Journal of Open Humanities Data

Mining an English-Chinese parallel Dataset of Financial News

RESEARCH PAPER

NICOLAS TURENNE
ZIWEI CHEN
GUITAO FAN
JIANLONG LI
YIWEN LI
SIYUAN WANG
JIAQI ZHOU

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Parallel text datasets are a valuable for educational purposes, machine translation, and cross-language information retrieval, but few are domain-oriented. We have created a Chinese–English parallel dataset in the domain of finance technology, using the *Financial Times* website, from which we grabbed 60,473 news items from between 2007 and 2021. This dataset is a bilingual Chinese–English parallel dataset of news in the domain of finance. It is open access in its original state without transformation, and has been made not for machine translation as has been used, but for intelligent mining, in which we conducted many experiments using up-to-date text mining techniques: clustering (topic modeling, community detection, *k*-means), topic prediction (naive Bayes, SVM, LSTM, Bert), and pattern discovery (dictionary based, time series). We present the usage of these techniques as a framework for other studies, not only as an application but with an interpretation.

CORRESPONDING AUTHOR:

Nicolas Turenne

BNU-HKBU United International College, UIC, Division of Science and Technology, Zhuhai Guangdong, China

nicolas.turenne@univ-eiffel.fr

KEYWORDS:

English-Chinese; text mining; clustering; classification; patterns

TO CITE THIS ARTICLE:

Turenne, N., Chen, Z., Fan, G., Li, J., Li, Y., Wang, S., & Zhou, J. (2022). Mining an English-Chinese parallel Dataset of Financial News. *Journal of Open Humanities Data*, 8: 9, pp. 1–12. DOI: *https://doi.org/10.5334/johd.62*

1 INTRODUCTION

The investigation of classical and new text mining methods using a bilingual dataset can enhance the meaningfulness of comparisons of these techniques. The original way to use a parallel text dataset is to benefit from its construction, by which the texts are supposed to be strictly similar, leading us to expect that exploratory results from text mining will be similar too. We decided to explore a parallel dataset from a domain to extract knowledge from a technical area (e.g., finance). The choice of the pair Chinese–English has several motivations: firstly, the data is more easily available; secondly, there is a demand for English and Chinese tools and datasets, as English is already the lingua franca in many areas (political, economical, cultural, and scientific), and we also see an increasing interest in Chinese, which is now being taught at schools in western countries. One can keep in mind 1.41 billion people speak Chinese as their first or second language, while this is 1.35 billion for English (the overlap is no more than 20%). Secondly, China and the USA, as the areas of the native speakers, are drivers for the world economy. The language of business and finance has always attracted interest, since the movement of stock indexes can be an indicator, a 'barometer,' of the general trend in the economy. When we look at the availability of domain-specific parallel corpora, the majority of them are constructed around the following drivers: biomedicine (Neves, Yepes, & Névéol, 2016), digital humanities/culture (Christodoulopoulos & Steedman, 2014), city, transport (Lefever, Macken, & Hoste, 2009), food, the environment (Xiong, 2013), ICT (Labaka, Alegria, & Sarasola, 2016), digital humanities/law, and governance (Steinberger et al., 2006). Concerning Chinese–English, Chang (2004) from Peking University made one of the first large scale Chinese–English parallel corpora from HTML files with alignments at the paragraph and sentence levels, leading to a size of 10 million Chinese characters about different genres (news, technical articles, subtitles). Concerning the domain of finance, there are some small corpora for different pairs of languages, but not Chinese–English, (Arcan, Thomas, de Brandt, & Buitelaar, 2013; Bick & Barreiro, 2015; Smirnova & Rackevičienė, 2020; Tiedemann, 2012; Volk, Amrhein, Aepli, Müller, & Ströbel, 2016). The largest one is the SEDAR dataset,[1] containing 8.6 million French–English sentence pairs in the finance domain from PDF files of the regulations of the province of Quebec (Ghaddar & Langlais, 2020). To our knowledge, the dataset discussed in our article represents new available material for the community. The question we address is to consider the state of the art techniques and the main contemporary approaches to text mining, and see what finally we can extract from a dataset of news in a specialized domain such as fintech. Knowing that each news item contains the same version in Chinese and English, another question to explore is the following:"are the efficiency and extraction exactly the same or do some cultural aspects influence the translation and so the lexical and semantic content? In this way, the general dataset we present in this article can be seen as a gold standard for the output of calibrated measures for all kinds of techniques. In general, studies use text collection within the framework of a specific method such as disinformation analysis (Turenne, 2018) or the development of medical drugs (Kolchinsky, Lourenco, Wu, & Rocha, 2015), or for a specific task such as part of speech (POS) tagging (Akbik, Blythe, & Vollgraf, 2018) or named entity extraction (Chiu & Nichols, 2016). In this article, we also take a domain dataset (namely, fintech) and a specific genre of document (news), but we do not have a specific task to improve. We try easy tasks intuitively and directly usable on such a dataset: clustering (named entity and word), classification (topic and sentiment), and pattern extraction (word life and citation). We made the dataset using the *Financial Times* website from which we grabbed 60,473 news items from between 2007 and 2021, each containing a version in English and its translation into Chinese. We focus on three families of techniques within the text-mining framework: (i) pre-processing techniques; (ii) supervised approaches involving deep learning techniques such as LSTM, BERT, CNN and also SVM, naïve Bayes, and random forest; and (iii) unsupervised techniques involving *k*-means, community detection, biclustering, co-cord analysis, and topic modeling (Turenne et al., 2020). This paper is divided into the following sections: we discuss the dataset and its sub-datasets, describe the state-of-the-art research based on bilingual corpora, machine learning, and natural language processing, and then present the results of our experiments.

