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

Abstract. This paper describes a new data governance knowledge extraction system 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 prototype system addresses the urgent need to convert unstructured information flows about data assets in an organisation into structured knowledge that can easily be queried, transformed and support data governance inference or audit trails. By basing the approach on an open OWL data governance ontology it supports a wide array of use cases and tools. 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 trained NER model was capable of identifying entities in a Slack channel with some precision but low recall. In addition we created an annotated training set of enterprise Slack channel data and a set of annotation guidelines for describing training data based on the data governance ontology. This study has shown that with relatively little training data 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 · Named Entity Recognition · Data Management · Systems of Engagement.

1 Introduction

Data governance is increasingly important in organisations, 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

* 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.

system. A number of recent events demands a further extension of data governance concerns from within the business ecosystem to society as whole. After a decade of year-on-year records in data breaches[3], the EC published the General Data Protection Regulation[4] (GDPR) which all companies with EU citizens in their customer base must implement.

Natural language processing (NLP) techniques have matured greatly over the last decade and are available to turn 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 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. Nonetheless mastering the use of natural language to interact with a formal data governance system would create opportunities to build data governance systems of engagement that use lightweight interactions yet retain greater control than inflexible data governance systems of record.

Given the lack of standard training data for this task and the vasts 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) [14]. 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. Details of these entities are described in the knowledge model in section 5.

First an annotated corpus was created from a dump of a public enterprise Slack channel from Collibra. This was annotated based on a the data governance ontology developed as part of this work but based upon the Collibra Data Governance Operating Model. An annotation scheme was developed and the Brat annotation tool[5] used to create the annotations. This annotated dataset was then used for training and evaluating our data governance NER system based on Stanford NER [9]. To fairly compare different NER configuration results, we set up a standard testing procedure. A Python script was created to comprehensively evaluate the performance of the NER as a whole, as well as a detailed breakdown of its performance for each entity type.

This paper provides the following contributions: a new, open, standards-based data governance ontology, Slack channel data governance annotation guidelines, a trained data governance NER system and evaluation of the system performance using real-world enterprise Slack data. This evaluation focuses on understanding the impact of abstract data governance entities on traditional NER

[3] https://digitalguardian.com/blog/history-data-breaches
[4] https://www.eugdpr.org/
[5] http://brat.nlplab.org

techniques. This is of interest to both semantic web and data governance researchers.

The rest of this paper is structured as follows: section 2 describes our use case, section 3 extracts requirements for the system, section 4 discusses the related work, section 5 discusses our approach to data governance knowledge extraction, section 6 presents a lab-based evaluation of the prototype system and finally section 7 provides conclusions and directions for future work.

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

The work presented in this paper is a first step in linking semantics-driven AI and data governance Systems of Engagement (SoE) to support new interactions that are natural to knowledge workers and data scientist teams. This applies to all kinds of data, all within the established Collibra platform that ties all governance concerns of those interactions together. 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. The diagram was adapted from our work on community-based business semantics management (De Leenheer et al., 2009)[6] which was foundational for Collibra . We also adopt the terminology from the FRISCO semiotic framework for information system concepts by (Falkenberg et al. 1998)[8].

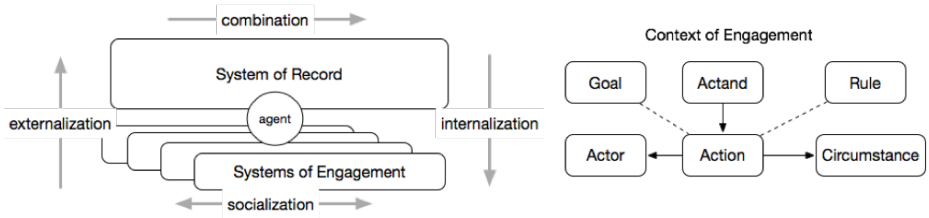

Fig. 1. LHS: SECI knowledge conversions between one SoR and many SoEs, through intelligent agents (on e.g., Slack). RHS: engagement contexts.

