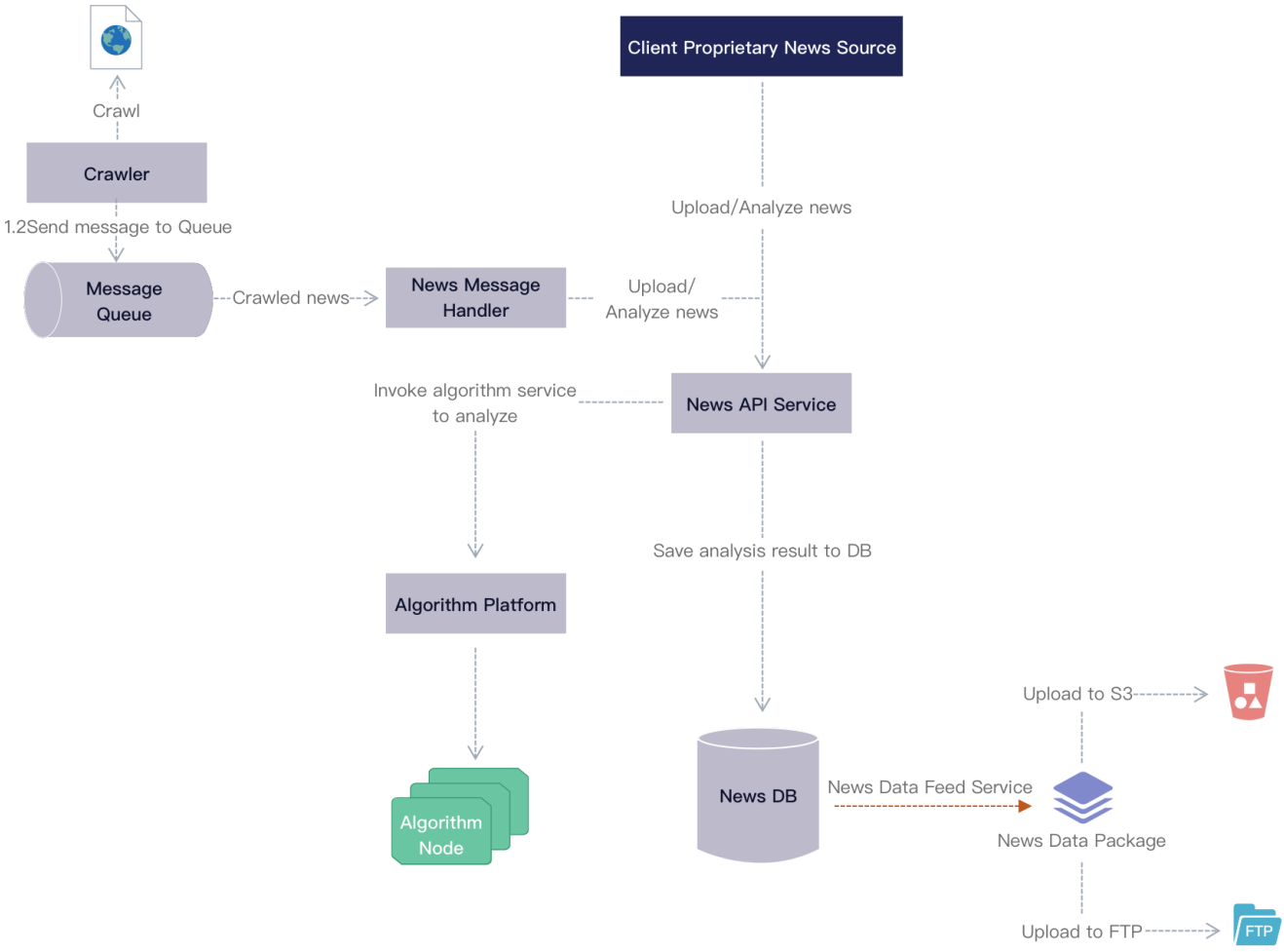
ChinaScope's regional knowledge base includes regional names, the corresponding unique IDs, and their hierarchical relationships.

The extraction algorithm first classifies the input news, then finds related words in the regional knowledge base based on the word segmentation results, and counts the distribution of related words. According to the distribution of related words in the news, the relevance weighting of the region in the text is calculated. The region that appears in the title will have higher weighting. After ambiguity is eliminated, the output result is generated after taking into account of the relevance weighting.