1 *https://github.com/autorite/sedar-bitext* (last accessed: 01.03.2022).

2 RELATED WORK

2.1 PARALLEL LANGUAGE DATASET BUILDING

Zhao and Vogel (2002) is probably one of the pioneering studies about combining a parallel Chinese–English dataset and mining approach. They used 10 years of the Xinhua bilingual news collection, but that is not available. Koehn (2005) is a large-scale document multilingual and parallel dataset containing ~60 million words on average per language for 21 European languages, but nothing about Chinese. In the same way, we find a topic detection and tracking repository.[2] It contains 30K in Chinese and English, but not in parallel. Christodoulopoulos and Steedman (2014) and Sturgeon (2021) are open data repositories and digital humanities projects. They contain books with English–Chinese versions but their content is closely related to philosophy, religion, and difficult-to-understand contemporary thinking: for example, manual annotation for classification is not easy. The UCI Machine Learning Repository (Dua & Graff, 2017) and Kaggle[3] are repositories of datasets, and many of them are used for the evaluation of algorithms. There are no English–Chinese parallel corpora. Zhai, Liu, Zhong, Illouz, and Vilnat (2020) made a dataset considering 11 genres (constructed based on existing work: art, literature, law, material for education, microblogs, news, official documents, spoken, subtitles, science, and scientific articles) and made a parallel English–Chinese dataset with 2,200 sentences to test the translation of literals. Tian et al. (2014) presents UM-Corpus,[4] designed for sentence machine translation (SMT) research. It contains 15 million English–Chinese parallel sentences and treats eight genres: News, Spoken, Laws, Theses, Educational Materials, Science, Speech/Subtitles, and Microblog. Globally, the dataset contains 2.2 million sentences in both languages (450,000 for news alone). This dataset is freely available but named entities are anonymized.

2.2 BUILDING DOMAIN-SPECIFIC PARALLEL DATASETS

In this section we present an extensive literature review of domain-specific datasets, their language pairs, and topics. We observed an increased interest in domain-specific parallel datasets in the past year. The main use of such material is to make a specialized learning dataset to improve a statistical machine translation system and to do cross-lingual information retrieval (McEnery & Xiao, 2007) from a computational point of view, to extract automatically or semi-automatically a specialized lexicon in different languages (Rosemeyer & Enrique-Arias, 2016) from a linguistic point of view. In the following review, we consider as domain-specific a dataset focused on all aspects of one topic. A text genre, such as news or technical publications, is considered as a domain.

2.2.1 Digital Humanities: culture

In this domain we have found 20 datasets, of which large pair datasets are as follows. In the area of religious studies, Christodoulopoulos and Steedman (2014) is about the Bible in 100 languages. We also find the Chinese–English (Sturgeon, 2021), the Arabic–English (Hamoud & Atwell, 2017), a presentation of the same ancient religious texts in different Germanic dialects (Dipper & Schultz-Balluff, 2013), and a parallel dataset of English and Persian religious texts (Beikian & Borzoufard, 2016). In literary studies Fraisse, Tran, Jenn, Paroubek, and Fishkin (2018) created a massively parallel dataset of translated American literary texts, with 23 languages. Altammami, Atwell, and Alsalka (2020) present a bilingual parallel English–Arabic dataset of narratives reporting different aspects of Muhammad's life. In the domain of tourism and traveling, Espla-Gomis et al. (2014) built a domain-specific English——Croatian parallel dataset from different websites, Ponay and Cheng (2015) made an English–Tagalog dataset, Bureros, Tabaranza, and Roxas (2015) created a English–Cebuano dataset, Woldeyohannis, Besacier, and Meshesha (2018) made an Amharic–English dataset, Srivastava and Sanyal (2015) made a small parallel English–Hindi dataset, and Boldrini and Ferrández (2009) got 4500 questions/answers from customers about tourism in Spanish translated into English. About literary texts, Rovenchak (2021) published a Bamana–French analysis concerning Bamana tales, Kenny (1999) describes GEPCOLT, an electronic collection of some fourteen works of contemporary German-language fiction alonside their translations into English,

2 *http://projects.ldc.upenn.edu/TDT3-TDT4* (last accessed: 01.03.2022).

3 *https://www.kaggle.com/datasets* (last accessed: 01.03.2022).

4 *http://nlp2ct.cis.umac.mo/um-corpus/index.html* (last accessed: 01.03.2022).

Giouli, Glaros, Simov, and Osenova (2009) made a Greek–Bulgarian dataset about cultural, literary and folk texts, Kashefi (2020) made a Persian–English dataset with masterpieces of literature, Frankenberg-Garcia (2009) built a parallel dataset of English and Portuguese literary texts, Miletic, Stosic, and Marjanović (2017) made Paracolab a dataset of English, French and Serbian literary books, Guzman (2013) describes a dataset of literary texts with versions in Spanish, French, German, and Catalan.