Both components consist of multiple instances of data governance operating concepts. We distinguish six key concepts: actor, actand (i.e., resources on which actions can be performed), action, goal, rule and circumstance. Actions are triggered/paused/terminated by circumstances and performed by actors on actands, with certain goals (or intents) in mind. Actions can be governed by rules, such as roles and affordances. In the Collibra platform SoR these three components are shared, explicit and understood, i.e. based on a shared ontology. we refer to this ontology as the Collibra data governance operating model. The components have been fully computerized, i.e. we have a well-defined and structured digital record of what actions (sequenced by workflows) have been performed

on which data asset (actand), and by whom (actors). Yet the ontology of these data governance concepts may differ widely in the various SoE applications we must consider 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 use vocabulary and grammar, requiring (named) entity resolution.

Enterprise data management has traditionally focused on centralizing formal management of operational and analytical data for inward purposes such as optimizing internal coordination and predicting the customers next transaction. Thereby, narrowing the focus down on structured subsets of the enterprise data universe.

This conservative telescope on the data universes inhibits us from seeing the underlying fabric that glues all the data records together. In the end, the bulk of operational and analytic data are mere records summarizing the transaction history, not including the more complex (often physical) personal interactions. This data is scattered across engagement platforms and often largely unstructured, usually expressed by humans in context heavy natural language. However, when observed as a whole, these core interactions generate a network effect digitized as social capital. More and more systems of engagement digitalize social capital. The next generations of data governance will tap into these systems to record a richer context for any given data asset that was used or produced.

Our data governance "universe" therefore consists of an insofar undefined dark energy that also could explain its expansion. Similarly, the Big Data bang consists of social capital that we know exists but is hard to understand because its scattered and unstructured. Capturing it will help us understand how the data universe will further expand. This will bring us greater insights into how people, workplaces, and perhaps entire societies interact rather than just a snapshot of the mere transactional data that is the traditonal scope of data governance systems of record.

Slack has become an increaslingly important channel for unstructured communication in the enterprise. It is a key corporate system of enagagement 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 system of record to engage with that unstructured context.

3 Requirements

In order to realise the technical vision of the use case it is necessary to extract a number of requirements that any suitable system should conform to. These are briefly described as follows:

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.
2. This ontology should leverage existing data governance-related ontologies and vocabularies already established in the community.
3. The ontology should also have explicit and preferably open licensing available to enable an eco-system of tools and solutions to develop around the model.
4. Ability to integrate multiple systems of engagement with a governance system of record. It should be possible to create data governance tool-chains that publish and consume the ontology.
5. 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).
6. Support for Named Entity Recognition (NER) of data governance entities. It should be noted that unlike many NER tasks the challenge of distinguishing between abstract data governance concepts like the difference between a data defintion asset (e.g. a master data dictionary), a data asset (e.g. a specific database) and a data hosting asset (e.g. a database server) is necessary.
7. Support for NER from noisey and highly informal text-based communication. A Slack channel is unlike the training data for most standard NLP models, which often use newspaper articles.

4 Related Work

In this section we discuss relevant work in the three fields of: knowledge models for data governance, NLP for data governance and NER for Chat communications channels.

4.1 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 [15], Collibras Data Governance Operating Model, INFAI/DBpedias DataID metadata specifications [**?**]to describe data assets, the H2020 ALIGNED project work on knowledge models of data lifecycles and tools [12]. The W3C provenance (PROV) model standard 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 based on

PROV AQ (access and query) 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) [2] could act as basis for describing data value metrics and dimensions.

Thus while there is still a need for an upper governance ontology to glue together many of these individual initiatives there is already a very rich set of RDF-based ontologies and linked data vocabularies available to describe the data governance domain as a knowledge model.

4.2 NLP for Data Governance

There are various ways in which information can be automatically extracted from text. One such way is the NER framework, which aims to identify individual words or phrases in running text that refer to information units such as person names, organisation names, locations, numeric expressions and dates [18]. These information units are commonly referred to as 'entities' in the NLP literature. In the present work, we adapt this framework by training a state-of-the-art NER system on data governance information and actions instead of the above-mentioned traditional entities.