Disambiguation:

There are two common ambiguities in region extraction. The first is that the region appears as a part of a company or an organization, and the second is the phenomenon of a region having the same name as another region (e.g. the equivalent of London in the UK and London in Ohio). For the situation, the identification of companies and institutions using an entity identification model can eliminate this ambiguity. For the second situation, it is necessary to use the affiliation between the regions to determine which region the name specifies.

3.6. SmarTag Overall System Architecture



3.7. SmarTag Quality Evaluation

3.7.1. Company, People, Sector, Product, Event, Concept Tag Quality

As an ongoing part of quality assurance, ChinaScope performs a manual check on the precision and recall rates of SmarTag results every two weeks based on randomly selected samples. The following table reflects the statistical analysis based on 2,000 randomly selected news sample between 4 March 2019 and 19 April 2019.

	Sector	Product	Company	Event	Concept	People
100% Accurate	934	733	955	614	674	226
Partially missing	3	10	51	12	2	3
Erroneous	79	28	53	29	6	2
100% Missing	4	22	53	102	5	8
Precision	92.22%	96.37%	95.00%	95.57%	99.12%	99.13%
Recall	99.26%	95.82%	90.18%	84.34%	98.97%	95.36%

The above summary does not include situations where both algorithm and human judgement determine the absence of tags. (Not identifying information correctly is technically an accurate matching of machine and human endeavors, but it does not yield meaningful information in for the purpose of our quality assessment)

Due to the fact that an article can have multiple tags, so evaluation of quality is based on the following 4 dimensions:

1. "100% Accurate" represents news articles where all the tags extracted from them are completely free of error and no tags are missing ($\mathrm{N}_{100\%\ \mathrm{Accu.}}$);
2. "Partially missing" represents news articles where all of the extracted tags are accurate, but some of the tags that should be extracted have not been ($\mathrm{N}_{\mathrm{Partial\ miss}}$);
3. "Erroneous' represents news articles where the tags extracted are incorrect based on the information in the text ($\mathrm{N}_{\mathrm{Error}}$);
4. "100% missing" represents news articles where 100% of the tags that should be extracted have not been identified ($\mathrm{N}_{100\%\ \mathrm{miss}}$)

Precision Rate Formula:

$$Precision = \frac{\mathrm{N}_{100\%\ \mathrm{Accu.}} + N_{\mathrm{Partial\ miss}}}{N_{100\%\ \mathrm{Accu.}} + N_{\mathrm{Partial\ miss}} + N_{\mathrm{Error}}}$$

Recall Rate Formula:

$$Recall = \frac{N_{100\%\ \mathrm{Accu.}}}{N_{100\%\ \mathrm{Accu.}} + \mathrm{N}_{\mathrm{Partial\ miss}} + N_{100\%\ \mathrm{miss}}}$$

The quality results can vary with different news samples. The results also vary with the scope of identification. For instance, private companies are subject to a larger error rate than listed company identification, because news skew tremendously toward covering well-known companies.

3.7.2. Geographical and Regional Tag Quality

Due to the recent addition of regional tags to the SmarTag system, for the purpose of this paper, we have randomly selected 750 news articles in the month of March 2019, using the same evaluation mythology as the other tags. The results are as follows:

	Region
100% Accurate	705
Partially missing	12
Erroneous	29
100% Missing	3
Precision	96.11%
Recall	97.92%

3.7.3. Factors that Affect Precision and Recall Rates

Precision and recall rates are subject to change based on the following factors:

- Sample Data
  1) Scope of Topical Content
  SmarTag algorithms are built specifically for the finance and economic vertical. If content strays from this area, such as entertainment or sports, the identification and extraction quality would precipitate markedly.

  2) Scope of Textual Style
  As of the authoring of this paper. SmarTag is suited for Chinese language news. Should the textual style stray into other forms such as social media tweets the results would suffer.

- Knowledge Base
  1) Quality and Comprehensiveness of Knowledge Base
  SmarTag relies heavily on different knowledge bases as foundation for identification and extraction, therefore the quality of data in the different knowledge bases as well as the scope of coverage of them will have significant impact on the precision and recall of the tags.

2) Timely Maintenance of Knowledge Base

The continuously changing tides of the world we live places tremendous demand on the timely upkeep of all of our different knowledge bases to ensure an accurate reflection of the real world. Anything less than a studious effort on the maintenance front would ensure a deterioration in precision and recall.

