

Isolation and Comparison of Company Names in the News Text with Tickers of Companies Traded on Stock Exchanges

Stanislav Bachinsky[1] and Alexey Stankus[2(✉)]

[1] Dubna State University, 19 Universitetskaya street, Dubna 141980, Russia
[2] Saint Petersburg State University, 7-9 Universitetskaya Emb., Saint Petersburg 199034, Russia
alexey@stankus.ru

Abstract. Determining the names of companies in the news text is an important task in predicting changes in the prices of companies' shares. There is a large number of works devoted to this problem, a large number of algorithms have been created. On the other hand, in order to understand which stocks of which company may be affected by this or that news, it is necessary to understand exactly which ticker corresponds to the name of the company extracted from the text of the news.

Behind the seeming simplicity of the question, there is the difficulty of comparing the descriptions of tickers that exchanges provide with incomplete company names that are indicated in ordinary, non-financial news. Often in the news, companies are written in abbreviated form or abbreviations. At the same time, the description of tickers contains deliberately redundant information, which also needs to be isolated and removed to improve the results of matching names.

In this article, we will analyze several possible approaches to solving this problem and present an algorithm developed by us for comparing company names found in news texts with stock tickers.

Keywords: Company names · News · Tickers · Matching · Neural networks · Gestalt pattern matching · NASDAQ · NYSE · Sequential elimination method

1 Introduction

News is a change and reflection of some fragment of reality, timely covered in the media, which is of interest and value to a large number of audiences.

Thousands of news are published daily, containing information about everything that happens around us. News analysis is the cornerstone for predicting price changes and volatility in stocks traded on the stock exchange. Identifying the company that is being talked about in the news is the first step towards predicting the stock price using the news.

There are number of works aimed to extract company names from news [1–4]. But in this article discusses approaches that allow us to select the names of companies from the text of general news, as well as a method we have developed that allows us to compare the resulting name with a possible ticker (a short name in the exchange information of quoted instruments) of a company.

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12956, pp. 370–380, 2021.
https://doi.org/10.1007/978-3-030-87010-2_27

2 The Essence of the Problem

Unlike financial news text, company names mentioned in general news text may not contain the full company name or designator words such as Inc, Corp, Limited, Common Stock. Such words allow you to distinguish the company name from other proper names, or ordinary words. For example, the sentence "Apple is the healthiest fruit" is about apples, but "Apple is an American corporation" refers to the name of the company. The method of extracting company names from text should be applied universally to general and financial news.

Unfortunately, company names in general news and their official names often differ. Differences can be in punctuation marks, casing, references to subsidiaries, and inaccuracies in the selection methods.

1) …. expect to see Nvidia's new pro-grade desktop GPUs in…
2) … in a McDonalds parking lot in the…
3) …. and Walmart Inc's Flipkart benefit only…

Such differences lead to the omission of company names in news texts, which later gives an error when trying to create a predictive model for a particular company or have grown.

This brings us to one requirement - False Negative minimization. At the same time, the presence of a large number of False Positives suits us, since all unnecessary words or phrases that would be mistaken for company names will be discarded at the stage of comparison with stock tickers.

On the other hand, the description of a ticker on the exchange contains redundant information and is broader than just the name of the company to which the ticker refers. For example - "International Business Machines Corporation Common Stock or Summit Hotel Properties, Inc. 9.25% Series A Cumulative Redeemable Preferred Stock". This feature of the ticker descriptions leads us to the need to process the ticker descriptions in order to increase the likelihood of subsequent matches and reduce the False Negative results of the matches.

Such disagreements in spelling lead to the fact that solving the problem of matching the names of companies mentioned in news texts and tickers of these companies traded on exchanges requires the solution of 3 subtasks:

a. Highlight the names of companies mentioned in the news;
b. Remove unnecessary text from tickers, leaving only the core.
c. Compare highlighted company names in news texts and optimized ticker descriptions.

3 Extraction of Company Names from News Text

In this section, we will consider several existing methods for solving the problem of highlighting company names in the text.

3.1 Linguistic Analysis Method

This method is based on an article in Proceedings The Seventh IEEE Conference on Artificial Intelligence Application [5]. The solution to the problem of highlighting company names in the text, described in this article, is a heuristic: testing and improving the method on the texts of real news.

The method is based on sequential iteration over the words of the news text. The main components of the method are shown in Fig. 1. The condition for stopping the search is the detection of a word from the list of trigger words, such as: Inc, Corp, Co, Plc, Ltd, etc. In case of a stop, the previous word will be read and checked:

a) mixed case;
b) whether the word is a union;
c) whether the word is a sentence/phrase boundary.

