Iterative Approach for Information Extraction and Ontology Learning from Textual Aviation Safety Reports

Lama Saeeda(✉)

Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
saeeda.lama@fel.cvut.cz

Abstract. Textual aviation safety reports are one of the main resources that contain valuable information to understand incidents and accidents in a high-risk industry such as the aviation domain. The reporting process, hence, is essential to provide these reports. Most of the time, the reporting process is done manually, and typically, poorly structured data are provided by the reporters. Automated content analysis for these reports has attracted researchers to extract the required information to perform many tasks, and they used several techniques to achieve it. Ontologies provide formal and explicit specifications of conceptualizations and play a crucial role in the information extraction process. In this paper, we propose a novel iterative ontology-based approach of information extraction and semantic annotations for aviation safety reports and augmenting back the aviation safety ontology with new concepts and relations depending on the terms already annotated in the discovered report model.

Keywords: Information extraction · Domain ontology · Ontology learning · Safety reports

1 Introduction

Aviation safety reports play a crucial role in data-driven safety oversight in the aviation safety field. Although the content of the reports is typically highly informative, its transformation into structured form (e.g. by means of dedicated reporting forms) is lossy and imprecise which negatively influences their potential for proper safety analyses.

Initial incident and accident reports are the best source of information for extracting the most important knowledge to feed the preliminary[1] reports' building process [20]. One of the main tasks in the process of building the preliminary report is to detect the type of event described in the initial aviation safety report so that an appropriate form can be displayed to the user to capture additional required information regarding the specific accident or incident.

[1] A preliminary report is created by the safety department of an organization and sent to the authority.

E. Blomqvist et al. (Eds.): ESWC 2017, Part II, LNCS 10250, pp. 236–245, 2017.
DOI: 10.1007/978-3-319-58451-5_18

Text processing for such reports is essential for simplification of the safety reporting process. Many techniques have been proposed for the purpose of classification, cause identification, and knowledge extraction from the aviation safety reports. Most of these techniques are mainly following linguistic or statistical methods. Semantic web technologies provide the advancement in information systems by assigning semantics to information by means of shared formal ontologies. These ontologies are also used as the main resource for semantic annotation authoring. Ontologies are especially useful because they support the exchange and sharing of information as well as reasoning tasks [14]. An ontology containing a class hierarchy relevant to the aviation safety domain can simplify the reporting process by enabling the reporter to process the data in a controlled way by the mean of the ontology, and hence, a more relevant and accurate data will be provided by the reporter. This will assure a better experience for the safety management of the statistics business intelligent (BI) user, who will benefit from the targeted, without noise, and less biased statistics, which will improve the quality of the data, the speed of the reporting process on the general level and provide more precise results of the BI. However, using such ontologies depends directly on the availability of this ontology in the target domain. The domain ontology construction process consumes a lot of time and resources and requires a lot of efforts by knowledge engineers and human experts. Researchers have been discussing the process of automatically as well as semi-automatically building ontologies. Ontology learning from the textual corpus is the set of methods and techniques used for building an ontology from scratch, enriching, or adapting an existing ontology in a semi-automatic fashion using several knowledge and information sources [16].

In this paper, we propose to use a knowledge-based approach for extracting useful information from the aviation safety reports, as well as we propose a methodology for semi-automated ontology learning for the aviation safety ontology (Fig. 1).