4. Connecting SmarTag to Clue

As mentioned in the previous chapters, ChinaScope's SmarTag system is one with ChinaScope's data schema as a whole. This is demonstrable through the tags interconnection with ChinaScope's knowledge graph system "Clue". Clue is the downstream product of ChinaScope's proprietary automated data extraction system "DAS", which curates stock exchange filings and automatically extracts, synthesizes and standardizes data on financial and business content. Clue tracks 30 types of relationships between entities which aggregate into the following dimensions:

- Shareholdings
- Related party transactions
- Competitors
- Product-to-product supply chain
- Company-to-company supply chain
- Co-mentions in news and exchange filings
- Employment and directorships
- Guarantees
- Investments

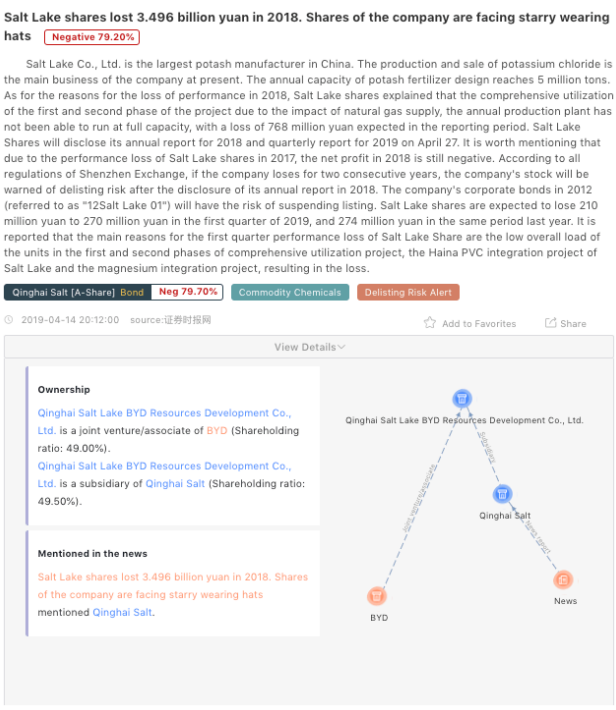
4.1. Relationship Mapping to Companies

Through knowledge graph connections, we can discover direct and indirect risk and value pathways from documents to portfolio assets.

Examples: The following are examples of connecting SmarTag through Clue knowledge graph system to individual assets in any given portfolio.