Based on the results of the checks, the following outcomes are possible:

a) The current phrase (the trigger word and the previous words that passed the test) is recognized as the name of the company.
b) The current phrase is not recognized by the name of the company.

The disadvantage of this approach lies in its essence: the method is heuristic. To effectively identify titles, a large list of trigger words is required, which can be generated by manually marking news texts. However, the main disadvantage is that the company name must contain a trigger word in order to be identified. The phrase Amazon Inc, the considered method, would successfully define the name of the company, but the words Amazon or Amazon.Com would be omitted. This method is effective for highlighting company names in the text of financial news, in which full company names are indicated, containing suffixes and other indicators. This method is not suitable for general news that uses partial company names (for example, without the Inc: Amazon.Com or Amazon suffix).

3.2 Method Using Neural Networks

Another approach to solving the problem of extracting company names from the text is the use of neural networks called NER (Named Entity Recognition). This article discusses a pipeline developed by Explosion [6]. With the help of the pipeline (Fig. 2), we will carry out lexical analysis of the text. The pipeline consists of blocks [7]:

1. word2vec [8] - a component for converting words into a vector;
2. tagger - a component that predicts the likelihood of tags based on vectors of tokens;
3. parser - a component intended for a dependency parser;
4. attributer_ruller - a component that allows you to set attributes;
5. lemmatizer - a component for assigning basic forms to tokens using rules based on part-of-speech tags;
6. ner - named object recognition component.

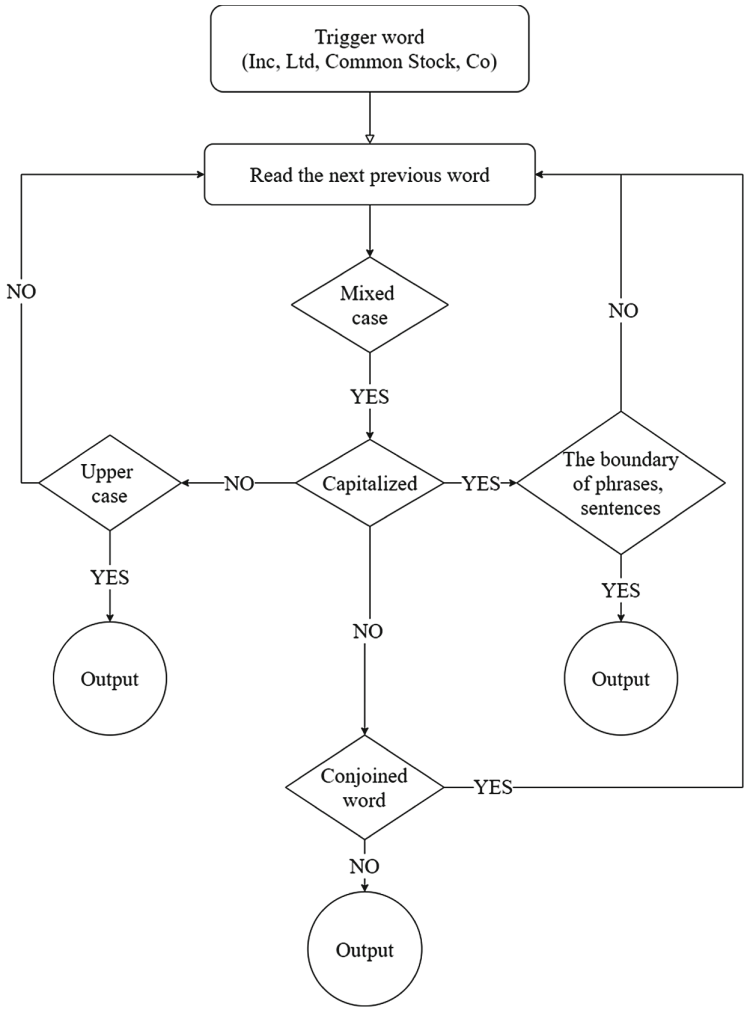

Fig. 1. Linguistic analysis method.

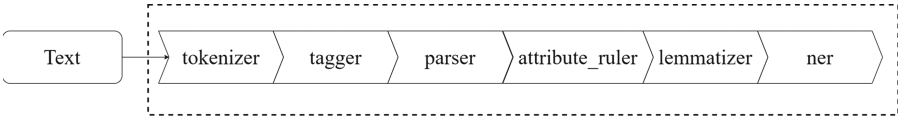

Fig. 2. NER data flow.