To the best of our knowledge, this is the first usage of an NER approach to extract data governance concepts. In particular, we employ a state-of-the-art machine learning NER method called conditional random fields (CRF) [14]. In addition to being able to extract traditional entities, CRF has been shown to be able to successfully extract other types of information units, such as headers, citations and key phrases from research papers [19, 3], temporal information from clinical records [21], opinions [13], as well as generic, open-ended information units [4]. An advantage of machine learning methods such as CRF (as opposed to rule-based systems) is that the same algorithms can be trained on data from different domains and different languages, provided that adequate annotated training data is available. For example, [16] trained a CRF system to identify multi-word verb units (phrasal verbs, idiomatic expressions, etc.) in 15 languages, performing competitively against bespoke machine-learning systems such as transition systems [1]. Given the success of CRF-based methods in such a diverse array of problems and languages, we consider them to be a good candidate for extracting data governance information, as well.

4.3 NER for Chat

CRF systems studied in the NLP literature are typically trained and used on formal and well-formatted texts such as news articles and academic papers. There has been less attention on informal text, for example, from chat logs. Characteristics inherent to this informal environment such as incomplete sentences, non-standard capitalisation and misspellings due in part to technical limitations in touch screens and predictive text, will generally lead to a loss of accuracy for an NER system.

Nevertheless, there have been several papers investigating the application of NER systems to informal texts, typically on social media. For example, [7] formally investigates what the main sources of error are in extracting entities on tweets using state of the art NER systems, and how these errors could be addressed in the future. This paper 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 noisiness in tweets, but ultimately found that precision and recall scores remained low using NER algorithms developed for use with formal texts. Similarly, [5] explores the use of word representations to improve the effectiveness of a NER in labelling Twitter messages. This work found that general NER systems trained on formal texts performed very poorly in labelling tweets and sought to explore means of improving this performance.

Of note is that while the data governance chat data is informal, messages are generally more formally structured than social media text, with better adherence to sentence structure and correct spelling, so feature extractors reliant on properties present in formal texts are likely to perform better than in [5].

5 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.

In order to create the corpus of annotated data required to train a NER, a dump of chat data from an internal public data governance Slack channel was provided by Collibra. This was first processed by a binary classifier previously developed by Collibra citeshah17 to filter out the messages not related to data governance entities. This left a subset of chat messages suitable for manual annotation.

The chosen subset of data was then formatted for annotation, which included cleaning of some character corruption caused by encoding changes and the separation of entities such as full stops and brackets. BRAT [20], a web annotation tool, was then used to collaboratively annotate the text in accordance with the guidelines we developed (see below).

This annotated dataset was then used for training and evaluating our data governance NER tool. Our NER is based on Stanford NER [10], a part of the Stanford CoreNLP package. To fairly compare different NER configuration results, we set up a standard testing procedure. A Python script was created to comprehensively evaluate the performance of the NER as a whole, as well as a detailed breakdown of its performance for each entity type.

5.1 Knowledge Architecture

The knowledge model used to classify the data governance entities and relationships to be detected in the Slack channel was based upon the Collibra data governance operating model but generalised as an OWL ontology, linked to key data governance ontologies and published as open data. These steps are described below.

Collibra Data Governance Operating Model The Data Governance Operating Model[6] is an open model and has been implemented by hundreds of companies. As depicted in the diagram below, a data governance operating model establishes the foundation for and drives all your data stewardship and data management activities. A model can be subdivided into three categories 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 - through applications such as Business Glossary, Data Catalog, Data Helpdesk, Reference Data Management, and Policy Manager - align and coordinate data management operations. Data Management concerns the integration of the ins and outs of Collibra stewardship activities with third-party applications (such as data profilers, scanners, metadata repositories, etc.) though Collibra Connect or the API.

The Data Governance Center is packaged with a foundational set of about 45 Asset Types from which you can choose to configure your data governance operating model with and is currently used by more than 300 companies. An overview diagram can be found in this footnote[7].