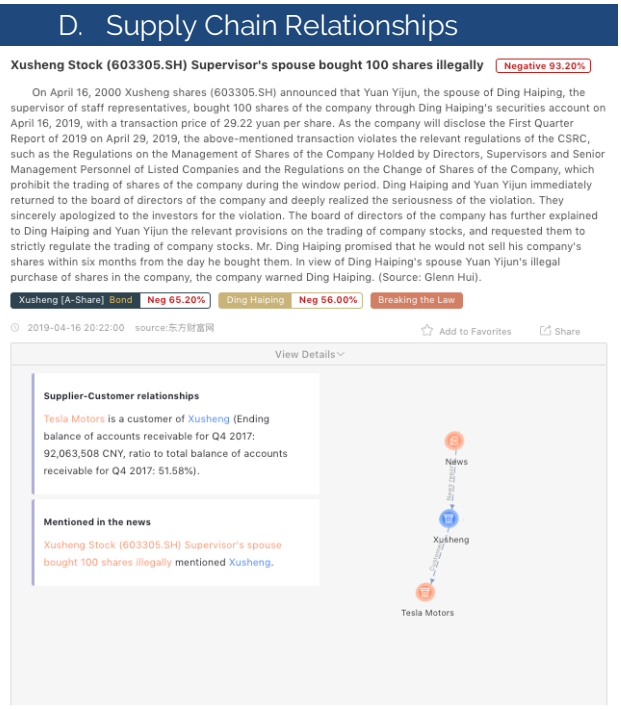
A. Industrial Value Chain Relationships

B. Ownership Relationships



C. Personnel Relationships

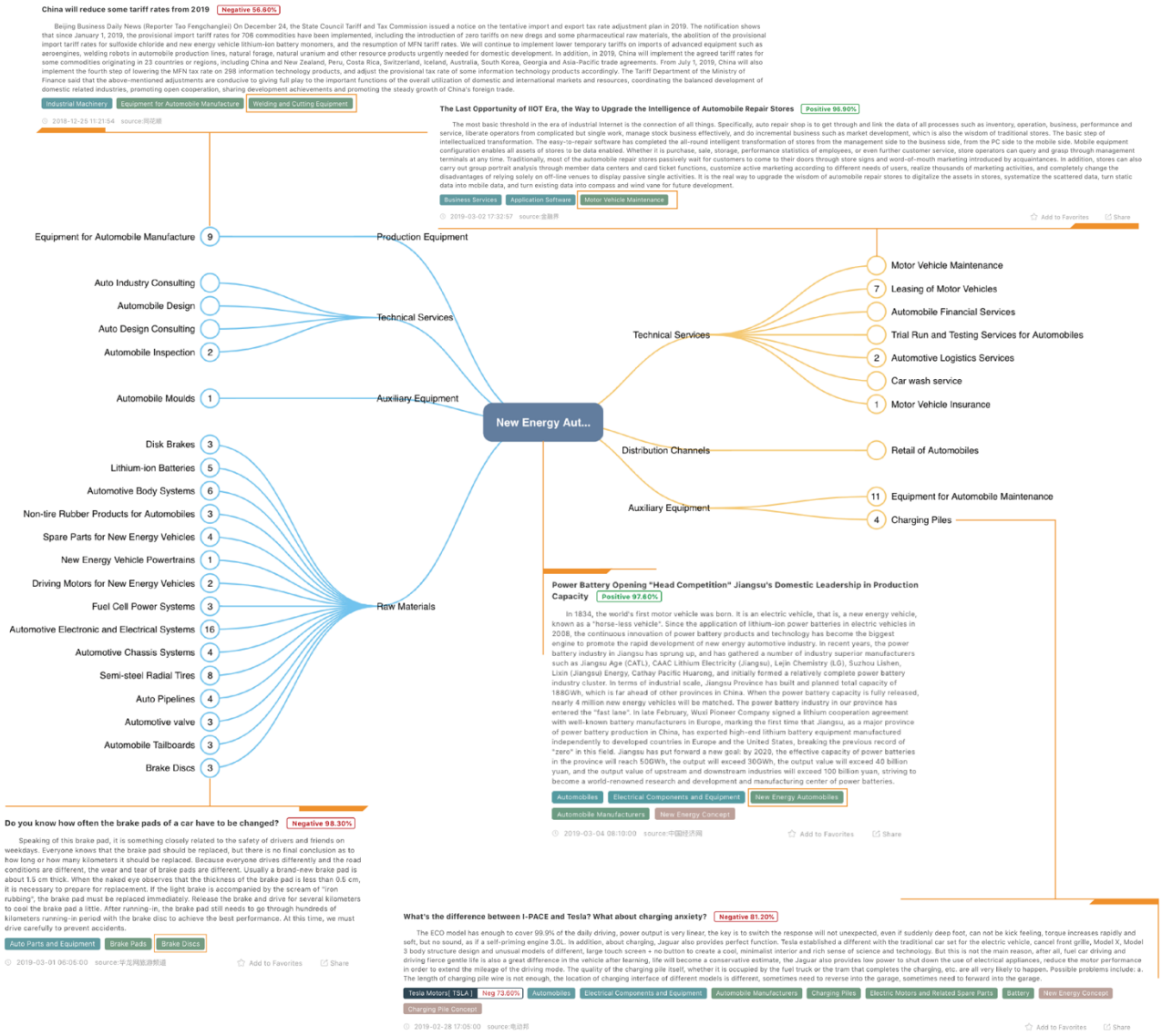
D. Supply Chain Relationships



4.2. Relationship Mapping to Industrial Value Chain (Product-to-Product Supply Chain)

An important part of the Clue knowledge graph system is ChinaScope's Industrial Value Chain data structure. This is based on the product to product relationships established on the SAM product tree. Since the industry and product tags map directly to SAM data, it can naturally aggregate in accordance to the SAM taxonomy, which in turn allows for tracking of news along the industrial value chain.

Example: The following illustration is based on the supply chain vertical for New Energy Automobiles (i.e. Electric cars).