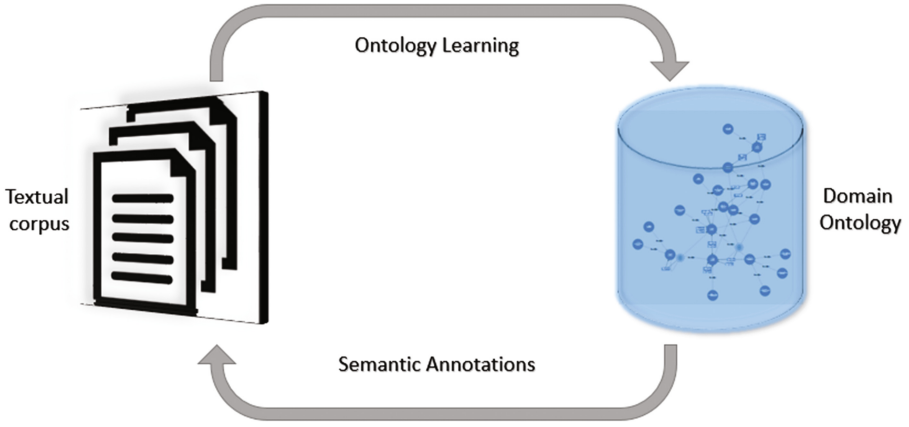

Fig. 1. The iterative approach

2 State of the Art

In the recent years, researchers have paid increasing attention to automatically analyzing textual safety reports in different domains.

For example, **transportation domain**, like analyzing maritime accident investigation reports [8], where text mining methods were applied to extract causal relations from maritime accident investigation reports collected from the Marine Accident Investigation Branch, and railroad accident investigation reports [3], where accident reports were analyzed using the text mining techniques of probabilistic topic modeling and k-means clustering to identify the recurring themes in major railroad accidents. Also, the performance of four machine learning paradigms applied to modeling the severity of injury that occurred during traffic accidents were performed in [17].

Also, safety reports in aviation domain were carefully studied. In [1] the authors presented a method of automatic Aviation Safety Reporting System (ASRS) shaping factor classification based on the most relevant words from a subjectivity lexicon, and [7] where the problem of cause identification from aviation safety reports was introduced to the NLP community as a multi-class, multi-label text classification task, and a bootstrapping algorithm was presented that automatically augments a training set by learning from a small amount of labeled data and a large amount of unlabeled data. Also, a survey of different NLP techniques designed and used to manage and analyze aviation incident reports was performed in [2].

Safety reports were also studied in other domains, such as **construction safety domain**, where in [5] the methodology consists of pre-processing the accident reports and weighting terms in order to apply a data-driven unsupervised K-Means-based clustering approach, in order to classify the collected reports in four clusters, each reporting a type of accident. and in [4], a natural language processing system based on hand-coded rules and dictionaries of keywords was proposed to extract precursors and outcomes from unstructured injury reports.

In [9] it is noticed that the authors analyzed safety reports in **medical safety domain**, where they employed natural language processing and network analysis to identify effective categories of medical incident reports.

In most of these research work, researchers have developed/adapted several techniques, such as natural language processing, machine learning, and rule-based techniques, without taking into consideration the idea of combining semantic web technologies in their strategies.

A survey of ontology-based approaches of semantic data mining was performed by [6], that investigate why ontology has the potential to help semantic data mining, and how formal semantics in ontologies can be incorporated into the data mining process, showing the advantages in performing data mining task that is not achievable with traditional data mining methods. A similar survey was performed in [11] to provide an introduction to ontology-based information extraction and to review the details of different OBIE systems developed.

While [10] focuses only on the terms mapping to ontology's concepts, performing simple NLP pre-processing and string matching to the labels,

other studies made use of the hierarchy (structural nature) of the ontology, and the use of the relations between the extracted terms as in [22,23].

Using domain-ontologies relies on the availability of this ontology in the domain of study. However, the automatic ontology construction is not a trivial task and requires lots of human intervention in some stages of ontology construction. There are various approaches and tools available for automatic construction of ontology from a textual corpus. In [13] they Focus on presenting a method for learning axioms from text based on named entity recognition. [14] describes a new ontology learning approach that consists of a method for the acquisition of concepts and its corresponding taxonomic relations, where also axioms disjointWith and equivalentClass are learned from text without human intervention. [15] focuses on identifying the relationships between medical concepts as defined by the REMed (Relation Extraction from Medical documents) solution that aims at finding the patterns that lead to the classification of concept pairs into concept-to-concept relations.

Creating large annotated textual corpus for the training and evaluating tasks is, however, prohibitively expensive. To mitigate this problem, the notion of supplementing labeled data with features derived from large amounts of unlabeled data has been explored in [12].

