Study of Corpus Classification Management Method in Construction Correspondence Files

Dan Li[1, a], Qiqi Zhang[1, b], Jiadong Cao[1, c] and Yihua Mao[1,2, *, d]

[1]Institute of Construction Management, Zhejiang University, Hangzhou 310058, China;

[2]Binhai Industrial Technology Research Institute of Zhejiang U0niversity, Tianjin 300301, China.

[a]dannl@zju.edu.cn, [b]21812255@zju.edu.cn, [c]657714480@qq.com, [d, *] maoyihua@zju.edu.cn

Abstract. The construction correspondence files, including letters, emails and meeting summaries, are used for information exchange during the construction stage. Information about design adjustment, project change and on-site coordination among several parties are included in these files. However, searching for the record is labor intensive and inefficient in the settlement and claim process. Two actual-based corpus classifications were brought out and tested by literature and statistics method in this paper, which could be the guidance for machine language learning. Finally, the application of contract corpus classification is proposed.

Keywords: Engineering Management; Construction Stage; Corpus Classification.

1. Introduction

1.1 Research Background

Various problems and disputes were raised during the construction stage due to factors such as on-site condition, weather influence, design adjustment and material change. The contractors, subcontractors and owners are all involved in conflict coordination, which requires solving issues and dividing responsibilities [1]. Practically, the information exchange among them was often completed by a large number of meeting summaries, texts, e-mails, letters, and other documents. When it comes to the project settlement or a claim, consulting contract and the corresponding documents are quite necessary [2]. As a result, the excessive number of change files not only consumes manpower, material and financial sources but also suffers inefficiency and high error rate. This paper brought out two corpus classification methods that were suitable for machine language learning. Furthermore, automatic labelling and classification by computer could be achieved through file screening program based on neural network training [3].

1.2 Corpus Classification

Natural language sentences can be classified into unsupervised, semi-supervised and fully supervised [4]. To guarantee the accuracy and applicability of clustering results, the semi and fully supervised method require manually classifying the sentences into predefined categories. This set of sentences are often called corpus [5].

In general, several steps are contained in training a machine learning model. Define the standards of annotation, obtain artificially marked corpus, design the machine learning model and train it based on the dataset and evaluate the tested model [6]. In the following part, two kinds of annotation rule were proposed and tested separately.

Several facets should be checked to test the rationality of corpus. Specifically, an appropriate and reasonable purpose of corpus labelling, a suitable corpus and their high match degree are necessary. Also, different categories shall be defined and described clearly, for the labeller to classify one sentence into a category as accurate as possible.

2. Original Classification Method

2.1 Classification and Content

Conflicts, adjustment and problems are revealed in actual construction projects and should be classified properly. Through analyzing existing engineering examples, we concluded the correspondence files and classified the raised problems into four categories, namely resources, security, technology and personnel. Table 1 exhibited their definition and content.

Table 1. Problem Classification and Definition

Category	Definition
Resources	Tangible resources like building materials, water and electricity and construction equipment; intangible resources like drawings, construction documents, data, construction schedule and others.
Security	Construction safety, including foundation, structure, water and electricity, HAVC engineering; fire safety, equipment safety and personal safety, etc.
Technology	Technical equipment, including foundation, structure, water and electricity, HAVC engineering during construction stage; construction technology and process, etc.
Personnel	Personnel technical quality, attendance frequency, interpersonal communication, construction personal work efficiency, etc.

After settling the definition, several researchers were invited to label and classify construction problems. Their consistency was tested by the Kappa coefficient, which required two evaluators to label issues based on given types without communication.

2.2 Consistency Test for Original Method

In our occasions, two people were invited to divide the issues shown in correspondence files into the above four categories sentence by sentence. Afterwards, the Kappa coefficient, varying between 0 and 1, was used for a preliminary consistency test on the proposed method. Table 2 shows its calculation rules. A result above 0.75 means proper consistency and below 0.4 refers to poor.

Table 2. Calculation Rules of Kappa Coefficient

	Assessor 2			
		Category A	Category B	Total
Assessor 1	Category A	1A2A	1A2B	1A2A+1A2B
	Category B	1B2A	1B2B	1B2A+1B2B
	Total	1A2A+1B2A	1A2B+1B2B	tot:1A2A+1A2B+1B2A+1B2B
Pe=[(1A2A+1A2B)*(1A2A+1B2A)+(1B2A+1B2B)*(1A2B+1B2B)]/tot/tot				
Pa=(1A2A+1B2B)/tot				
Kappa=(pa-pe)/(1-pe)				

Note: 1A2A refers to the number of issues that assessor 1 marks to category A while assessor 2 also marks to category A, 1B2A refers to the issue number that assessor 1 marks to B while assessor 2 marks to A, other symbols are likewise.