5. Sentiment Analysis

5.1. Article-Level Sentiment

Article-level sentiment, as the name indicates, relates to a sentiment score assigned to a news article as a whole. The sentiment score falls into three buckets: positive, negative and neutral. Each bucket would carry a probability score between 0 to 1.

Input / Output

Algorithm Input:

Title and text body of news article.

Algorithm Output:

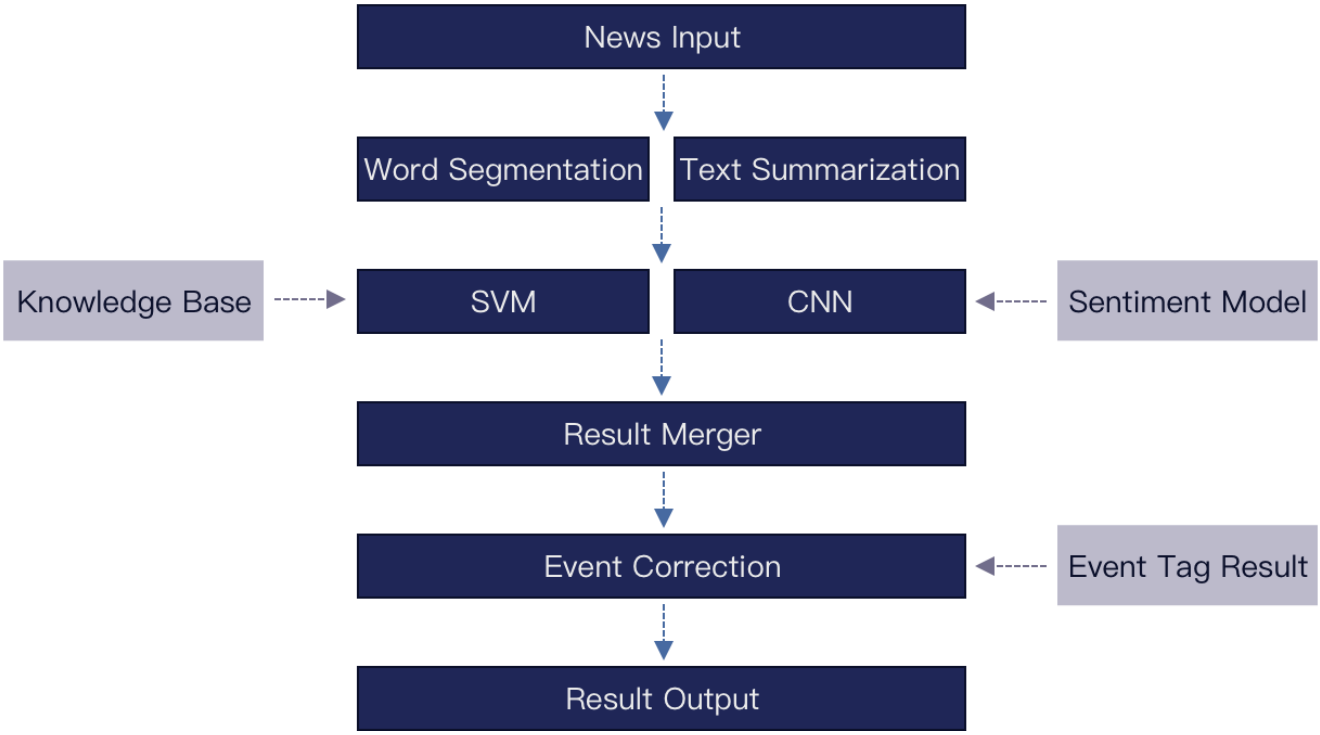
```
{
    'stat': 0,
    'message': [{'pos': {'0': 0.0325, '1': 0.9675, '2': 0.0}}],    #news' sentiment and probability
    "version": version
}
```

Technical Principle

Technical Background:

There are various methods for sentiment orientation analysis of text, such as a bag of polarity words method, machine learning model methods, and deep learning model-based methods. Deep learning is the mainstream method adopted by the market at present. ChinaScope uses a combined approach based on convolutional neural network and support vector machine model, with predetermined polarity event database as correction agent.

System Architecture:



Algorithm Architecture:

ChinaScope's sentiment algorithm first extracts summary abstract from news, and then analyzes the headline and body of the text using support vector machine model, and then uses the convolutional neural network to analyze the headline and the abstract of the news, after which the aforementioned event extraction algorithm is used to extract events from the news headline. Finally, the three sets of results are combined together to generate a sentiment score for the article as a whole.