2.2.2 Finance

D.-Y. Lee (2011) used an interesting approach, for Korean and English, to improve financial phrase translation, but the corpora are comparable without being really parallel. There are some parallel corpora about finance, with a limited size, such as Smirnova and Rackevičienė (2020), who made a dataset of European documents in English translated to French and Lithuanian related to finance, but the size is relatively small, consisting of 154 documents from 2010 to 2014. Bick and Barreiro (2015) made a Portuguese–English parallel dataset of about 40,000 sentences in the Legal-Financial domain, coming from a company translation memory. We will next mention four notable parallel corpora about finance, for which we will give the details below: the ECB dataset,[5] the DBpedia-linguee dataset, the CSB dataset,[6] and the SEDAR dataset.[1] All of them have been made for automatic translation and cross-lingual information retrieval purposes. In the Opus project (Tiedemann, 2012), we can find the ECB dataset, employing 19 European languages and concerning financial and legal newsletters from the European Central Bank. As an example, it contains 113,000 English–German pairs of sentences. Arcan et al. (2013) used DBpedia datasets to extract the titles of relevant Wikipedia articles, and the Linguee database, obtaining 193,000 aligned sentences (English–German, English–French, and English–Spanish) to find translations of financial terms. The Credit Suisse Bulletin dataset (CSB) is based on the world's oldest banking magazine, published by Credit Suisse since 1895 in both German and French (Volk et al., 2016). The SEDAR dataset (i.e., the System for Electronic Document Analysis and Retrieval) contains 8.6 million French–English sentence pairs in the finance domain from PDF files of regulations of the province of Quebec (Ghaddar & Langlais, 2020). However, all these datasets are about pairs of European languages. Guo (2016) describes how it can be feasible to make a domain-specific Chinese–English parallel dataset in the financial service domain, but it is restricted to giving guidelines about which tool to use to get raw data and how to use a parallel dataset, with the description and availability of the dataset. We have seen in this review that, firstly, domain-specific datasets are for different topics of societal challenges. Secondly, although the finance domain is not lacking in datasets, English–Chinese is not covered yet.

2.3 PARALLEL LANGUAGE DATASET EXPLORATION

Parallel corpora have been investigated to make alignments between sentences. Wu and Xia (1994) is a pioneering work using parallel sentences in the framework of automatic translation. They used literal translations of sentences from the parliamentary proceedings of the Hong Kong Legislative Council, with five million words, to predict the Chinese translation of each English entry. In Yang and Li (2003), an alignment method is presented at different levels (title, word, and character) based on dynamic programming (DP). Lu, Tsou, Jiang, Kwong, and Zhu (2010) used a non-open dataset of 157,000 files, with both Chinese and English versions. More recently, Schwenk, Chaudhary, Sun, Gong, and Guzmán (2021) have made an alignment process over 85 languages and 135 million sentences from Wikipedia (available as open data), but they found only 790 sentences for English–Chinese, which is very few for a text mining workflow. Li, Wang, Huang, and Zhao (2011) used a linear combination and minimum sample risk (MSR) algorithm to make a matching between named entities (Person, Organization) and obtained an *F*-score of 84%. A pioneering work in text mining and English–Chinese texts is probably C.-H. Lee and Yang (2000), who used a neural network clustering method called Self-Organizing maps to extract clusters from an English–Chinese parallel dataset (this parallel dataset is made with Sinorama magazine articles with 50,000 sentences)[7] but their conclusion only reveals the potential of the

5 *https://www.ecb.europa.eu/press/key/html/downloads.en.html* (last accessed: 01.03.2022).

6 *http://csb.access.ch* (last accessed: 01.03.2022).

7 *https://www.taiwan-panorama.com/en/Home/About* (last accessed: 01.03.2022).

approach. Lan and Huang (2017) construct a bilingual English–Chinese latent semantic space and also select *k*-means initial cluster centers, but the interpretation of the clustering is not very clear.

3 THE DATASET

3.1 DATA COLLECTION

We extracted news from the Financial Times and FT Chinese, both freely available news located at the financial times website.[8,9,10]

The news was collected for the period **from 2007 to 2021**.

After collating the links, the pages were downloaded with 'wget' and stripped of HTML. The encoding of the files was normalized to UTF-8 (R package 'httpr'). Cloud computing under the SLURM framework was used to parallelize the NLP preprocessing.

In all, we got an uncleaned raw **text dataset with 90,003 documents**.

3.2 DATA PREPROCESSING

We carried out sentence segmentation, word splitting, and named entity extraction. For linguistic preprocessing, we used regular expressions for field extraction, sentence and paragraph splitting. We used Jieba and spaCy algorithms for tokenization and tagging, and the Stanford NER framework for named entity extraction.

The use of HTML was helpful to automatically extract from each news item its timestamp, title (in both languages), text body (in both languages), and topic tags. But in some cases, a translation was not available, so we took it as is. We tried to carry out a paragraph alignment between two equivalent documents in Chinese and English. Splitting into paragraph is also quite easy using line break markers. However, in some cases the number of paragraphs does not match, and we did not achieve this alignment because of the expensiveness of a human validation.

We proceeded to clean the documents using two rules: (1) each one had to have both English and Chinese versions; (2) only files with a text body containing more than two characters were kept.

We got a cleaned raw **text dataset of 60,473 documents**.

The dataset is available at *https://doi.org/10.5281/zenodo.5591908*

3.3 DATA STATISTICS

LANG.	TOKEN	NP	MULTIWD	PARAG.S	SENT.	NE	HANZI
English	2,598,309	1,672,577	2,376,424	272,756	597,372	1,190,682	0
Chinese	7,480,139	1,491,790	3,466,453	258,213	572,185	1,268,674	21,679,815

Table 1 Linguistic features of the text collection ('Lang.' is language, 'NP' is noun phrases, 'MultiWD' is multiwords, 'Sent.' is sentences, 'NE' is named entities, 'Hanzi' is Chinese characters.