3 Problem Statement and Contributions

Reporting process of aviation safety incidents and accidents needs to be done clearly and easily. To achieve that, a reporting tool has been built on the top of the aviation safety ontology in [18]. In order to make reporting process more user-friendly, as well as make it easy and logical, a smart form generation based on the event-type and other attributes is needed, in order to support the reporting process by reducing the list of attributes that have to be filled, only to those related and relevant for a specific event type. In order to detect the event-type in the initial safety report, a comprehensive textual analysis process has to be performed, taking into consideration the unstructured nature of the initial input report, which is usually full of jargon.

So far, our work has focused on text annotations. It has achieved high precision semantic annotations for aviation safety reports, detecting the main event and event-type in the safety report based on the aviation safety ontology as well as all the participants, temporal and spatial information, including all the information that can help to construct a dynamic form suited to the actual report. However, a large space of improvement can be done regarding the detection of the terms which have not been yet introduced in the ontology, based on the iterative method of detection and learning. The full scenario is explained in the Fig. 2.

To overcome these problems, we are planning to proceed to perform several tasks, including indicate a more accurate mapping to the corresponding concept in the ontology on the level of concepts, as well as to make use of the ontology hierarchy to detect the relations between the concepts for the better understanding of the aviation safety reports. Also, making use of the already-detected

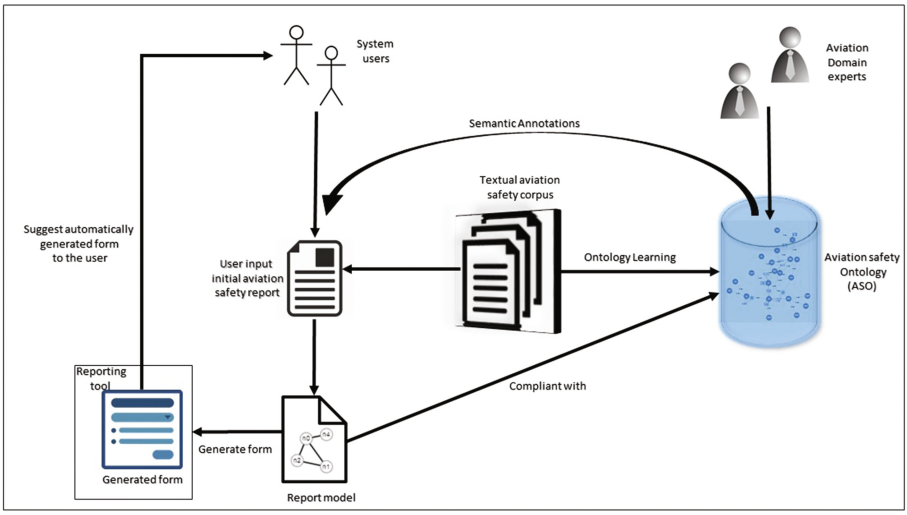

Fig. 2. Full scenario of the iterative approach of annotating and learning back the ontology

concepts for augmenting the ontology with new concepts and adding them in the proper position in the ontology hierarchy with a context-based approach.

Achieving these tasks will help to investigate several research questions as following:

- What is the impact of utilizing the domain-specific ontology on classification performance?
- How this iterative approach of recognition of the terms and augmenting back the ontology will improve the background knowledge and hence, the event type detection.

4 Research Methodology and Approach

We aim to improve the results of semantic annotations in our previous work [20]. This includes improving annotations of the terms, disambiguating them, and detecting the relations and the facts. This will be done by optimizing the pipeline described in Fig. 3, tuning the parameters of the tools and exploring new algorithms to be included in the pipeline.