Multi-Model Combination:

The sentiment algorithm adopts a multi-model combination method, which consolidates a convolutional neural network model and a support vector machine model. The convolutional neural network can extract the features in the text well and obtain accurate analysis results through nonlinear calculations. At the same time, the computational complexity is reduced by parameter sharing. The sentiment algorithm also introduces an expansion convolution technique, which can cover more features than the generic convolution kernel without increasing the amount of computation and can capture features that are distant from the text. However, the convolutional neural network model must specify the length of the text when analyzing it. When the length of the text exceeds the limit value, the text should be truncated or extracted. In the sentiment algorithm, the text is compressed by extracting the news summary abstract. However, text compression inevitably brings information loss, so the news sentiment algorithm adds a support vector machine model.

The support vector machine model is a prototypical machine learning model, which is characterized by good generalization performance and robustness in combatting overfitting. Moreover, the input of the support vector machine model is a one-hot vector, which can vectorize the news full text and then input it into the model, thus avoiding information loss. The results of the support vector machine model are less correlated with the results of the convolutional neural network model, so combining the two models can improve model effect.

Event Correction Agent:

Due to the wide spread nature of news, manually labeled training data cannot cover the full distribution of all news data. Therefore, in order to improve the generalization ability of the algorithm and adapt better to the data changes, the sentiment algorithm adds a bag of words approach based on event polarity for result correction. For news articles that have clearly delineated events, the correction feature can be quite evident.

Evaluation Metrics

In a randomly selected sample of 1,000 news articles taken between February and March of 2019, eliminating noisy content (e.g. advertising), 858 articles remained for testing. The following table shows the statistical distribution of the evaluation results.

		Algorithm Output		
		Neutral	Positive	Negative
Human Inspection	Neutral	81	18	53
	Positive	41	426	54
	Negative	4	1	180

Algorithm accuracy rate is 80.07% based on the following formula:

$$accuracy = \frac{81 + 426 + 180}{81 + 18 + 53 + 41 + 426 + 54 + 4 + 1 + 180}$$

Due to the fact that investors tend to show preference on certainty around positive news and that they don't want to miss negative news, ChinaScope focuses on precision for positive sentiment and recall for negative sentiment.

Positive precision rate ($Precision_{pos}$) is 95.73% :

$$Precision_{pos} = \frac{426}{18 + 426 + 1}$$

Negative recall rate ($Recall_{neg}$) is 97.30%:

$$Recall_{neg} = \frac{180}{4 + 1 + 180}$$

Even though there are three sentiment buckets, they are not perceived to be equal. Positive and negative sentiment attract much more attention than neutral. Also, the subjective understanding of what's positive versus what's negative is a lot less prominent than that between neutral and positive and between neutral and negative. Therefore, it is more important to pay attention to the error rate associated with misidentification between positive and negative sentiment. This is denoted as $Error_{pos\_neg}$, which is calculated below.

$$Error_{pos\_neg} = \frac{54 + 1}{81 + 18 + 53 + 41 + 426 + 54 + 4 + 1 + 180}$$

Article-level $Error_{pos\_neg}$ is 6.41%

5.2. Entity-Level Sentiment

ChinaScope's entity-level sentiment algorithm tracks the sentiment scores for companies and persons that are mentioned in news articles. The sentiment score falls into three buckets: positive, negative and neutral. Each bucket comes with a probability score of between 0 and 1.

Input/ Output

Algorithm Input:

Title and body text of news article.

Algorithm Output:

```
{
    'message': [
        {
            'com_list':                          # company information list
                [
                    {
                        'com': company abbreviate,
                        'code': ticker,
                        "comcode": company unique code,
                        "chiname": company full name
                        'pos': [label, ...],    #label = 0/1/2, polarity of sentences with the company
                        'sentence':[ sentence, ...] #sentences with the company
                        'sen_idx':[ int, ...] #indexes of sentences with the company
                        'general_pos': 0/1/2 # polarity of the news for the company
```