The dataset contains various metadata, such as title and text body both in English and Chinese, the time of publication, and some topic tags. ***Table 1*** shows the extraction of elementary linguistic features.

3.4 CATEGORIES OF FINANCE DOMAIN

We made different samples for topic prediction using classification methods. This is the list of the 10 topic-metadata tags contained in the documents, used by the *Financial Times* to annotate the area of each news item. A news item can contain several tags: book, business, culture, economy, lifestyle, management, markets, people, politics, or society. There were

8 *https://www.ft.com/* (last accessed: 01.03.2022).

9 *https://www.ftchinese.com/* (last accessed: 01.03.2022).

10 This is an example of a parallel archived news link: *http://www.ftchinese.com/story/001015037/ce?archive* (last accessed: 01.03.2022).

57,584 documents containing topic metadata. This is the list of the 10 tags from the *Financial Times* websites about the economic sector we used for manual annotation: technology, consumer services, health care, consumer goods, basic materials, industrials, oil & gas, and telecommunications. There are 2,993 documents that were tagged manually.

The top influential media in Finance are: 1. *The Wall Street Journal*. 2. Bloomberg. 3. *The New York Times*. 4. The *Financial Times*. 5. CNBC. 6. Reuters. 7. *The Economist*.

Five items of the Financial Times website can be clearly identified as related to the "economy" (equities, currencies, commodities, bonds, funds & ETFS) and the item world market can be associated with "markets," company as "business," and director dealings as management. The economy, management, markets, and business are among the tags contained in each document as metadata. However, we also find other tags, such as lifestyle, politics, and people. In fact, many influential people have an impact on the evolution of markets.

Other items as sectors and industrials can be further split into:

id01 – Technology (Software & Computer Services, Technology Hardware & Equipment)
id02 – Consumer Services (General Retailers, Travel & Leisure, Food & Drug Retailers, Media)
id03 – Health Care (Health Care Equipment & Services, Pharmaceuticals & Biotechnology)
id04 – Consumer Goods (Automobiles & Parts, Leisure Goods, Personal Goods, Food Producers, Household Goods, Tobacco, Beverages)
id05 – Basic Materials (Industrial Metals, Mining, Chemicals)
id06 – Industrials (Support Services, Electronic & Electrical Equipment, Industrial Transportation, Aerospace & Defense, Construction)
id07 – Financials (Real Estate Investment & Services, Financial Services, General Financial, Life Insurance, Banks, Nonlife Insurance)
id08 – Oil & Gas (Alternative Energy, Oil & Gas Producers, Oil Equipment, Services & Distribution)
id09 – Utilities (Gas, Water & Multi-utilities, Electricity)
id10 – Telecommunications (Fixed Line Telecommunications, Mobile Telecommunications)

Sectors, in finance, act both as a guide to make promising investments in the right places and as representation of areas of activity.

Topics id01 to id10 are used for manual annotation so their representation is less important than topics inserted into each document as metadata. From the manual annotation, the most frequent topics are: financials, consumer goods, consumer services, and technology. From the metadata, the most frequent topics are: business, the economy, markets, management, politics, lifestyle, and society.

3.5 MANUAL ANNOTATION

To carry out the manual annotation, we made a set of document batches, each one containing 100 distinct documents. A population of 31 students (year-3 level in computer science, with B1 to C1 level of English) received one batch each. Multiple annotation was possible, and the format of the annotation was quite elementary, such as document id followed by class id, one annotation by line, e.g.:

1014550; id07
1014871; id11

An extra annotator assessed the annotations by choosing randomly 10 files for each batch. If the annotation done by the extra annotator showed more than four differences from those produced by the annotator (i.e., >40% disagreement), the batch had to be revised by the annotator. Nineteen batches were revised. Finally, after the second round, we compiled all the batches together.

3.6 DATA USAGE

As mentioned in the previous section on the literature, there are several ways to use a parallel dataset. The same is true for our Chinese–English parallel dataset for the domain of finance. So here are five main key points as possible usage:

- The influence of the language on the knowledge discovery
  We present the results of different clusterings for topic discovery and classification for topic detection. Here, the algorithm is not supposed to take into account specificities of the language (i.e., it is to be language-independent). This dataset can be useful to study how a language-dependent algorithm could be more efficient.
- Keyword in context
  Concordances of a word in the domain of finance can be extracted. In such a usage case, different contexts make possible the study of the meaning of a phrase and its variation.
- Automatic translation
  A classical usage case is to exploit such a dataset to make automatic translations of documents in the domain of finance, using this dataset as a training set for a statistical machine translation system (SMT)
- Neologism translation
  Translation is always a challenge, especially for new words. A usage case of the dataset is the study of neologisms. For example, to find the Chinese equivalent to about a new named entity in English (company name, people name).
- Time series of a domain-specific word
  The last case can be the study of the distribution of words or phrases over time and see their popularity.

4 DISCOVERY OF SOME FREQUENT INTERESTING TERMS

In this section, we will search for some interesting words or phrases in the dataset and count their frequency of occurrence, which will be conducive to our further understanding of the dataset. Next, this section will be divided into three parts to explore the frequency of English proverbs and Chinese idioms, important finance related terms, and globally famous companies in the dataset. We made some experiments about lexical variation over time and proverb analysis (see appendix A for more details).

