Automatic Extraction of Data Governance Knowledge from Slack Chat Channels*

Rob Brennan[1], Simon Quigley[1], Pieter De Leenheer[2], and Alfredo Maldonado[1]

[1] ADAPT Centre, Computer Science and Statistics, Trinity College Dublin, Ireland
{rob.brennan, siquigle, maldona}@scss.tcd.ie
[2] Collibra Research Lab, New York, USA
pieter@collibra.com

Abstract. This paper describes a data governance knowledge extraction prototype for Slack channels based on an OWL ontology abstracted from the Collibra data governance operating model and the application of statistical techniques for named entity recognition. This addresses the need to convert unstructured information flows about data assets in an organisation into structured knowledge that can easily be queried for data governance. The abstract nature of the data governance entities to be detected and the informal language of the Slack channel increased the knowledge extraction challenge. In evaluation, the system identified entities in a Slack channel with precision but low recall. This has shown that it is possible to identify data assets and data management tasks in a Slack channel so this is a fruitful topic for further research.

Keywords: Ontologies · Data Management · Systems of Engagement.

1 Introduction

Data governance is increasingly important, and formal systems of data governance that audit and channel communication about data have become widespread. However large amounts of intra-organisational communication, including data governance information, is carried over unstructured channels such as Slack, and thus is not easily captured by a traditional data governance system.

Natural language processing (NLP) techniques have matured greatly over the last decade and can convert this unstructured human communication into machine-processable structured data for analysis and audit. Transformation into open knowledge models, such as RDF and OWL, provides the greatest flexibility to support inference, interlinking and global knowledge sharing. However data governance knowledge extraction from Slack chat has many challenges: short interactions, informal use of language, lack of standard test corpora, small datasets

* This research has received funding from the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded by the European Regional Development Fund.

compared with global Twitter feeds, expert domain knowledge required to annotate training data and the abstract nature of data governance concepts compared with traditional NLP concepts used for named entity recognition (NER) tasks.

Given the lack of published training data for this task and the vast data requirements for neural NLP approaches, it was decided to investigate the performance of a state-of-the-art NER system based on conditional random fields (CRF). Thus the following research question is proposed: *To what extent can CRF-based Named Entity Recognition be used to extract data governance knowledge from an enterprise chat channel?* Data governance information is defined here as a set of data governance assets, processes, rules, roles and users.

This paper provides the following contributions: a new, open, data governance ontology and a trained data governance NER system and evaluation of the system performance using real-world enterprise Slack data.

The rest of this paper is structured as follows: section 2 use case and requirements, section 3 related work, section 4 our approach to data governance knowledge extraction, section 5 evaluation of the prototype system and finally section 6 provides conclusions.

2 Use Case: Slack Channels as Data Governance Systems of Engagement

This paper is a first step in linking semantics-driven AI and user-centred data governance Systems of Engagement (SoE) [10]. The following diagram shows the systemic interaction between the Collibra DGC (Data Governance Centre) platform, being the System of Record (SoR), and a set of systems of engagement e.g. based on a Slack channel. The diagram was adapted from our work on community-based business semantics management [4] which was foundational for Collibra .

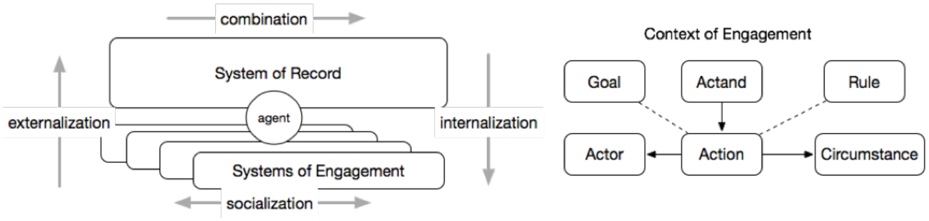


Fig. 1: LHS: SECI knowledge conversions between one SoR and many SoEs, through intelligent agents (on e.g., Slack). RHS: engagement contexts.