In Fig. 2 we are showing the full scenario. Following, the most important relevant components are explained:

Semantic annotations: We are currently developing the Aviation Safety Text Analyzing Tool. Entity recognition tools that allow the usage of custom background knowledge inside were tested separately, then the most accurate tools that served our purpose the best were chosen. We combined these tools in one

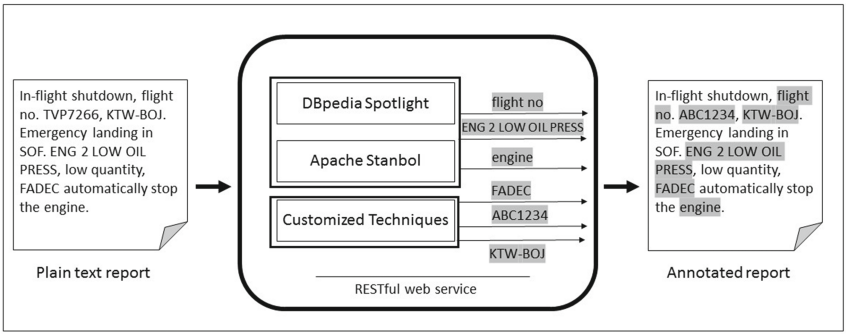

Fig. 3. Text annotation processing pipeline

pipeline to cooperate together. As shown in Fig. 3, Apache Stanbol[2] is one of the tools that provides the ability to work with custom vocabularies and creating custom indexes upon it. It also comes with a list of enhancement engines implementations, with the ability to build a specific one to get the most benefit out of the tool. This allowed us to build a chain of enhancement engines that fits perfectly to the aviation-safety concepts detection. DBpedia Spotlight[3] is another entity recognition tool that offers to create a spotlight model on the user's own server to model the occurrences of resources with the context in which they are mentioned. We are also taking into consideration the entities that are not possible for the current tools to detect, in spite of their ability to detect mentions from a specific terminology. For this complication that stems directly from the nature of the aviation domain, such as callsigns, registration marks, flight numbers, airport names abbreviations, etc., we are using special techniques for every case as discussed in [20]. The output of Apache Stanbol, DBpedia spotlight and the techniques for the special terms were parsed, merged, and optimized in a RESTful web service and output the mentions being detected with their proper mapping to the ontology hierarchy.

The clear line we are walking through now is the relationships recognition. Taking into consideration the ontology subclasses and subproperties, the detected entities that are mapped to the ontology, and the corpus as an input, then using the relations between the subclasses to detect the relationships and the facts between the entities.

Ontology learning: In [21], researchers proposed learning from structured knowledge (XML documents of card design). They used some ontological refinements. They designed an ontology on the top of the documents, then based on the data, the ontology was used to refine the learning task so they can learn according to the hierarchy. In this work, we have a more difficult task because of the unstructured nature of the input. Based on the entities and relations

[2] https://stanbol.apache.org.
[3] https://github.com/dbpedia-spotlight/dbpedia-spotlight.

between the entities that are detected, we expect to reconstruct the model trying to detect additional entities and relations to learn the ontology. In this case, the refinement operators for the subsumption classes should be a good solution because of the huge hierarchies in the Aviation Safety Ontology (ASO).

In our approach we will rely on the nature of the **Aviation Safety Ontology (ASO)**, which is designed based on domain terminology found in aviation safety standards and manuals, as well as safety data found in incident and audit reports. In [19], researchers combine the domain terminology with a relevant and well-defined state of art conceptualizations. The ASO was built on the top of the Unified Foundational Ontology (UFO), which is one of the top-level ontologies that has a good modeling language and it is supported by useful tools. UFO presents a level of abstraction that forms a perfect point to start with the annotation and learning process. Then, as mentioned in [19], during the analysis of the aviation safety domain, several modules and relationships among them could be identified. For example, the Aviation Safety Core (ASC) ontology which contains the Aviation ontology, that consists of the common aviation domain terms, such as Aircraft and Flight. and Safety ontology, which consists of the fundamental conceptualizations necessary for the management of safety information. This way of design can lead to a big potential of improvement in semantic annotations of the aviation safety reports, and potentially, makes the ontology learning process accurate and very specific task.

This iterative text annotating and ontology learning methodology should guarantee more reliable understanding for the new reports that need to be processed, as well as enriching the Aviation Safety Ontology with new terms and relations, and supporting the complicated and expensive process of building the domain ontology.