In this work we only extracted subclasses of Asset Types. An Asset is the capital building block in the Data Governance Center. An Asset Type formally defines the semantics of an asset in terms of attribute types and relation types that can be instantiated for it. In other words, it serves as a template. Therefore, all Asset Types are specializations on of five 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'

[6] https://university.collibra.com/courses/introduction-to-the-operating-model-5-x/
[7] book link

– 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'.

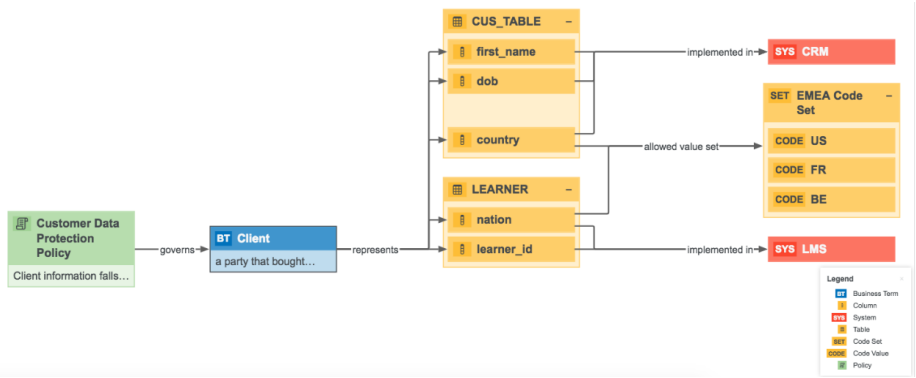

Fig. 2. Rendering of assets and their relations forming a traceability diagram.

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[8].

5.2 NLP/NER Toolchain

Stanford Named Entity Recogniser [9] was the NER tool 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 chat dump was tokenised using the Stanford Tokeniser (part of the Stanford CoreNLP toolkit [17]). In addition, the authors developed Python scripts for additional data pre-processing, conversion, experiment automation and evaluation. We have made these scripts freely available online[9].

The annotation of the chat dump was done through the Brat annotation tool by the authors themselves. Section 6.1 details the annotation scheme.

6 Evaluation

We seek to evaluate the effectiveness of the NER system described in this paper to extract data governance information and actions from a real Slack chat channel. This evaluation focuses on experiments determining the labelling accuracy of the NER system on a representative annotated dataset containing data governance information items. We start in section 6.1 by describing the slack chat dataset and the scheme used to annotate it. In section 6.2 we present the actual experiment protocol. As shall be seen in the results (section 6.3), 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.

6.1 Data Annotation

As previously mentioned, the dataset is a raw dump of messages from a Data Governance team at Collibra's Slack chat. Messages containing any particularly sensitive information were removed before being released, and the resulting data consisted of 7,022 messages totalling about 300,000 tokens. Since a large proportion of these messages were not directly related to data governance, a filtering was performed in order to remove messages not related to data governance. This produced a final dataset of 800 messages, totalling 4,749 tokens.

The entities and relations annotated in the dataset were based on the Asset Types from the Collibra Operation Model detailed in section 5.1. This approach

[8] http://theme-e.adaptcentre.ie/odgov
[9] https://github.com/simonq80/datagovernancener

lead to the initial entity types: `Gov`, `Bus`, `Data`, `Tech` and `Issue`, representing Governance Assets, Business Assets, Data Assets, Technology Assets and Issues respectively. However, upon annotating a sample of the dataset with this scheme, some tokens were found not to fit under any of the entity types defined, but were still considered useful for the NER system to label. To address this, two additional entity types were devised: `Role` was created to label text representing a data governance role, such as an administrator or a domain expert, whilst the `Dmtask` label was created to label text representing a data management task, such as upgrading or backing up a database. As previously mentioned, the actual annotation work was conducted by the authors using the BRAT annotation tool.

Table 1 shows the number of annotated tokens for each data governance information category. Out of the total 4,749 tokens, 3,011 were annotated as non-DG related (i.e. non-entities).

