PROCEEDINGS OF SCIENCE

Towards A Semantic Tagger for Analysing Contents of Chinese Corporate Reports

Scott Piao[1]
School of Computing and Communications, Lancaster University
Lancaster LA1 4WA, United Kingdom
E-mail: s.piao@lancaster.ac.uk

Xiaopeng Hu[2]
China Center for Information Industry Development (CCID)
CCID Plaza, 66 Zizhuyuan Rd, Haidian District, Beijing, China
E-mail: huxp@ccidtrans.com

Paul Rayson
School of Computing and Communications, Lancaster University
Lancaster LA1 4WA. United Kingdom
E-mail: p.rayson@lancaster.ac.uk

In this paper, we report on an experiment in which we explore the feasibility of applying a semantic tagger for analysing the textual contents of Chinese corporate reports, focusing on the contents of corporate strategy. In recent years, Natural Language Processing (NLP) research has been giving increasing attention to automatic analysis of the textual contents of corporate reports using NLP approach on a large scale. We test the assumption that semantic annotation tools can be useful for such a purpose, and study the feasibility by testing a Chinese semantic tagger developed by UCREL, Lancaster University for extracting core Chinese terms and semantic concepts from Chinese corporate annual disclosures, focusing on three main USAS semantic categories, *Money & Commerce*, *Architecture & Buildings*, and *Science & Technology*, which we assume are closely relevant to corporate strategy description, and use these categories and tags to extract core strategy terms. Our study shows that, by carefully applying selected semantic categories, our semantic annotation tool is capable of extracting core Chinese terms which can facilitate further content analysis of Chinese corporate reports.[3]

ISCC 2015
18-19, December, 2015
Guangzhou, China

[1]Corresponding Author
[2]Speaker
[3]This study is supported by NSFC Dean's award 2016 of China, UCREL of Lancaster University, UK and The China Center for Information Industry Development (CCID), Beijing, China.

1. Introduction

Recent Natural Language Processing (NLP) research has been giving increasing attention to the automatic analysis of the textual contents of corporate business reports on a large scale, such as in Big Data context, in order to provide useful information for stakeholders such as investors, regulators and governments, who are concerned with the performance of commercial companies.

The importance of analysis of corporate reports has long been recognised in business and economic areas, and a substantial amount of research has been carried out in this regard. For example, Jones and Shoemaker analysed a collection of accounting narrative papers for the types of linguistic features and readability [1]. Rutherford analysed the UK Operating and Financial Review as a genre of accounting narrative, and applied a word-frequency based corpus linguistics approach to identify genre rules [2]. Beattie and Thomson investigated the issue of applying content analysis approach to the analysis of intellectual capital disclosures [3]. More recently, Brennan et al. investigated the issue of impression management in accounting communication through the analysis of accounting narratives [4], Davies et al. examined typology of text analysis approaches in corporate narrative reporting research [5], and Hoberg and Lewis applied linguistic analysis and statistical metrics to detect fraud corporate disclosures [6]. But most of the above mentioned studies are based on small sets of data or manual analysis. With increasing amounts of corporate reports become available, such analysis need to be automated in order to apply it on a large scale.

A recent new development is the application of the NLP approach to facilitate automatic content analysis of corporate business reports on a large scale. For example, El-Haj et al. developed techniques and a tool to automatically parse and annotate document structures of the UK corporate reports [7]. Young suggests that NLP techniques can be used to address various issues of corporate communications including 1) Detecting deception and fraudulent reporting; 2) Measuring sentiment (market and individual stock; 3) Opinion mining; 4) Measuring information content of corporate narratives, etc. [8]

So far, the majority of research on automatic textual content analysis of corporate reports has been conducted on English data. Similar research on Chinese data is still very limited. Our research aims to address this gap and to develop a semantic analysis tool for automatically analyzing corporate business reports written in Chinese language. Specifically, in this experiment we focus on the issue of automatic identification and extraction of core Chinese terms and concepts which can be used to search for business-strategy related information in a large collection of Chinese corporate business reports.