Both components consist of multiple instances of data governance operating concepts. In the Collibra platform SoR these concepts are shared, explicit and understood, i.e. based on a shared ontology. we refer to this ontology as the Collibra data governance operating model. Yet the ontology of these data governance concepts may differ widely in the various SoE applications we wish to integrate. On the SoE side, instances of these concepts are typically less explicit and usually scattered. They can be more of a socio-technical of nature, i.e. tacitly shared among humans, resulting in a poor unified record for supporting data governance as opposed to a SoR. E.g., your actor identity in Slack may be different from Confluence and Collibra. Also references to actands and actions may suffer wide differences in vocabulary and grammar, requiring (named) entity resolution. Slack has become a key corporate system of engagement and a hence a source of vital data governance context as data assets are discussed, evaluated, located and exchanged through Slack. Now it becomes imperative to enable the data governance SoR to engage with that unstructured context.

Enterprise data management traditionally focuses on centralizing formal management of operational and analytical data. This conservative view inhibits us from seeing the underlying fabric that glues all the data together. This data is scattered across engagement platforms and largely unstructured, usually expressed by humans in context-heavy natural language. Data governance systems must tap into this unstructured data and interactions to bring greater insights into how people, workplaces, and perhaps societies interact.

Derived Requirements: (1) A common ontology of data governance concepts and context that can span data governance in both systems of record and systems of engagement. For widespread adoption it is important that this uses an open, standards-based model such as W3C's OWL/RDF; and (2) Ability to convert unstructured communications into machine-readable data. This requires a knowledge extraction framework that is specialised both for the data governance domain and for the style and content of the communications channel(Slack).

3 Related Work

Here we discuss relevant work in: knowledge models for data governance, and NER for data governance in Chat channels.

Knowledge models for Data Governance: Data governance is defined here as the organisational function aimed at the definition and enforcement of data policies to enable data collaboration, understanding and trust. To our knowledge there is no over-arching semantic model for data governance, e.g. ISO 38505-1 addresses foundations for data governance, but it does not provide a knowledge model of the domain.

However there are many existing standards-based metadata vocabularies that are important for data governance, e.g. the W3C provenance (PROV) standard [7], DBpedia's DataID to describe data assets, the H2020 ALIGNED project knowledge models of data lifecycles and tools. W3C PROV can be used as a basis for specifying activities, agents and entities in a data governance model. This would enable interoperability with standard PROV services such as meta-data repositories and wider enterprise workflow and information integration applications. The W3C data quality vocabulary (DQV) standard can be used to describe a dataset's quality, whilst the data value vocabulary (DaVE) [1] could act as basis for describing data value metrics and dimensions. Thus there is a need for an upper governance ontology to glue together these individual vocabularies to describe the data governance domain as a knowledge model.

NER for data governance in Chat communications channels: To the best of our knowledge, this is the first usage of an NER approach to extract data governance concepts. NER aims to identify individual words or phrases in running text that refer to information units such as person names. We employ a state-of-the-art machine learning NER method called conditional random fields (CRF). In addition to being able to extract traditional entities, CRF has been shown to be able to successfully extract other types of information, e.g. headers, citations and key phrases from research papers [2]. Given the success of CRF-based methods in such a diverse array of problems and languages [8], we consider them to be a good candidate for extracting data governance information.

CRF systems in the NLP literature are typically trained and used on formal and well-formatted texts like news articles. There has been less attention on informal text, e.g. Slack chat logs. Informal text characteristics such as incomplete sentences, non-standard capitalisation and misspellings generally lead to a loss of accuracy for NER systems. There are some papers on the application of NER systems to informal texts, typically on social media. For example, [5] investigates the main sources of error in extracting entities in tweets, and how these errors could be addressed. It found that non-standard capitalisation had a particularly negative impact on NER performance, with greater impact than slang or abbreviations. It investigated the use of part-of-speech tagging and normalisation to reduce the impact of noise in tweets, but ultimately found that precision and recall scores remained low using NER standard algorithms. Similarly, [3] explores the use of word representations to improve the effectiveness of a NER in labelling Twitter messages. This work found that general NER systems performed very poorly in labelling tweets.