4.1 DISCOVERY OF FREQUENT TERMS OF FINANCE DOMAIN

The first step is deciding how to choose some commonly used financial terms. Our decision was to use Fundera. Fundera is an online marketplace that connects small business owners with the best providers of capital for their businesses. It offers product marketplaces that cover everything from loans to legal services, free financial content, and one-on-one access with experienced lending experts. Based on the founding editor and vice president of the Fundera Ledger Meredith Wood's "60 Business and finance terms you should definitely know",[11] we selected the top 20 financial terms that appear most frequently in the dataset. The results are shown on ***Table 2***.

Table 2 20 most frequently used financial terms.

Capital	9383	Net Worth	195
Asset	3086	Liability	141
Liquidity	1704	Business Plan	126
Interest Rate	1036	Fixed Asset	101
Bankruptcy	616	Debt Financing	97
Balance Sheet	522	Working Capital	83
Principle	382	Financial Statements	72
Collateral	371	Equity Financing	64
Depreciation	368	Line of Credit	46
Cash Flow	209	Appraisal	42

11 *https://www.fundera.com/blog/business-finance-terms-and-definitions* (last accessed: 01.03.2022).

Next, we imitated the method used above to detect the most frequent idioms and proverbs, extracting the statements in the dataset and calculating the frequency of occurrence of each financial term (see appendix file).

4.2 DISCOVERY OF FREQUENT COMPANY NAMES

We used the same method to collect statistics on the frequency of occurrence of company names in the dataset. Among them, we find the Chinese company Huawei, which shows that with the increase of China's international influence, Chinese technology enterprises are increasingly favored by global business people.

5 TEXT-MINING APPROACHES AND THE DOMAIN OF FINANCE

The first point for people interested in finance or natural language processing about such a dataset as this, is that we provide a full analysis taking into account state of the art text-mining technology. These experiments were of three kinds (see appendix B and appendix C for technical details):

(1) lexical extraction (words, noun phrases, names of people, names of companies)
(2) classification (rervised learning)
(3) clustering (unsupervised learning)

As we showed in the section on the discovery of lexical items, this dataset is useful for identifying the important concepts and actors of the domain. These concepts are not new for an expert working in finance everyday, but the dataset can be used as an educational tool for students at school or college to understand what is finance through real life events and practical information. A list of frequent noun phrases (such as 'asset,' 'interest rate'), a list of famous and influential people (such as Elon Musk, Xi Jinping), a list of names of famous organizations (such as the IMF and the Fed) were extracted, and one hundred frequent items for these three categories can easily serve as a basic framework of concepts for educational purposes. We also studied and compared the properties of the English and Chinese languages through the use of proverbs, which is one of the high-level linguistic patterns of any language. We discovered that in the domain of finance, which is highly related to technology and also to society, in the Chinese language, people used more freely proverbs but not at all in English. We do not have an explanation for this except that it may be an important cultural difference in how people use language to disseminate information (even in a technological area). We have shown that using this classification technique some potential readers could process new documents (unseen from the dataset), which may be interesting for them, according to the ontology of 20 topics described in Section 3.4.

Clustering, by definition, relies mainly on organizing knowledge about a set of unstructured data. We have carried out several experiments and clustering has revealed some classical topics of finance, such as business or markets, but also surprising topics in the finance domain, such as lifestyle, art and life, politics, and British education, which seem to play a big role. This shows that finance is not just an activity in society, like sports for example, but also seems to be an ideological model. Secondly, the clusters show that even if finance is globalized, a polarity about the specific relationship between China and US appears to emerge as more important than all others.

6 CONCLUSION

Chinese and English is an interesting combination of languages for testing algorithms and mining. Finance is a hot area of activity in our contemporary world. We made a text dataset using the *Financial Times* website from which we grabbed 60,473 news items from between 2007 and 2021. This dataset is a bilingual Chinese–English parallel dataset of news in the domain of finance, and is open access. We used a text mining analytical framework. As a future perspective, our dataset can be used to infer the translation of new terms from English to Chinese (i.e., company names), to extract the distribution of occurrences of new concepts for time series analysis (i.e., neologisms) or to apply a more innovative clustering approach to discover new concepts (i.e., ontology learning).

ADDITIONAL FILES

The additional files for this article can be found as follows:

- **Appendix A.** Discovery of some frequent interesting terms. DOI: *https://doi.org/10.5334/johd.62.s1*
- **Appendix B.** Classification. DOI: *https://doi.org/10.5334/johd.62.s2*
- **Appendix C.** Clustering. DOI: *https://doi.org/10.5334/johd.62.s3*

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Nicolas Turenne: Conceptualisation and writing original draft

Ziwei Chen: Methodology, Classification section

Jianlong Li: Methodology, Classification section

Guitao Fan: Methodology, lexical analysis and pattern section

Jiaqi Zhou: Methodology, lexical analysis and pattern section

Siyuan Wang: Methodology, clustering section

Yiwen Li: Methodology, clustering section

AUTHOR AFFILIATIONS

Nicolas Turenne *orcid.org/0000-0003-1229-5590*
BNU-HKBU United International College, UIC, Division of Science and Technology, Zhuhai Guangdong, China

Ziwei Chen
BNU-HKBU United International College, UIC, Division of Science and Technology, Zhuhai Guangdong, China