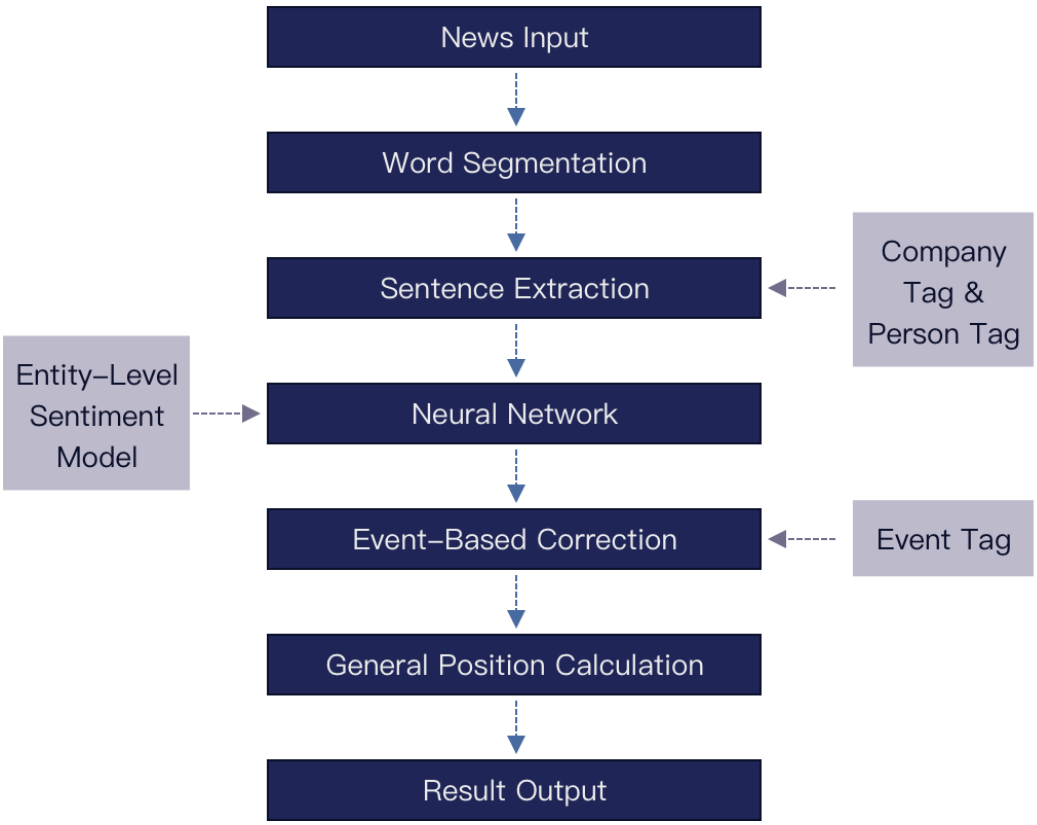
```
                    'weights': {'0': 0.3406, '1': 0.0207, '2': 0.6387}    #weight of general_pos
                ],
            ],
        'per_list':                         #person information list
            [
                {
                    'per': person name,
                    'com': company (person service) name,
                    'code': ticker,
                    'pcode': person unique code
                    'comcode': company unique code
                    'pos': [label, ...],    #label = 0/1/2 polarity of sentences with the person
                    'sentence':[ sentence, ...] # sentences with the person
                    'sen_idx':[ int, ...] #indexes of sentences with the person
                    'general_pos': 0/1/2 # polarity of the news for the person
                    'weights': {'0': 0.3406, '1': 0.0207, '2': 0.6387}    #weight of general_pos
                },
            ]
        }
    ]
    'stat': 0
    "version": version
}
```

Technical Principle

Technical Background:

Extracting sentiment orientation for entities is quite different from performing this task for the entire news article as a whole. Typically, we see this exercise being done for comments on commercial goods, e-commerce sites and movie reviews. Assigning sentiment to entities in news is a relatively recent phenomenon driven by investment needs. ChinaScope applies a neural network model for analysis and adds business information corrections based on the model results to improve the generalization ability of the algorithm.

System Architecture:



Algorithm Architecture:

The entity-level sentiment algorithm utilizes the company and people tags as foundation, match up the entities with the sentences within which they are contained, and then inputs the match into a convolutional neural network model for computation. We then analyze the sentences that contain entities based on events. We combine the results to generate sentiment scores for these sentences.