4 Data Governance Knowledge Extraction Approach

Before integrating the knowledge extraction tool into the Slack channel as a bot it was necessary to train and evaluate a NER system capable of detecting the data governance entities defined by the new open data governance ontology based upon the Collibra data governance operating model and the state of the art semantic web data governance ontologies.

4.1 Knowledge Architecture

The knowledge model used to classify the data governance entities and relationships detected in the Slack channel was based upon the Collibra data governance operating model but generalised as an OWL ontology.

Collibra Data Governance Operating Model This Model[3] has been implemented by hundreds of companies. It establishes the foundation for and

[3] https://university.collibra.com/courses/introduction-to-the-operating-model-5-x/

drives all data stewardship and data management activities. It has three subcategories each addressing a key design question.

1. What is to be governed in terms of Structural Concepts, including asset types, (complex) relation types, attribute types.
2. Who governs it, in terms of Organisational Concepts. These include Communities, domains, users, user groups.
3. How is it to be governed in terms of Execution and Monitoring Concepts, including role types, status types and workflow definitions.

Data stewardship activities align and coordinate data management operations. Data Management concerns the integration of stewardship activities with third-party applications (such as data profilers, scanners, metadata repositories, etc.). In this work we only extracted (data) Asset Types. An Asset is the capital building block in data governance. An Asset Type formally defines the semantics of an asset. There core asset types, or asset classes as illustrated below. An asset captures the authoritative lifecycle metadata, in terms of attributes and relations with other assets, for one of the following five classes of assets:

- a governance asset (such as a policy or data quality rule): e.g., ‘Customer Data Protection Policy’ is the name of an asset of type ‘Policy’
- a business asset (such as a business term or metric): e.g., ‘Client’ is the name of an asset of type ‘Business Term’;
- a data asset (such as e.g., reports or predictive models): e.g., ‘first_name’ is the name of an asset of type ‘Column’;
- a technology asset (such as a database or system): e.g., ‘CRM’ is the name of an asset of type ‘System’
- an issue (such as a data quality issue): e.g., ‘Customer Lifetime Value Report data is of too low quality’ is the name for an asset of type ‘Data Issue’.

The Open Data Governance Ontology (odgov) Conversion of the entire Collibra Data Governance Operating Model into an OWL ontology is a large task beyond the scope of this paper. However here we have created the first upper data governance ontology that serves the knowledge extraction and annotation needs of the data governance NER system.

This required the creation of eight main OWL classes (GovernanceAsset, BusinessAsset, DataAsset, TechnologyAsset, Role, Issue, and User) and parent classes for Assets and data governance execution and monitoring concepts. In addition a data management task class was created to hold the frequent references to data management activities (e.g. importing, copying, and backing up data) that appear in the Slack channel. This last class was an extension to Collibra data governance operating model as these activities are not separately modelled from business processes within that model. In addition three relation types from the Collibra model are included: the generic relation between assets, the uses asset relation and the is governed by relation.

Then these upper data governance terms were linked to the W3C provenance ontology by defining all odgov:Asset as subclasses of prov:entity, odgov:DataAssets as subclasses of dataid:dataset, dgov:DataManagementTask as a subclass of prov:Activity and odgov:User as a subclass of prov:Activity. Then a set of machine-readable metadata fields were defined so that the ontology is publishable via the live OWL documentation (LODE) environment. The final ontology and html documentation is available on the web[4].

4.2 NLP/NER Toolchain

Stanford NER [6] was used for the experiments reported in this paper. It is a widely used open source implementation of a CRF system that performs well with minimal fine-tuning requirements as it includes many built in feature extractors to enhance performance. The Collibra slack chat dump was tokenised using the Stanford Tokeniser (part of the Stanford CoreNLP toolkit [9]). In addition, the authors developed Python scripts for additional data pre-processing, conversion, experiment automation and evaluation. We have made these scripts available online[5]. The annotation of the chat dump was done through the Brat annotation tool by the authors. Section 5.1 details the annotation scheme.