Guitao Fan
BNU-HKBU United International College, UIC, Division of Science and Technology, Zhuhai Guangdong, China

Jianlong Li
BNU-HKBU United International College, UIC, Division of Science and Technology, Zhuhai Guangdong, China

Yiwen Li
BNU-HKBU United International College, UIC, Division of Science and Technology, Zhuhai Guangdong, China

Siyuan Wang
BNU-HKBU United International College, UIC, Division of Science and Technology, Zhuhai Guangdong, China

Jiaqi Zhou
BNU-HKBU United International College, UIC, Division of Science and Technology, Zhuhai Guangdong, China

REFERENCES

Akbik, A., Blythe, D., & **Vollgraf, R.** (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics (coling)* (pp. 1638–1649). New Mexico: Paparazzi Press. Retrieved from *https://aclanthology.org/C18-1139*

Altammami, S., Atwell, E., & **Alsalka, A.** (2020). The Arabic-English parallel corpus of authentic hadith. *International Journal on Islamic Applications in Computer Science And Technology*, *8*(2). DOI: *http://www.sign-ific-ance.co.uk/index.php/IJASAT/article/view/2199*

Arcan, M., Thomas, S. M., de Brandt, D., & **Buitelaar, P.** (2013). Translating the FINREP taxonomy using a domain-specific corpus. In *Proceedings of Chinese translation summit XIV*. Nice, France. Retrieved from *https://aclanthology.org/2013.mtsummit-posters.1.pdf*

Beikian, A., & **Borzoufard, M.** (2016). *Mizan: A large persian-english parallel corpus.* Retrieved from *https://cdn.ketabchi.com/products/175402/pdfs/ketab-general-book-sample-wybml.pdf*

Bick, E., & **Barreiro, A.** (2015). Automatic anonymisation of a new portuguese-english parallel corpus in the legal-financial domain. *Oslo Studies in Language, 7*(1), 101–124. Retrieved from *https://journals.uio.no/index.php/osla/article/view/1460/1357*. DOI: *https://doi.org/10.5617/osla.1460*

Boldrini, E., & **Ferrández, S.** (2009, March 1–7). A parallel corpus labeled using open and restricted domain ontologies. In *Proceedings of 10th international conference CICLing.* Mexico City, Mexico. DOI: *https://doi.org/10.1007/978-3-642-00382-0_28*

Bureros, L. L., Tabaranza, Z. L. B., & **Roxas, R. R.** (2015). Building an English-Cebuano tourism parallel corpus and a named-entity list from the Web. In *Proceedings of workshop on computation: Theory and practice* (pp. 158–169). DOI: *https://doi.org/10.1142/9789813202818_0012*

Chang, B. (2004). Chinese-English parallel corpus construction and its application. In *Proceedings of the PACLIC* (pp. 201–204). Tokyo: Waseda University, Dec. 8–10. Retrieved from *https://aclanthology.org/Y04-1030.pdf*

Chiu, J. P. C., & **Nichols, E.** (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* (pp. 357–370). DOI: *https://doi.org/10.1162/tacl_a_00104*

Christodoulopoulos, C., & **Steedman, M.** (2014). *The Bible in 100 Languages*. Retrieved from *https://github.com/christos-c/bible-corpus*

Dipper, S., & **Schultz-Balluff, S.** (2013). The Anselm Corpus: Methods and perspectives of a parallel aligned corpus. In *Proceedings of the workshop on computational historical linguistics at NODALIDA. NEALT* (pp. 27–42). Retrieved from *https://ep.liu.se/ecp/087/ecp13087.pdf#page=35*

Dua, D., & **Graff, C.** (2017). *UCI machine learning repository*. Retrieved from *http://archive.ics.uci.edu/ml*

Espla-Gomis, M., Klubička, F., Ljubešić, N., Ortiz-Rojas, S., Papavassiliou, V., & **Prokopidis, P.** (2014). Comparing two acquisition systems for automatically building an English-Croatian parallel corpus from multilingual websites. In *Proceedings of the ninth international conference on language resources and evaluation* (pp. 1252–1258). European Language Resources Association (ELRA). Retrieved from *http://www.lrec-conf.org/proceedings/lrec2014/pdf/529 Paper.pdf*

Fraisse, A., Tran, Q.-T., Jenn, R., Paroubek, P., & **Fishkin, S.** (2018, May). TransLiTex: A parallel corpus of translated literary texts. In *Proceedings of the eleventh international conference on language resources and evaluation* (pp. 201–204). Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from *https://hal.archives-ouvertes.fr/hal-01827884/file/11 W34.pdf*

Frankenberg-Garcia, A. (2009). Compiling and using a parallel corpus for research in translation. *Babel: International journal of translation, 21*(1), 57–71. Retrieved from *https://openresearch.surrey.ac.uk/esploro/outputs/journalArticle/Compiling-and-using-a-parallel-corpus-for-research-in-translation/995416816302346#file-0*

Ghaddar, A., & **Langlais, P.** (2020). Sedar: a large scale French-english financial domain parallel corpus. In *Proceedings of the language resources and evaluation conference* (pp. 3595–3602). Marseille, France: European Language Resources Association. Retrieved from *https://aclanthology.org/2020.lrec-1.442*