2. USAS semantic annotation system

In our study, we use a Chinese semantic tagger developed by UCREL, Lancaster University, UK to automatically identify core Chinese terms related to corporate strategy. It is a part of the Lancaster USAS semantic annotation software system which is based on a set of semantic lexicons and applies a set of coarse-grained word sense disambiguation rules [9]. This system employs a semantic classification scheme derived from Tom McArthur's Longman Lexicon of Contemporary English [10], as well as a set of tags for denoting the semantic fields

of the classification scheme. In detail, the USAS semantic scheme contains 21 main semantic fields which are denoted by 21 letters. They are further divided into 232 sub-fields, such as *I1.2* for "Money: Debts", *K5.1* for "Sports", *N5* for "Quantities" etc.[4] In addition to the formal tagset, it also employs a set of auxiliary codes, such as ***m/f*** (male/female), +/- (positive/negative) etc. to distinguish important attributes of semantic fields. For example, the antonyms "happy" and "sad" are tagged with "E4.1+" and "E4.1-", which indicate positive and negative sentiment respectively. Furthermore, it is designed to identify and tag multi-word expressions as single terms, such as phrasal verbs, noun phrases, named entities and non-compositional idioms, which is highly significant for identifying contextual meaning.

Originally developed for processing English textual data, it has been ported to a number of other languages, including Chinese [11]. The current version of Chinese semantic tagger is capable of annotating Chinese words using the USAS semantic tagset with a good lexical coverage. Although it is still under improvement, we hypothesise that it is suitable for our study, where we focus on only three main USAS semantic categories.

The Chinese semantic tagger software incorporates a Chinese word segmenter and a POS tagger developed in Stanford University (http://nlp.stanford.edu/software/tagger.shtml), based on which the Chinese semantic package carries out semantic annotation. It employs Chinese semantic lexicons derived from the English semantic lexicons via automatic translation using a Chinese-English bilingual dictionary and a corpus-derived bilingual lexicon [11].

3. Experiment of corporate strategy related core term extraction

As mentioned previously, we aim at examining the feasibility of identifying core Chinese terms in the Chinese corporate reports that are related to the description of corporate business strategy. A collection of such terms can help to generate macro summaries regarding the corporate strategies reflected by their business reports as well as provide data searching points for further analysis. Obviously not all semantic fields are related to the business strategy information. Therefore, for our experiment, we chose three main USAS semantic fields which we consider are closely relevant to the strategic contents in the corporate reports. The three selected fields (denoted by three letters) include:

1) *I – Money & Commerce*,
2) *H - Architecture, Buildings, Houses & Home*,
3) *Y - Science & Technology.*

In the USAS semantic annotation scheme, the above three categories are further divided into eighteen sub-categories (tags), as shown in Table 1 below. As indicated by their definitions, the categories under the *I* major category are closely related to financial and commercial entities and activities. In addition, those under the *H* major category can cover terms related to corporate office buildings and factories etc. while those under the *Y* major category can cover terms related to the research and development activities in corporates. Therefore, we assumed that these three USAS semantic major fields can be useful in identifying and extracting core Chinese terms related to corporate entities, events and strategy. In practice, we collected and analysed the

[4] For further details about USAS system, see http://ucrel.lancs.ac.uk/usas/

Chinese terms annotated by these USAS tags listed in Table 1 and investigate to what extent our approach can facilitate the automatic semantic content analysis of the corporate reports.

I	***H***	***Y***
I1 - Money generally *I1.1* - Money: Affluence *I1.2* - Money: Debts *I1.3* - Money: Price *I2* - Business *I2.1* - Business: Generally *I2.2* - Business: Selling *I3* - Work and employment *I3.1* - Work and employment: Generally *I3.2* - Work and employment: Professionalism *I4* - Industry	*H1* - Architecture, kinds of houses & buildings *H2* - Parts of buildings *H3* - Areas around or near houses *H4* - Residence *H5* - Furniture and household fittings	*Y1* - Science and technology in general *Y2* - Information technology and computing

Table 1: USAS semantic fields/tags under three main categories *I*, *H* and *Y*.