5 Evaluation

We evaluated the effectiveness of the NER system at extracting data governance information from a real Slack chat channel. section 5.1 describes the slack chat dataset and the scheme used to annotate it. Next we present the experimental protocol. The results (section 5.3) show the accuracy of the NER system varies according to the actual Data Governance Information Category it seeks to predict. We present a correlation analysis to explain these variations.

5.1 Data Annotation

The test dataset is a raw dump of messages from Collibra's Data Governance team Slack. The resulting data consisted of 7,022 messages totalling about 300,000 tokens. These messages were first filtered to detect those most directly related to data governance using Shah et al.'s binary classifier[11], this produced a final dataset of 800 messages, totalling 4,749 tokens. The entities annotated in the dataset were based on our data governance ontology detailed earlier. This approach used the initial entity types: Governance Assets(`Gov`), Business Assets(`Bus`), Data Assets(`Data`), Technology Assets(`Tech`), Governance Roles(`Role`), Users(`User`) and Issues(`Issue`). However, upon annotating a sample of the dataset with this scheme, the Business Asset class was found to be overloaded and an additional entity types was used: `Dmtask` to label text representing a data management task, such as upgrading or backing up a database. The actual annotation work was conducted by the authors using the BRAT annotation tool. The number of annotated tokens for each data governance entity

[4] http://theme-e.adaptcentre.ie/odgov

[5] https://github.com/simonq80/datagovernancenter

was as follows: Governance Assets 182, Business Assets 196, Data Assets 503, Technology Assets 236, Users 153, Governance Roles 14 and Issues 310 for a total of 1738 word tokens. Of the total 4,749 tokens, 3,011 were not related.

5.2 Experimental Protocol

We evaluate the accuracy of our system using standard precision, recall and F-1 scores, which are commonly used for evaluating NER systems. We compute these scores on each entity type as well as overall scores for all categories.

The computation of these scores require the dataset to be partitioned into training and test sets. In order to produce robust evaluation scores, we followed the k-fold cross validation evaluation scheme. Under this scheme, the dataset is divided into k equally sized sections. Each of the k sections is used as the test set once, with the remaining $k - 1$ sections used as the training set. This results in k test results which are averaged to get a performance estimate of the model. Larger values of k result in a smaller test set and larger training set for each fold. Cross validation tends to have low variance and generally low bias.

10-fold cross validation (i.e. $k = 10$) is typically used as it is generally considered to be optimal for reducing bias and variance for accuracy estimation. However, due to the small size of our dataset, test portions tended to be too small in 10-fold cross validation to represent all Data Governance Information Categories reliably. So we experimented as well with 5- and 4-fold cross validation variants and 4-fold validation was found to have the best F1 score. Evaluation results for these experiments are presented in the following section.

5.3 Results

Results for each Data Governance Information category from 4-fold cross-validation can be seen table 1. Across all metrics, the NER performed by far the best on the `User` category. It also performed well on both `Data` and `Tech`, achieving relatively high precision, but with worse performance in recall. Aside from `Role`, which was never predicted due to its rarity in the dataset (hence the N/A values in the table), the NER performed the worst on the `Gov`, `Issue` and `Bus` categories, all of which had very low recall and relatively low precision. With the exception of `User`, all entity types had notably higher precision than recall.

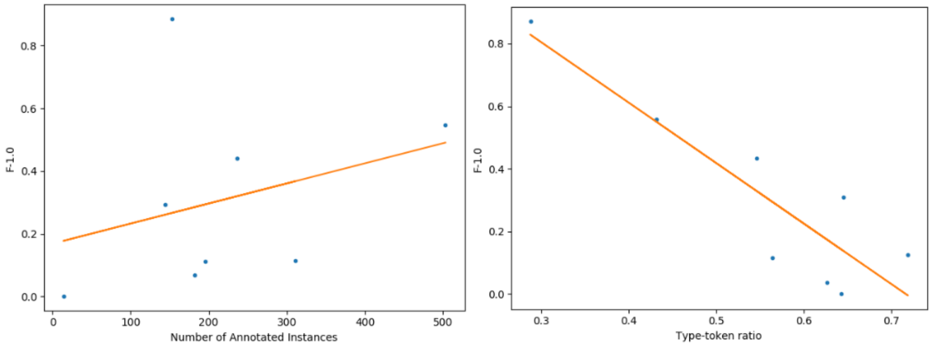
Different Data Governance Information Categories perform differently. We find that this variation in performance correlates with the number of annotated instances for each category (the more instances a category has, the better its performance) as well as with its type-token ratio (the lower the category's type-token ration, the better its performance). We now look into these two correlations.