Giouli, V., Glaros, N., Simov, K., & **Osenova, P.** (2009). A web-enabled and speech-enhanced parallel corpus of Greek-Bulgarian cultural texts. In *Proceedings of the of the EACL workshop on language technology and resources for cultural heritage, social sciences, humanities, and education* (pp. 35–42). Athens, Greece: Association for Computational Linguistics. Retrieved from *https://aclanthology.org/W09-0305.pdf*. DOI: *https://doi.org/10.3115/1642049.1642054*

Guo, X. (2016, November 17–18). Drawing a route map of making a small domain-specific parallel corpus for translators and beyond. In *Proceedings of translating and the computer* (pp. 88–99). London, UK. Retrieved from *https://aclanthology.org/2016.tc-1.9.pdf*

Guzman, J. R. (2013). El corpus COVALT i l'eina d'alineament de frases Alfra-COVALT. In L. Bracho Lapiedra (Ed.), *El corpus COVALT: un observatori de fraseologia traduïda* (pp. 49–60). Aachen: Shaker.

Hamoud, B., & **Atwell, E.** (2017). Evaluation corpus for restricted-domain question-answering systems for the holy Quran. *International Journal of Science and Research, 6*(8), 1133–1138. Retrieved from *https://eprints.whiterose.ac.uk/125920/*

Kashefi, O. (2020). *MIZAN: A large Persian-English parallel corpus*. Retrieved from *https://arxiv.org/pdf/1801.02107v3.pdf*

Kenny, D. (1999). The German-English parallel corpus of literary texts (GEPCOLT): A resource for translation scholars. *Teanga, 1*, 25–42.

Koehn, P. (2005). *Europarl*. Retrieved from *http://www.statmt.org/europarl/*

Kolchinsky, A., Lourenco, A., Wu, H.-Y., & **Rocha, L. M.** (2015). Extraction of pharmacokinetic evidence of drug-drug interactions from the literature. *PLOS ONE*. DOI: *https://doi.org/10.1371/journal.pone.0122199*

Labaka, G., Alegria, I., & **Sarasola, K.** (2016). Domain adaptation in MT using Wikipedia as a parallel corpus: Resources and evaluation. In *Proceedings of the tenth international conference on language resources and evaluation* (pp. 2209–2213). Portoroz, Slovenia: European Language Resources Association (ELRA).

Lan, H., & **Huang, J.** (2017, February). Chinese-English cross-language text clustering algorithm based on latent semantic analysis. In *Proceedings of information science and cloud computing* (pp. 1–7). Retrieved from *https://pos.sissa.it/300/007/pdf*

Lee, C.-H., & **Yang, H.-C.** (2000). Towards multilingual information discovery through a SOM based text mining approach. In *PRICAI workshop on text and web mining* (pp. 80–87). Melbourne, Australia. Retrieved from *https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.8800&rep=rep1&type=pdf*

Lee, D.-Y. (2011). A corpus-based translation of Korean financial reports into English. *Journal of Universal Language, 12*(1), 75–94. Retrieved from *https://www.sejongjul.org/download/downloadpdf?pid=jul-12-1-75*. DOI: *https://doi.org/10.22425/jul.2011.12.1.75*

Lefever, E., Macken, L., & **Hoste, V.** (2009, 30 March – 3 April). Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th conference of the European Chapter of the ACL* (pp. 1746–1751). Athens, Greece. Retrieved from *https://aclanthology.org/E09-1057.pdf*. DOI: *https://doi.org/10.3115/1609067.1609122*

Li, L., Wang, P., Huang, D., & **Zhao, L.** (2011). Mining English-Chinese named entity pairs from comparable corpora. *ACM Transactions on Asian Language Information Processing, 10*, 1–19. DOI: *https://doi.org/10.1145/2025384.2025387*

Lu, B., Tsou, B. K., Jiang, T., Kwong, O. Y., & **Zhu, J.** (2010). Mining large-scale parallel corpora from multilingual patents: An English-Chinese example and its application to SMT. In *Proceedings of the 1st CIPS-SIGHAN joint conference on Chinese language processing* (pp. 79–86). Beijing. Retrieved from *https://aclanthology.org/W10-4110.pdf*

McEnery, T., & **Xiao, Z.** (2007). Parallel and comparable corpora – the state of play. In N. T. Y. Kawaguchi T. Takagaki & Y. Tsuruga (Eds.), *Proceedings of the international conference on Asian language processing* (pp. 131–146). Amsterdam: Benjamin. DOI: *https://doi.org/10.1075/ubli.6.11mce*

Miletic, A., Stosic, D., & **Marjanović, D.** (2017). ParCoLab: A Parallel Corpus for Serbian, French and English. In K. Ekštein & V. Matoušek (Eds.), *Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science*, *10415*, 201–204. Berlin: Springer-Verlag. DOI: *https://doi.org/10.1007/978-3-319-64206-2*

Neves, M., Yepes, A. J., & **Névéol, A.** (2016). The Scielo Corpus: A parallel corpus of scientific publications for biomedicine. In *Proceedings of the 15th international conference on language resources and evaluation*. European Language Resources Association. Retrieved from *https://aclanthology.org/L16-1470*

Ponay, C. S., & **Cheng, C. K.** (2015). Building an English-Filipino tourism corpus and lexicon for an ASEAN language translation system. In *Proceedings of the international conference ASIALEX* (pp. 201–204). Hong Kong: Polytechnic University. Retrieved from *https://www.researchgate.net/profile/Charmaine-Ponay-2/publication/27994689223BuildinganEnglish-FilipinoTourismCorpusandLexiconforanASEANLanguageTranslationSystem/links/559f2fee08ae97223ddc602f/23-Building-an-English-Filipino-Tourism-Corpus-and-Lexicon-for-an-ASEAN-Language-Translation-System.pdf*