Table 1. Word counts of each Data Governance Information Category

Category	Word Tokens	Word Types	Length Mean	Length Std. Dev.
`Bus`	196	141	1.6752	1.2665
`Data`	503	217	1.7964	1.4459
`Dmtask`	144	93	1.1707	0.5054
`Gov`	182	114	3.7143	3.2262
`Issue`	310	175	3.4444	2.3623
`Role`	14	9	1.5556	0.8315
`Tech`	236	129	1.4937	0.9727
`User`	153	44	3.1875	1.8892
Total	1738	922	1.8242	1.7729

6.2 Experiment 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 evaluation 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. Bias is a general tendency for the resampling method to over or underestimate the performance of a classifier, while variance is degree to which results can vary between runs, typically measured by the standard deviation of the resampling methods estimates over many runs.

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 [11]. 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. Evaluation results for all of these experiment variants are presented in the following section.

6.3 Results

Table 2 shows the results of 4-, 5- and 10-fold cross validation experiments. Means of precision, recall and F-1 scores, along with their standard deviations are given. Although the results are very similar across all three fold variants, there is a slight increase in variance when using 10-fold cross validation across all three scores.

Table 2. Mean and Standard Deviation of performance metrics across 4-, 5- and 10-fold Cross-Validation

k-fold CV	Precision	Recall	F-1
4	0.4516±0.01255	0.2487±0.08224	0.3143±0.06584
5	0.4334±0.04266	0.2444±0.06982	0.3092±0.06380
10	0.4540±0.09512	0.2612±0.09045	0.3262±0.08908

Results for each Data Governance Information category from 4-fold cross-validation can be seen table 3. 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.

Table 3. 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

As mentioned in the introduction to this section, 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 3 plots this correlation for the F-1 measure (precision and recall show a 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.

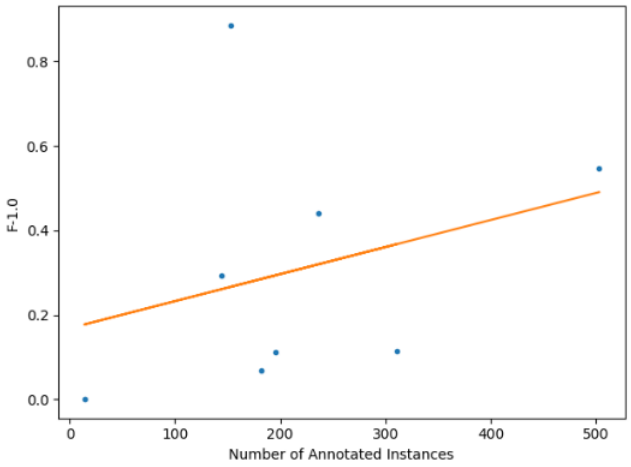

Fig. 3. Correlation between the number of annotated instances of a category and its F-1 score

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 4 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.

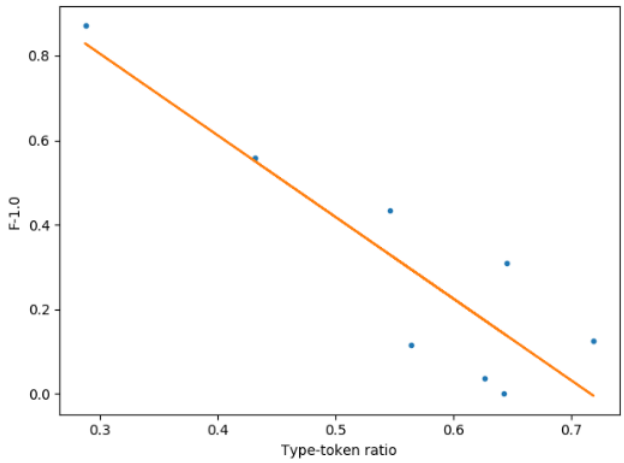

Fig. 4. Correlation between the type-token ratio of a category and its F-1 score

7 Conclusions and Future Work

This paper has demonstrated that CRF-based Named Entity Recognition is a promising approach for extraction of data governance knowledge described in an open ontology. 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. Although the recall scores are disappointing it is our hope that precision is more important for the first planned application as an interactive data governance bot on the Slack channel system of engagement who must minimise their number of incorrect interventions to avoid frustrating the user instead of helping them.