Number of annotated instances per category As expected, Data Governance Information Categories that have more annotated instances in the dataset will tend perform better. This is simply because the CRF algorithm is exposed to more examples and is thus able to learn relevant features more reliably. Figure 2a plots this correlation for the F-1 measure (precision and recall show a

Table 1: 4-Fold Cross-Validation Results per Category

Category	Precision	Recall	F-1 Score
Bus	0.3611	0.0663	0.1121
Data	0.6139	0.493	0.5469
Dmtask	0.4918	0.2083	0.2927
Gov	0.3684	0.0385	0.0697
Issue	0.3889	0.0675	0.1151
Role	N/A	0.0000	N/A
Tech	0.6423	0.3347	0.4401
User	0.8831	0.8889	0.886

similar correlation). The Pearson correlation coefficient is 0.32. A least-squares polynomial line is shown in the figure to make this correlation more visible.



(a)

(b)

Fig. 2: Correlation between F-1 score and (a) the number of annotated instances of a category and (b) the type-token ratio of a category

Type-token ratio is the number of unique words (types) of a category divided by the total number of words (tokens) of that category. It is a measure of word diversity in each category: the higher the type-token ration, the more word diversity there is in the category. Categories with low type-token ratios tend to use more or less the same words (little word diversity). So it is not surprising that figure 2b shows a very strong negative correlation between the type-token ratio of categories and their F-1 score. The Pearson correlation coefficient is -0.89. Again, a least-squares polynomial line is shown to visualise the correlation. Precision and recall plots show similar correlations.

6 Conclusions and Future Work

This paper has demonstrated that CRF-based NER is a promising approach for extraction of data governance knowledge based on a new open upper ontology for data governance. Given the limitations of the current training data set (c. 5,000 annotated tokens) it is a positive result to see two categories of governance entity detected with over 0.6 precision and one at 0.88. The recall scores are poor, but we hope that precision is more important for the planned application as an interactive bot on the Slack channel system of engagement who must minimise their number of incorrect interventions to avoid frustrating the user.

References

1. Attard, J., Brennan, R.: A semantic data value vocabulary supporting data value assessment and measurement integration. In: Proc. 20th International Conference on Enterprise Information Systems,. pp. 133–144. INSTICC, SciTePress (2018)
2. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. arXiv preprint arXiv:1704.02853 (2017)
3. Cherry, C., Guo, H.: The unreasonable effectiveness of word representations for twitter named entity recognition. In: Proc. 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 735–745 (2015)
4. De Leenheer, P., Debruyne, C., Peeters, J.: Towards social performance indicators for community-based ontology evolution. In: Workshop on Collaborative Construction, Management and Linking of Structured Knowledge at ISWC (2009)
5. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Information Processing & Management **51**(2), 32–49 (2015)
6. Finkel, J.R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proc. 43nd Annual Meeting of the Association for Computational Linguistics. pp. 363 – 370 (2005)
7. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. Tech. rep. (2012), http://www.w3.org/TR/prov-o/
8. Maldonado, A., Han, L., Moreau, E., Alsulaimani, A., Chowdhury, K.D., Vogel, C., Liu, Q.: Detection of Verbal Multi-Word Expressions via Conditional Random Fields with Syntactic Dependency Features and Semantic Re-Ranking. In: Proc. 13th Workshop on Multiword Expressions. pp. 114–120. Valencia (2017)
9. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014)
10. Moore, G.: Systems of engagement and the future of enterprise it-a sea change in enterprise it. Tech. rep., AIIM, Silver Spring, Maryland (2011)
11. Shah, J.: Utilizing natural language processing and artificial intelligence to identify plausible data requests on slack and linking it to collibras system of record tool dgc. Tech. rep., Collibra (2017)