This method allows you to extract company names from the text of the news. Unlike the previously described method, the pipeline identifies companies regardless of the completeness of their name. However, it gives a high percentage of false positive errors, which suits us at this stage, due to the fact that in the future, the list of words and phrases

that are taken for company names will undergo one more check and the excess will be discarded.

Below is an example of words and phrases that appear in company names:

a. Footfall
b. Cesar basanta
c. USA
d. The African Basketball League
e. Bluetooth

However, for us, it is more important not to miss the loss of the company name in the news than to get unnecessary meaning that will be discarded in the next step - matching the highlighted prospective company names with stock tickers.

4 Optimizing Description Tickers

The names of companies used on the exchange consist of four parts: a distinctive element, a descriptive element, a corporate element, and an additional element (Fig. 3).

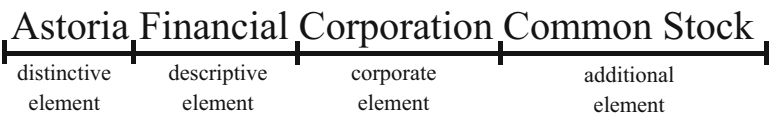

Fig. 3. Ticker company name description.

A distinctive element can be a personal name, a place name, or a made-up word. Its purpose is to distinguish companies with similar descriptive elements.

The descriptive element is needed to describe the nature of the business and allows you to distinguish between companies that have the same or similar distinguishing element: United States Steel Corporation and United States Cellular Corporation.

The corporate element is a suffix that carries information about the type of company: private limited company, public limited company, group of companies, and others.

Additional element - additional information about the stock: its type, duration, percentage, and more.

Company names mentioned in general news are often abbreviated with only distinctive and descriptive elements of the name. The additional element identifies a stock, not a company, therefore, to correctly match company names from the news text and company names from the stock exchange, the additional element should be removed.

To remove an additional element, you need to find the border between the additional and the descriptive element, that is, the corporate element. There is a limited number of words used as a corporate element, so we extract the 50 most frequently mentioned words in the entire list of companies from the stock exchange. Most of the words on the list are corporate words:

{Inc, Corporation, Shares, Class, Fund, Corp, Holdings, Trust, Depositary, Group, Ordinary, Limited,…}.

The list also contains words denoting the type of stock {Shares, Common Stock}, as well as popular words in the descriptive element {Energy, Oil}, therefore, words that are not related to the corporate element have also been removed.

5 Matching the Highlighted Company Name to the Ticker

5.1 Comparison Method for Hashed Sequences

The sequence comparison algorithm was published back in the late 1980s by John W. Ratcliff and John A. Obershelp under the hyperbolic title "Gestalt pattern matching". [9] The idea is to find the longest contiguous matching subsequence that does not contain "unnecessary" items. These "unnecessary" elements are in some sense uninteresting, for example, blank lines or spaces. (Garbage handling is an extension of Ratcliff and Obershelp's algorithm.) The same idea is then applied recursively to the portions of sequences to the left and right of the corresponding subsequence. This does not result in minimal edit sequences, but tends to produce matches that "look right" to humans.

The similarity of two strings S_1 and S_2 is determined by the formula, calculating twice the number of matching characters K_m divided by the total number of characters of both strings. The matching characters are defined as the longest common substring (LCS) plus recursively the number of matching characters in the non-matching regions on both sides of the LCS:

$$D_{ro} = \frac{2K_m}{|S_1| + |S_2|}$$

where the similarity metric can take a value between zero and one:

$$0 \leq D_{ro} \leq 1$$

The algorithm is effective for comparing slightly different words or sequences, but it is not suitable for comparing the results of methods for identifying potential companies with stock tickers, since the phrase representing a potential company found by the neural network method often contains words that are not included in the name of the company. For example, "Morgan Stanley Tactical Value Investing" extracted from news and its corresponding "Morgan Stanley have" a weak similarity ratio, and the phrase "Root Blocks in has" a strong similarity to "Riot Blockchain".

5.2 Sequential Elimination Method

The developed algorithm allows us to remove restrictions on the presence of unnecessary words in the studied phrase. Figure 4 shows a diagram of the algorithm for matching the selected, from news, word or phrase, and the name of the company from the stock exchange. The algorithm uses processed company names from the stock exchange, stock tickers, and words and phrases extracted from news, presumably being company names.