However there is much work to be done if this system is to be deployed in live customer sites. The first is to enlarge the training and test dataset, the second is to explore alternative feature extraction approaches - both statistical and neural-based approaches are under evaluation but of course much more training data would be required to train a neural approach. One key to this may lay in the system of record, where the structured model of the enterprise could be mined for vector information in a way analogous to the rdf2vec approach to ontology vectorisation.

References

1. Al Saied, H., Constant, M., Candito, M.: The ATILF-LLF system for Parseme Shared Task: a transition-based verbal multiword expression tagger. In: Proceedings of the 13th Workshop on Multiword Expressions . pp. 127–132. MWE '17, Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/W17-1717

2. Attard, J., Brennan, R.: A semantic data value vocabulary supporting data value assessment and measurement integration. In: Proceedings of the 20th International Conference on Enterprise Information Systems - Volume 2: ICEIS,. pp. 133–144. INSTICC, SciTePress (2018). https://doi.org/10.5220/0006777701330144
3. 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)
4. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: IJCAI. vol. 7, pp. 2670–2676 (2007)
5. Cherry, C., Guo, H.: The unreasonable effectiveness of word representations for twitter named entity recognition. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 735–745 (2015)
6. 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 the International Semantic Web Conference (2009)
7. 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)
8. Falkenberg, E., Hesse, W., Lindgreen, P., Nilsson, B., Han Oei, J., Rolland, C., Stamper, R., van Assche, F., Verrijn-Stuart, A., Voss, K.: FRISCO: A framework of information system concepts : The FRISCO report (WEB edition). International Federation for Information Processing (IFIP) (1998)
9. Finkel, J.R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005). pp. 363 – 370. Ann Arbour, MI (2005). https://doi.org/10.3115/1219840.1219885
10. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics. pp. 363–370. Association for Computational Linguistics (2005)
11. Forman, G., Scholz, M.: Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter **12**(1), 49–57 (2010)
12. Gavin, O., Kontokostas, D., Koller, A., Davies, J., Francois, P., Marciniak, A., Bozic, B., Mendel-Gleason, G., Feeney, K., Brennan, R.: The aligned project aligned, quality-centric software and data engineering driven by semantics. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Paolo Ponzetto, S., Lange, C. (eds.) Project Networking Session at ESWC 2016 THE SEMANTIC WEB. LATEST ADVANCES AND NEW DOMAINS. LNCS, vol. 9678. Springer (2016), http://www.tara.tcd.ie/handle/2262/76242
13. Jakob, N., Gurevych, I.: Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In: Proceedings of the 2010 conference on empirical methods in natural language processing. pp. 1035–1045. Association for Computational Linguistics (2010)
14. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289 (2001). https://doi.org/10.1038/nprot.2006.61

15. 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., http://www.w3.org/TR/prov-o/
16. 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: Proceedings of The 13th Workshop on Multiword Expressions. pp. 114–120. Valencia (2017)
17. 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)
18. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Linguisticae Investigationes **30**(1), 3–26. (2007). https://doi.org/10.1075/li.30.1.03nad
19. Peng, F., McCallum, A.: Information extraction from research papers using conditional random fields. Information Processing and Management **42**(4), 963 – 979 (2006). https://doi.org/https://doi.org/10.1016/j.ipm.2005.09.002
20. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: a web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. Association for Computational Linguistics (2012)
21. Wang, W., Kreimeyer, K., Woo, E.J., Ball, R., Foster, M., Pandey, A., Scott, J., Botsis, T.: A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. Journal of Biomedical Informatics **62**, 78 – 89 (2016). https://doi.org/https://doi.org/10.1016/j.jbi.2016.06.006