We will also take into consideration the nature of the targeted **domain corpus** and the characteristics and the structure of the aviation safety reports, (i.e. common abbreviations, registration marks, pieces of controlled language, event chains, etc.).

5 Preliminary Results

To this end, different Linked Data Knowledge Extraction tools with respect to a domain-specific vocabulary had been tested, then we chose the tools that allow the best results of entity recognition, combining them into one pipeline, and making them working together, as well as with other features that we added, taking into consideration some very specific terms and abbreviations used in the aviation field.

We built a tool that integrates several techniques inside in order to provide high precision reports' annotations in aviation safety domain in order to be used directly in practice. In our preliminary results, the precision scores high rates in most of the cases, it even reaches to 100% rate for some reports. On the other hand, the recall scores low rates. You can check the preliminary results of comparison of ontology-based entity recognition to the non-ontological approach of samples of reports in Table 1 and Sect. 4 in [20].

Currently, we are facing the problem of improving the achievement of reliable entity recognition, with the iterative methodology which will, potentially, improve the recall in a significant percentage.

Also, relations detection between the concepts, and the results for the new potential concepts for semi-automatic ontology learning. Our target now is to proceed with the iterative methodology that we discussed before, of ontology learning and augmenting the ontology with new terms, and relations between the terms, in a graph building process within a context-based approach, and to measure how much improvement will be gained with the semantic annotations.

6 Evaluation Plan

Due to a noticeable fact in the aviation safety domain, many public safety reports are available. However, most of these reports are unannotated or poorly automatically annotated, while a very few reports are actually well annotated. This makes the process of corpus construction a very hard task for the evaluation process, that requires extensive time and effort of the experts.

As mentioned in [20], we are creating a high quality, very precise gold standard corpus out of, mainly, initial aviation safety reports taken from different authorities' resources, for example, UZPLN[4], where they have their public aviation investigation incidents and accidents reports. Also, the corpus contains confidential data provided by the partners of the current project that this work is included in.

Experts in aviation domain manually annotated domain terms (entities) in each report with respect to the Aviation Safety Ontology (ASO) that mentioned earlier. Technically, they used the General Architecture for Text Engineering (GATE) tool[5].

To this end, this corpus consists of 80 high quality annotated documents that will hopefully grow fast through time. We need this kind of corpus for the evaluation process of the annotation pipeline using the well-known recall, precession and F1 score metrics. Regarding the ontology learning evaluation task, we will only be able to evaluate the algorithms and the methods of the generated concepts and relations but not the quality of the ontology. Creating ontologies automatically doesn't guarantee the quality of these ontologies. We want the ontology that we are creating to keep some reliable level of quality. In order to achieve that, aviation ontology is revised by experts who understand the problem. This way experts will not be forced to process or to read every single report. Instead, they will be offered some typical notions, relationships, and data patterns that they can generalize into a logical knowledge.

7 Conclusions

In this proposal, a novel iterative methodology of semantic annotations and ontology learning has been presented, in order to improve the understanding

[4] http://www.uzpln.cz/.
[5] https://gate.ac.uk/.

of the huge jargon unstructured incidents and accidents safety reports in the aviation domain, and supporting the smart form generation for easier and more efficient reporting process.

Acknowledgments. I want to thank Dr. Petr Křemen for his support in accomplishing this proposal. Furthermore, this work was partially supported by grants No. TA04030465 Research and development of progressive methods for measuring aviation organization's safety performance of the Technology Agency of the Czech Republic, No. SGS16/229/OHK3/3T/13 Supporting ontological data quality in information systems of the Czech Technical University in Prague.