3.1 Collection of corporate business reports as test data

For the test data of our experiment, we collected some business annual reports from a central Chinese corporation website (http://www.nbdqw.com/), which lists publically accessible annual business disclosure reports of Chinese corporates. Such annual reports publicise the corporates' performances and achievements over the earlier year as well as announce their future plans and strategies for business promotion purposes. This type of data provide valuable resources for analysing and predicting corporate information via the NLP approach. In order to guarantee a wide representativeness and a high quality of the test data, we manually selected annual disclosure reports of ten represenative Chinese public companies published from 2007 to 2014 wherever available (pseudonyms are used for the companies in this paper). All together, we collected 57 reports containing 3,584,956 Chinese characters. In terms of industry area, they have a wide coverage of China’s industry, spanning high technology sector, building and construction, automobile production, travel and tourism, chemical industry, heavy industry, electronics industry, and petroleum industry. We use the sample reports to study and test our methodology to automatically extract information about business strategy of different types of corporates. In this particular experiment, we use the sample data to test our method for automatically identifying core Chinese terms that are related to the business strategy. In order to estimate the lexical coverage of the semantic tagger for the corporate data, we calculated the percentage of the Chinese words in the sample reports that are recognised by our semantic tagger. Table 2 below shows a breakdown of test data distribution and Chinese semantic tagger lexical coverage for the ten corporates' reports, where the columns represent individual corporates and the first, second and third rows list respectively the number of documents, the size of the documents in terms of Chinese characters, and the the lexical coverage.

	corp. 1	corp. 2	corp. 3	corp. 4	corp. 5	corp. 6	corp. 7	corp. 8	corp. 9	corp. 10	total
docs num	5	4	7	5	6	6	6	7	6	5	57
chi. chars	336,133	251,686	245,275	264,416	342,622	276,025	570,537	593,088	422,743	282,431	3,584,956
lex cov.	76.91%	75.41%	77.08%	75.57%	78.26%	75.95%	74.86%	75.57%	74.42%	78.16%	76.02%

Table 2: Breakdown of test data size and lexical coverage of Chinese semantic tagger for annual disclosure reports from each corporate.

As these corporate reports are published as PDF documents, we needed to extract the text data from them in order to apply our semantic tagger software. We used Foxit PDF reader software (https://www.foxitsoftware.com/) for extracting the textual contents of the reports into plain text files. It is well known that it is highly challenging to extract clean text data from PDF documents, and inevitably our extracted text files contain some broken lines and words. But assuming such noise does not significantly affect the results of our experiment, we only carried out minor noise cleaning process such as removing number matrices derived from tables, page breaking lines etc.

3.2 Data processing, core term extraction and manual rating

The text files extracted from the PDF documents were processed using the Chinese semantic tagger. The reports were processed separately for each corporate in order to investigate how the different types of corporate business influence the results of our approach. For each corporate, we separately collected the terms which were tagged with USAS tags falling under the three main USAS semantic fields *I*, *H* and *Y* (refer to Table 1). Next, we collected the frequencies of the terms along with their tags (e.g. 资产_H1/H3), and then selected the most frequent 100 term_tag pairs for each corporate and for each of the three main semantic fields. As a result, we obtained three term-tag frequency lists for each corporate, which contain entries in the format shown below (the brackets are for inputing manual raking scores later).

#Freq.	Word_Tag	Manual Rating
2866	资产_H1/H3	[ ]
1226	合并_H2	[ ]

Finally we asked human raters familiar with corporate business reports to rank the terms using a numerical scale and give a score to each term using the guidelines shown in Table 3.

rating scale	Description of rating criteria
5	Closely related to corporate strategy description, e.g. 资产, 合并.
4	Fairly related to corporate strategy description, e.g. 方法, 计划 etc.
3	Ordinary nouns, verbs, e.g. 目录 , 计算.
2	Meaningful single character words, such as measurement .
1	Irrelevant words.

Table 3: Description of rating criteria for human raters.

According to the above rating scale criteria, the rating score of 3 indicates neutral terms. Therefore, if a term is rated with scores of 5 and 4, it is considered to be relevant to the corporate strategy information with a certain degree. Furthermore, if a group of terms have an average rating score above 3, it indicates that the terms collectively carry certain information

about corporate strategy, and higher the score, the more strategy information they carry. Following this criteria, in our experiment we measure performance of a method by observing the average rating score obtained by the terms extracted by it, with an average score above 3 indicating certain level of success, and with the average score of 5 indicating the maximum success.