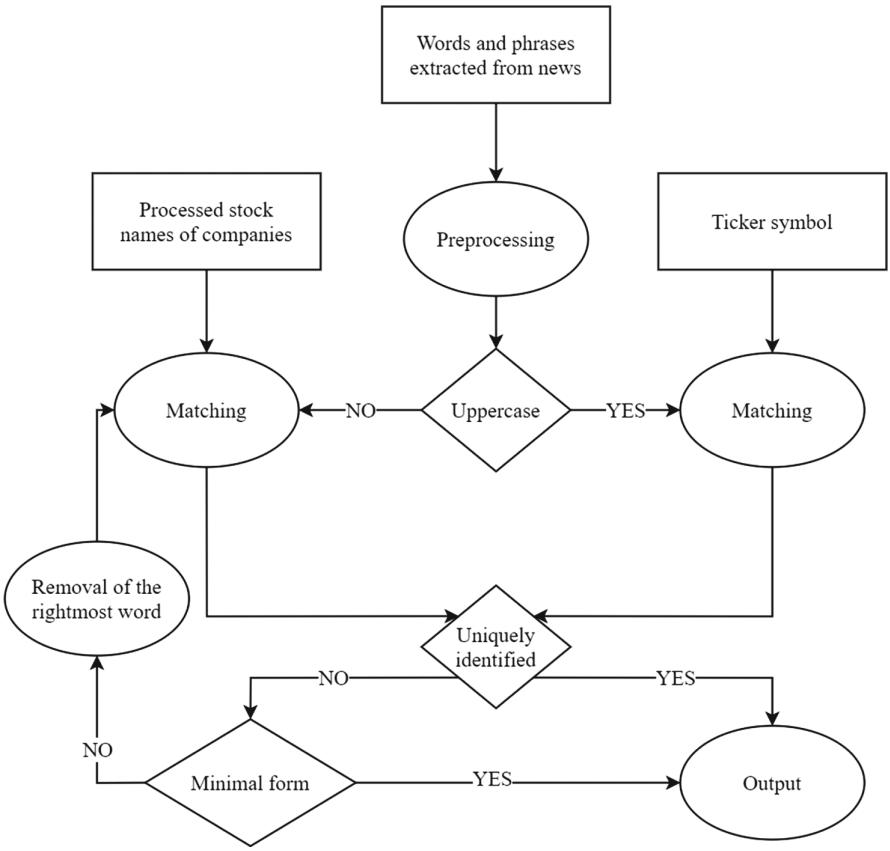

Fig. 4. Sequential elimination method.

Processing of Company Names Extracted from News Text

At the preprocessing stage, the problems associated with the operation of methods for highlighting words and phrases that are potentially company names in the text are minimized: the addition of insignificant spaces at the beginning or end of words and phrases, an error in the encoding of individual characters, incorrect addition of punctuation and quotation marks: "Mc Donadls", 'McDonald^{TM}s', 'McDonald's'. The solution is to use regular expressions and two-way encoding.

Processing Companies in Upper Case

The identified word or phrase, potentially a company name, in uppercase, may be a stock ticker (IBM, Bio), and may also be a stylistic feature of the news (APPLE instead of Apple). First, the word/phrase is compared with the stock ticker. If the comparison did not give any results, the entire word/phrase, except for the first letter, is converted to lower case.

Handling the Incorrect Selection of Potential Companies
There are situations when words that are not related to it are added after the actual name of the company in the identified phrase, which is potentially the name of the company, for example:

a. Fiat Chrysler Automobiles FCAU.N.Nissan replacing Fiat Chrysler Automobiles.
b. Morgan Stanley Tactical Value Investing instead of Morgan Stanley.

To process such cases, the rightmost word is removed from the phrase one by one, in the event that the part remaining after the deletion is minimally sufficient to unambiguously identify the company, after which it is compared with the exchange names. The minimum sufficient for an unambiguous definition is a phrase consisting of two words, one of which is not in common use (Astoria from Astoria Financial Corporation), or a phrase of 3 words.

Determination Results
As a result of the method of successive exclusions, the identified phrase is marked as unambiguously defined or as undefined. The resulting value is entered into the database with the ticker of the corresponding company.

6 Experiments

6.1 Initial Data

The experiments were based on news collected from BBC News, Business Wire, Breitbart News, CNN, The New York Times, Reuters, Washington Post, Trustnodes and Yahoo News from January 2018 to October 2020.

In total, 988 102 texts of various news were collected. To test the methods of allocation and comparison with the stock ticker of company names, 100 news were selected from the general list, 62 unique company names were manually marked in them, and 28 unique names of companies with shares on the NASDAQ or NYSE stock exchange were manually assigned the corresponding stock tickers.

To test the algorithm for matching phrases with a stock ticker, a list of companies trading on the NASDAQ and NYSE stock exchanges was used [10].