References

1. Switzer, J., Khan, L., Bin Muhaya, F.: Subjectivity classification and analysis of the ASRS corpus. In: 2011 IEEE International Conference on Information Reuse & Integration (2011)
2. Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., Raynal, C.: Natural language processing for aviation safety reports: from classification to interactive analysis. Comput. Ind. **78**, 80–95 (2016)
3. Williams, T., Betak, J., Findley, B.: Text mining analysis of railroad accident investigation reports. In: 2016 Joint Rail Conference (2016)
4. Tixier, A.J.-P., Hallowell, M.R., Rajagopalan, B., Bowman, D.: Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports. Autom. Constr. **62**, 45–56 (2016)
5. Chokor, A., Naganathan, H., Chong, W.K., Asmar, M.E.: Analyzing Arizona OSHA injury reports using unsupervised machine learning. Procedia Eng. **145**, 1588–1593 (2016)
6. Dou, D., Wang, H., Liu, H.: Semantic data mining: a survey of ontology-based approaches. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015) (2015)
7. Persing, I., Ng, V.: Semi-supervised cause identification from aviation safety reports. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL-IJCNLP 2009, vol. 2 (2009)
8. Tirunagari, S.: Data mining of causal relations from text: analysing maritime accident investigation reports. arXiv preprint arXiv:1507.02447 (2015)
9. Fujita, K., Akiyama, M., Park, K., (Nakagami) Yamaguchi, E., Furukawa, H.: Linguistic analysis of large-scale medical incident reports for patient safety. In: MIE, pp. 250–254 (2012)
10. Sfakianaki, P., Koumakis, L., Sfakianakis, S., Iatraki, G., Zacharioudakis, G., Graf, N., Marias, K., Tsiknakis, M.: Semantic biomedical resource discovery: a natural language processing framework. BMC Med. Inform. Decis. Mak. **15** (2015)
11. Wimalasuriya, D.C., Dou, D.: Ontology-based information extraction: an introduction and a survey of current approaches. J. Inf. Sci. **36**, 306–323 (2010)
12. Henriksson, A., Kvist, M., Dalianis, H., Duneld, M.: Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J. Biomed. Inf. **57**, 333–349 (2015)

13. Rios-Alvarado, A., Lopez-Arevalo, I.: Ontology learning from text: method for learning axioms. Technical report (2012)
14. Rios-Alvarado, A.B., Lopez-Arevalo, I., Tello-Leal, E., Sosa-Sosa, V.J.: An approach for learning expressive ontologies in medical domain. J. Med. Syst. **39** (2015)
15. Barbantan, I., Porumb, M., Lemnaru, C., Potolea, R.: Feature engineered relation extraction - medical documents setting. Int. J. Web Inf. Syst. **12**, 336–358 (2016)
16. David Sanchez, R.: Domain ontology learning from the web. Ph.D. thesis, Universitat Politecnicade Catalunya (2007)
17. Chong, M., Abraham, A., Paprzycki, M.: Traffic accident analysis using machine learning paradigms. Informatica **29**(1) (2005)
18. Vittek, P., Lališ, A., Stojic, S., Plos, V.: Challenges of implementation and practical deployment of aviation safety knowledge management software. In: Communications in Computer and Information Science Knowledge Engineering and Semantic Web, pp. 316–327 (2016)
19. Kostov, B., Ahmad, J., Křemen, P.: Towards ontology-based safety information management in the aviation industry. In: 13th International Conference, IESD (2016)
20. Saeeda, L., Křemen, P.: Text analyzing of aviation safety reports. WIKT & Data a Znalosti (2016)
21. Žáková, M., Železný, F., Garcia-Sedano, J.A., Tissot, C.M., Lavrač, N., Křemen, P., Molina, J.: Relational data mining applied to virtual engineering of product designs. In: Muggleton, S., Otero, R., Tamaddoni-Nezhad, A. (eds.) ILP 2006. LNCS (LNAI), vol. 4455, pp. 439–453. Springer, Heidelberg (2007). doi:10.1007/978-3-540-73847-3_39
22. Wong, M.K., et al.: A multi-phase correlation search framework for mining non-taxonomic relations from unstructured text. Knowl. Inf. Syst. **38**(3), 641–667 (2012)
23. Reyes, J.A., Montes, A.: Learning discourse relations from news reports: an event-driven approach. IEEE Lat. Am. Trans. **14**(1), 356–363 (2016)