For the original classification results, calculate the Kappa coefficient and the process is shown in Table 3. Their Kappa coefficient values 0.4564, which is a little higher than the below boundary. Thus, a redefinition is required.

Table 3. Assessment Result of Kappa Coefficient for Original Method

	Assessor 2					
	Number	Resources	Security	Technology	Personnel	Total
Assessor 1	Resources	11	0	0	0	11
	Security	1	3	0	0	4
	Technology	9	0	2	2	11
	Personnel	0	0	0	2	2
	Total	21	3	2	2	28

Pe=(21*11+3*4+11*2+2*2)/28/28=0.3431

Pa=(11+3+2+2)/28=0.6429

Kappa=(pa-pe)/(1-pe)=0.4564

3. Updated Classification Method

3.1 Classification and Content

Again, the correspondence files were concluded and issues were classified into five categories, namely nature, material, information, technology and personnel. Table 4 exhibited their definition and content. Comparing with the original method, the number of issues belonging to resources is segregated into nature and material categories.

Table 4. Updated Problem Classification and Definition

Category	Definition
Nature	Nature force majeure and environmental factors that will change the construction process or in/decrease the construction content. Eg. severe weather conditions, changes in national policies, etc.
Material	All kind of physical resources appeared in construction projects, including quality, transport, replacement and other conditions. Eg. concrete quality defects, water and power poor supply, etc.
Information	All kind of engineering documents and electronic information during construction, including their request, change, storage and so on. Eg. engineering drawings, schedule information, etc.
Technology	Various technologies for construction, including operation and construction technology. Eg. tower crane control, decoration, etc.
Personnel	Personnel allocation and coordination in construction projects and others related to subjective actions.

3.2 Consistency Test for Updated Method

Two researchers who haven't been involved in either the discussion or the previous test were invited to label the same issues under the updated categories. Their responded numbers are shown in Table 5. During classification, they were told to find out 'root cause'. That's to say, assuming potential problems happened due to the phenomenon described, label the category that responsibility belongs to. For instance, when a certain building material arrived late on site and the contractor was told to follow, the delay is a matter of material resources rather than human-caused issues.

The Kappa coefficient value of the updated method is 0.60, referring to its modest performance. Although the factor is not larger than 0.75, an increment of nearly 0.15 indicated that the updated method is more acceptable than the original.

Table 5. Assessment Result of Kappa Coefficient for Updated Method

		Assessor 2					
	Number	Nature	Material	Information	Technology	Personnel	Total
Assess-or 1	Nature	0	0	0	0	0	0
	Material	0	9	0	0	0	9
	Information	0	0	5	1	0	6
	Technology	0	0	0	0	1	1
	Personnel	0	4	1	0	4	9
	Total	0	13	6	1	5	25

Pe=(0*0+13*9+6*6+1*1+5*9)/25/25=0.3148

Pa=(0+9+5+0+4)/25=0.72

Kappa=(pa-pe)/(1-pe)=0.60

4. Summary

In this paper, the purpose of establishing corpus is to minimize time, labor work and other costs when the project settlement or a claim happened. Two sets of categories were proposed and the consistency was tested. A cause-based principle is settled to match the classification purpose and the given corpus. Meanwhile, under a certain type, one sentence is segregated to a category only.

According to the increased value of the Kappa coefficient, the original classification method works worse than the updated method. That's to say, the difference between resources and security is not so obvious to tell. Thus, the problems recorded in the correspondence files during construction stage could be classified into nature, material, information, technology and personnel. Moreover, this method is also suitable for contract management and machine language learning.

References

[1]. Yuan Xiangyun. Management of correspondence between key projects and constructions [J]. Zhejiang Archives, 2011, 8:41.

[2]. CHEN Yongqiang, WANG Wenqiang, You Jingjia. Understanding the multiple functions of construction contracts: the anatomy of FIDIC model contracts[J]. Construction Management and Economics, 2018, Vol. 36(8): 472–485.

[3]. SHAMI M, VERHELST W. Automatic Classification of Expressiveness in Speech: A Multi-corpus Study [G]. Speaker Classification II: Selected Projects. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 43–56.

[4]. Dareen A. Salama, Nora M. El-Gohary, A.M.ASCE. Automated Compliance Checking of Construction Operation Plans Using a Deontology for the Construction Domain[J]. Journal of Computing in Civil Engineering, 2013, Vol. 27(6): 681–698.

[5]. Dareen M. Salama. Semantic deontic modeling and text classification for supporting automated environmental compliance checking in construction. University of Illinois at Urbana-Champaign, U.S., 2011, 17-81.

[6]. Liu Changchun, Sun Yuanyuan. Research and Application of Modularization of Lightweight Wood Structure Industrialized Houses[J]. Construction Technology, 2016, 45(04): 35-38.