Classification Model:

The entity-level sentiment algorithm performs a classification calculation on each sentence containing the entity. When there are multiple entities in a sentence, different words are embedded with different entities. After embedding, convolutional vectors are convolved through multiple convolutional check vectors of different sizes, and then maximum pooling is performed. After the pooling, the nonlinear layer is connected, and finally the classification result is obtained through the softmax layer. Over-fitting is controlled in the middle by the dropout layer.

Event-Based Correction Agent:

Event-based correction agent is added to adjust for the fact that labeled data may not be representative of the full distribution of situations in the news. By adding this correction layer, we can

mitigate the shortfall brought upon by limited labeled data.

Evaluation Metrics

In a randomly selected sample of 1,405 news articles taken between February and March 2019, we have 1,894 entities related texts. The following table tabulates the results of quality evaluation on entity-level sentiment.

		Algorithm Output		
		Neutral	Positive	Negative
Human Inspection	Neutral	588	143	110
	Positive	147	481	26
	Negative	54	26	319

Entity-level overall accuracy rate is 73.28%:

$$accuracy = \frac{588 + 481 + 319}{588 + 143 + 110 + 147 + 481 + 26 + 54 + 26 + 319}$$

The misidentification of positive and negative sentiment rate ($Error_{pos\_neg}$) is 2.75%:

$$Error_{pos\_neg} = \frac{26 + 26}{588 + 143 + 110 + 147 + 481 + 26 + 54 + 26 + 319}$$

Another issue with the accuracy calculation is the degree of polarity. Due to the fact that 100% probability is shared between the three sentiment buckets, the highest scoring bucket would emerge as the sentiment (e.g. if positive has the highest probability compared to the other two, then the sentiment tag would be positive). Since all three numbers need to add up to 100%, then the lowest possible probability number a bucket can have to emerge as the tag would be 34%. Clearly, in that case, the confidence level for that bucket wouldn't be very high. Therefore, another way we judge quality is by comparing accuracy rates based on probability scores of above 80% as a minimum confidence threshold. The following table tabulates the results of quality evaluation based on 80% as minimum probability.

		Algorithm Output		
		Neutral	Positive	Negative
Human Inspection	Neutral	354	38	35
	Positive	60	257	11
	Negative	33	5	175

Above 80% probability sentiment score accuracy rate ($accuracy_{80}$) is 81.20%:

$$accuracy_{80} = \frac{354 + 257 + 175}{354 + 38 + 35 + 60 + 257 + 11 + 33 + 5 + 175}$$

Above 80% probability sentiment error rate between positive and negative ($Error_{pos\_neg80}$) is 1.65%:

$$Error_{pos\_neg80} = \frac{11 + 5}{354 + 38 + 35 + 60 + 257 + 11 + 33 + 5 + 175}$$

6. Algorithm and Lexicon Base Management

6.1. Algorithm Version Management

When a new version of an algorithm is released, ChinaScope creates quick snapshot metatags to the algorithm. We separately manage each version of each algorithm, and the version management adopts Semantic Version specification (an example of the format is V0.4.0.9). This keeps the evolution of each algorithm independent. At the same time, the output result of each algorithm also records the corresponding algorithm version number, so that the analysis result can be retroactively analyzed by past versions of the algorithm.

Algorithm publication falls into small version releases and large new edition releases. Major edition releases usually do not surpass 3 times a year.

6.2. Lexicon Base Management

An important part of ChinaScope's overall knowledge base that underscores much of our NLP engine is our lexicon base. Given the constant evolution of vernacular, the expansion of our coverage, and the need to improve the accuracy of our NLP capability, we have a dedicated team of Lexicon stewards who constantly updates and prunes our thesauruses and word banks.

6.3. Quantitative Use

In the field of quantitative investment analysis, data noise or biases formed by any external factors will affect the integrity of data back testing results. Therefore, maintaining the continuity of data element input in the model is key. As NLP is still a developing science, the need to maintain consistency in methodology and underlying data would come at odds with the need for timely and robust upgrades. Therefore, it is important to strike a high degree of balance between technological development and data stability in the way of system management, ensuring consistent output for the purpose of quantitative investment analysis. Presently, most quantitative uses for news NLP falls in the application of company sentiment scores in extracting tradable signals. As such, in order to eliminate the inconsistency of underlying data change, ChinaScope does not include lexicon usage in the extraction of company sentiment scores, instead we adopt a pure machine learning approach. Through clear cut version management and controlled algorithm upgrade, ChinaScope minimizes inconsistency between back testing and live situations.