6.2 Testing Methods for Highlighting Company Names

Methods for extracting company names (linguistic and neural network) were tested on a test dataset of 100 news. Manual analysis showed that in these 100 news items, there are references to 62 companies. The experimental results are shown in Table 1.

It should be noted that the method using neural networks gives good results in terms of the number of correctly identified companies but produces a very large amount of unnecessary information that will have to be processed in the future.

Table 1. Highlighting company names.

Method	Total found	Detected	Not a company
Actual values	62	62	-
Linguistic analysis	32	8 (12.9%)	24
Neural networks	514	58 (93.5%)	456

6.3 Testing the Method of Matching Company Names and Stock Tickers

The method of comparing companies and stock tickers was tested on data obtained from the following initial processing:

- Checking the operation of the method of sequential elimination and the Gestalt Pattern method on the names of companies marked up manually;
- Checking the method of successive elimination and Gestalt Pattern method on the results of the linguistic method;
- Checking the method of successive elimination and the Gestalt Pattern method on the results of the neural network algorithm for extracting company names from the news text.

During manual checking, we received 62 unique company names, of which 28 names correspond to a specific ticker, 34 corresponds to an empty ticker (that is, the company's shares are not traded on NASDAQ and NYSE). The experimental results are presented in Table 2. All percentages are calculated from the correct number of tickers.

6.4 Experimental Results

The results of the experiments showed that the linguistic method is not suitable for solving the problem of separating the names of companies from the texts of ordinary, non-financial news, the result is less than 13% of the total number of companies mentioned.

The method using neural networks showed a significantly better result - 93.5%, but it is worth noting a very large number of unnecessary phrases that are perceived as company names and interfere with further processing.

To solve the matching of the ticker to the company name, we have the following result:

The Gestalt Pattern method shows an average result when you manually isolate company names (82.1%), but when you add random (obviously unnecessary) words that fall into phrases, the accuracy of the method drops significantly (57.1%). With a large amount of garbage data, the method gives a large number of errors.

The Sequential elimination method shows an excellent result (100%) with a good isolation of the company name from the news (manual method) and a good result with the presence of unnecessary words and phrases (89.2%).

Table 2. Matching company names and stock tickers.

Method	Number of companies	Correct ticker	Wrong ticker	Ambiguous tickers	No ticker
Actual values	62	28	0	0	34
Gestalt Pattern method (based on actual values)	62	23(82.1%)	7(25%)	3	29
Sequential elimination method (based on actual values)	62	28(100%)	3(10.7%)	0	31
Gestalt Pattern method (based on linguistic analysis values)	32	8(28.5%)	6(21.4%)	1	17
Sequential elimination method (based on linguistic analysis values)	32	8(28.5%)	4(14.2%)	0	20
Gestalt Pattern method (based on neural networks values)	514	16(57.1%)	27(96.4%)	66	405
Sequential elimination method (based on neural networks values)	514	25(89.2%)	5(17.8%)	56	428

7 Conclusion

The article examined and conducted tests of methods for separating company names from the texts of ordinary, non-financial news, as well as methods for comparing the selected company names to stock tickers.

The experimental results show that methods based on neural networks show good results in identifying company names, but they carry a large amount of garbage that contaminates the data obtained, which ultimately leads to a slight but increase in errors in the best of the algorithms for matching the company name to the exchange tickers.

To improve the results, it is necessary to concentrate on training neural networks for a more correct selection of company names, with a decrease in the amount of unnecessary information.

References

1. Loster, M., Zuo, Z., Naumann, F., Maspfuhl, O.: Improving company recognition from unstructured text by using dictionaries, March 2017
2. Oliya, A.: News to company linking with Bert, November 2019
3. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C., Jatowt, A.: YAKE! collection-independent automatic keyword extractor. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 806–810. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_80
4. https://www.analyticsinsight.net/company-names-standardization-using-a-fuzzy-nlp-approach/, 28 March 2020
5. Rau, L.F.: Extracting company names from text. In: Proceedings the Seventh IEEE Conference on Artificial Intelligence Application (1991)
6. Introducing spaCy. explosion.ai. Accessed 18 Dec 2016
7. Models & Languages. spaCy Usage Documentation. spacy.io. Accessed 08 Feb 2021
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
9. Ratcliff, J.W., John, A.: Obershelp Gestalt pattern matching. https://en.wikipedia.org/wiki/Gestalt_Pattern_Matching
10. Tickers list. http://www.nasdaqtrader.com/trader.aspx?id=symboldirdefs