3.3 Evaluation

Based on the human rater's scores, we assessed the performance of our approach in terms of the effectiveness of each USAS semantic category for the corporate strategy related term extraction. In detail, for each company, we calculated the average human raters' raking score for the *I*, *H* and *Y* semantic categories respectively in order to assess to what extent the terms falling under these categories carry corporate strategy related information. Table 4 below shows the results, where the columns represent the ten Chinese corporates and the rows represent the three major semantic categories.

sem cat	corp. 1	corp. 2	corp. 3	corp. 4	corp. 5	corp. 6	corp. 7	corp. 8	corp. 9	corp. 10	avg
I	4.33	3.52	3.66	3.8	4.18	3.44	3.27	3.32	3.28	3.25	3.60
H	3.43	3.07	3.02	3.15	3.06	3.06	3.28	3.07	3.4	3.28	3.18
Y	3.13	2.9	3.04	3.53	3.07	2.83	2.83	2.93	2.97	2.92	3.01

Table 4: Statistics of manual rating of terms for ten company reports.

As shown in table 4, the *I* semantic category produced the best average results, reflected by the average rating score of 3.60. On the other hand, the *Y* category produced the worst result, obtaining scores below 3 for six companies, indicating many terms extracted using this category are irrelevant to the corporate strategy information. Given that the rating score of 3.0 indicates the neutral terms, the average scores greater than 3.0 imply many of the extracted terms bear certain information about the corporate strategy information. We observed that some broken words caused by errors in the pdf-to-text conversion and Chinese word segmenting processes included in the term lists and we suspected they might affect the results. Therefore, we filtered out the single-character terms and re-calculated the statistics of the human rating. Table 5 below shows the results.

sem cat	corp. 1	corp. 2	corp. 3	corp. 4	corp. 5	corp. 6	corp. 7	corp. 8	corp. 9	corp. 10	avg
I	4.5	3.65	3.76	3.88	4.30	3.55	3.40	3.45	3.41	3.38	3.73
H	3.89	3.48	3.44	3.51	3.38	3.37	3.31	3.29	3.77	3.61	3.50
Y	3.45	3.30	3.33	3.83	3.45	3.17	3.15	3.21	3.15	3.18	3.32

Table 5: Statistics of manual rating of terms after removing single character words.

The comparison between Table 4 and Table 5 reveals that the broken words indeed affected the results. For example, by removing the Chinese single-character words, many of whom derived from broken words, the average rating score of *I* category was improved by 0.13. Figures 1, 2 and 3 illustrate the improvements achieved by the filtering, where points represent

the ten companies and the orange and grey lines represent respectively the average rating scales for the three main semantic categories obtained with and without single character words included. These graphs demonstrate a consistent noticeable improvement of the rating scores after the filtering except for *company 7* (building and construction) in *H* category. Overall, the filtering has a significant impact on *H* and *Y* categories. This result indicates that Chinese single-character words may not be suitable for key term extraction in the content analysis of corporate reports.

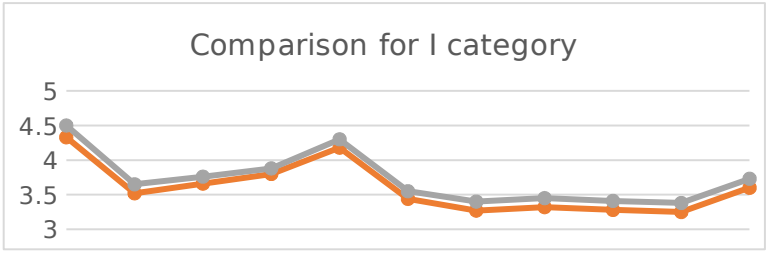

Figure 1: Rating scores comparison for *I* category.

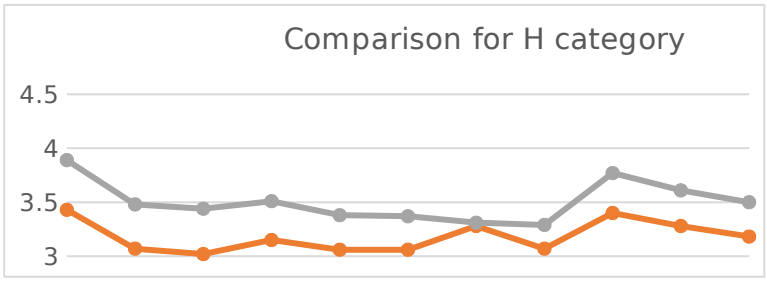

Figure 2: Rating scores comparison for *H* category.