Rosemeyer, M., & **Enrique-Arias, A.** (2016). A match made in heaven: Using parallel corpora and multinomial logistic regression to analyze the expression of possession in Old Spanish. *Language Variation and Change*, *28*(03), 307–334. DOI: *https://doi.org/10.1017/S0954394516000120*

Rovenchak, A. (2021). Bamana tales recorded by Umaru Nanankr Jara: A comparative study based on a Bamana-French parallel corpus. *Mandenkan*, *64*, 81–104. DOI: *https://doi.org/10.4000/mandenkan.2471*

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & **Guzmán, F.** (2021, April). WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: Main volume* (pp. 1351–1361). Online: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/2021.eacl-main.115*. DOI: *https://doi.org/10.18653/v1/2021.eacl-main.115*

Smirnova, O., & **Rackevičienė, S.** (2020). *English-French-Lithuanian parallel corpus of EU financial documents*. Retrieved from *http://hdl.handle.net/20.500.11821/35*

Srivastava, J., & **Sanyal, S.** (2015). POS-based word alignment for small corpus. In *Proceedings of international conference on Asian language processing* (pp. 37–40). DOI: *https://doi.org/10.1109/IALP.2015.7451526*

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & **Varga, D.** (2006, 24–26 May). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th international conference on language resources and evaluation* (pp. 2142–2147). Genoa, Italy. Retrieved from *https://arxiv.org/abs/cs/0609058*

Sturgeon, D. (Ed.). (2021). *Ancient Chinese Books Datasets (Chinese Text Project)*. Retrieved from *https://ctext.org/daoism*

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., & **Yi, L.** (2014). UM-Corpus: A Large English-Chinese parallel corpus for statistical machine translation. In *LREC*. Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from *http://www.lrec-conf.org/proceedings/lrec2014/pdf/774Paper.pdf*

Tiedemann, J. (2012, May). Parallel data, tools and interfaces in OPUS. In N. Calzolari et al. (Eds.), *Proceedings of the eighth international conference on language resources and evaluation* (pp. 2214–2218). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.673.2874&rep=rep1&type=pdf*

Turenne, N. (2018, January). The rumour spectrum. *PLOS ONE, 13*(1), 1–27. DOI: *https://doi.org/10.1371/journal.pone.0189080*

Turenne, N., Xu, B., Li, X., Xu, X., Liu, H., & **Zhu, X.** (2020). Exploration of a balanced reference corpus with a wide variety of text mining tools. In *Proceedings of ACAI 2020: 2020 3rd international conference on algorithms, computing and artificial intelligence* (pp. 1–9). New Mexico, USA: ACM Digital Library. DOI: *https://doi.org/10.1145/3446132.3446192*

Volk, M., Amrhein, C., Aepli, N., Müller, M., & **Ströbel, P.** (2016). Building a parallel corpus on the world's oldest banking magazine. In *Proceedings of the 13th conference on natural language processing (konvens)* (pp. 288–296). DOI: *https://doi.org/10.5167/uzh-125746*

Woldeyohannis, M. M., Besacier, L., & **Meshesha, M.** (2018). A corpus for Amharic-English speech translation: The case of tourism domain. In F. Mekuria, E. Nigussie, W. Dargie, M. Edward & T. Tegegne (Eds.), *Proceedings of information and communication technology for development for Africa. ict4da 2017. Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (Vol. 244). DOI: *https://doi.org/10.1007/978-3-319-95153-9*

Wu, E., & **Xia, X.** (1994). Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the first conference of the association for machine translation in the Americas* (pp. 206–213). Retrieved from *https://aclanthology.org/1994.amta-1.26.pdf*

Xiong, W. (2013). The development of the malaysian hansard corpus: A corpus of parliamentary debates 1959–2020. *New Technology of Library and Information Service,* Vol. Issue (6): 36–41. DOI: *https://doi.org/10.11925/infotech.1003-3513.2013.06.06*

Yang, C. C., & **Li, K. W.** (2003). Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol., 54*, 730–742. Retrieved from *https://aclanthology.org/A00-1004.pdf*. DOI: *https://doi.org/10.1002/asi.10261*

Zhai, Y., Liu, L., Zhong, X., Illouz, G., & **Vilnat, A.** (2020, May). Building an English-Chinese parallel corpus annotated with sub-sentential translation techniques. In *Proceedings of the 12th language resources and evaluation conference* (pp. 4024–4033). Marseille, France: European Language Resources Association. Retrieved from *https://www.aclweb.org/anthology/2020.lrec-1.496*

Zhao, B., & **Vogel, S.** (2002). Adaptive parallel sentences mining from web bilingual news collection. In zz (Ed.), *Proceedings of the IEEE international conference on data mining* (pp. 745–748). Beijing. DOI: *https://doi.org/10.1109/ICDM.2002.1184044*

TO CITE THIS ARTICLE:
Turenne, N., Chen, Z., Fan, G., Li, J., Li, Y., Wang, S., & Zhou, J. (2022). Mining an English-Chinese parallel Dataset of Financial News. *Journal of Open Humanities Data*, 8: 9, pp. 1–12. DOI: *https://doi.org/10.5334/johd.62*

Published: 18 March 2022

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.