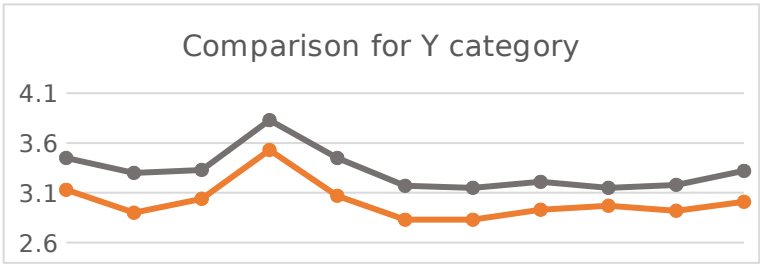

Figure3: Rating scores comparison for *Y* category.

Although limited and preliminary, our experiment and the analysis of results show that, with carefully selected semantic categories and proper text cleaning process, the Chinese semantic tagger can potentially facilitate rapid automatic extraction of corporate strategy related terms and concepts, which is crucial for achieving timely delivery of corporate business information to stakeholders and clients based in large-scale data, such as in big data context. Because as far as we know our experiment is the first study on automatic extraction of corporate strategy related Chinese terms, it is difficult to make direct comparison with other existing methods. As our research develops, standard test data will be produced and more methods will be tested in order to find an optimal solution.

4. Conclusion

In this paper, we have reported our experiment in which we tested the feasibility of automatically identifying and extracting corporate strategy related Chinese terms from corporate annual disclosure documents using a Chinese semantic tagger. The analysis of our experiment results demonstrates that it is feasible to use such a tool to automatically extract key Chinese terms for further analysis of corporate strategy information. In future work, we will improve the lexical coverage and accuracy of the semantic tagger and design a better approach for selecting appropriate semantic tags to improve the analysis of corporate reports.

References

[1] M.J. Jones and P.A. Shoemaker. *Accounting narratives: a review of empirical studies of content and readability.* Journal of Accounting Literature, 13, 142–85 (1994)

[2] B. A. Rutherford. *Genre analysis of corporate annual report narratives - a corpus linguistics-based approach*. Journal of Business Communication, Volume 42, Number 4, pp. 349-378 (2005)

[3] V. Beattie and S. Thomson. *Lifting the lid on the use of content analysis to nvestigate intellectual capital disclosures*. Accounting Forum, 31(2),129-163 (2007)

[4] N. M. Brennan and D. M.Merkl-Davies. *Accounting narratives and impression management*. In: Russell J. Craig, Jane Davison and Lisa Jack (eds.), The Routledge Companion to Accounting Communication. London: Routledge, pp.109-132 (2013)

[5] D. M. Merkl-Davies, N. M. Brennan and P. Vourvachis. *Content analysis and discourse analysis in corporate narrative reporting research: a methodological guide*. Centre for Impression Management in Accounting Communication (CIMAC) Conference 2014. Bangor, UK. (2014)

[6] G. Hoberg and M. L. Craig. *Do fraudulent firms strategically manage disclosure?*. The 8th LSE/LUMS/MBS Conference, ICAEW, London. (2014)

[7] M. El-Haj, P. Rayson, S. Young and M. Walker. *Detecting document structure in a very large corpus of UK financial reports*. In Proceedings of The 9th Edition of the Language Resources and Evaluation Conference, Reykjavik, Iceland. (2014)

[8] S. Young. *Textual analysis and investment decisions: an overview*. Sell-Side Meeting: Citi Investment Research & Analytics, 2014 Citi Quantitative Finance Conference. Valencia, Spain. (2014)

[9] P. Rayson, D. Archer, S. Piao and T. McEnery. *The UCREL semantic analysis system*. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp. 7-12 (2004)

[10] T. McArthur. *Longman Lexicon of Contemporary English*. Longman London. (1981)

[11] S. Piao, F. Bianchi, C. Dayrell, A. D'Egidio and P. Rayson. *Development of the multilingual semantic annotation system*. In Proceedings of The 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015), Denver, Colorado, USA. (2015)