Automatic Semantic Annotation: Towards Resolution of WFIO Incompatibilities

Chahrazed Tarabet, Meriem Mouzai and Ali Abbassene
Centre de Développement des Technologies Avancées, Baba Hassen, Algiers, Algeria

Keywords: IOWF, NLP, Automatic Semantic Annotation, Ontology.

Abstract: Inter-organizational workflows (IOWF) allow for orchestration of processes between different organizations, but the incompatibility they reveal poses a serious problem. Nevertheless, there are approaches that can remedy this problem, notably the semantic annotation. In this position paper, we will present a study whose objective is to address the detection and correction of these incompatibilities between workflow partners. For this purpose, amelioration, optimization and automation are necessary for the semantic annotation phase of inter-organizational workflows, in order to achieve the IOWF incompatibility resolution.

1 INTRODUCTION

An organization is a coordinating unit, with identifiable boundaries, working to achieve a goal shared by its participating members. Nowadays, the company is no longer industrial but commercial, it is no longer located in a single place, it is extended, it even happens not to be fully visible but to be (in part) virtual. These companies need inter-organizational cooperative systems, that is, across multiple network organizations. In this context, information and communication technologies play an essential role in enabling enterprises to exchange all types of data.

The inter-organizational workflows go in the same direction, allowing an orchestration of processes between several organizations. However, the incompatibility of the latter poses a serious problem (Abbassene, Alimazighi and Aouachria, 2015). Nevertheless, there are approaches that can remedy this problem, notably the semantic annotation. This approach represents a mechanism for linking a data to its semantic description represented by a concept derived from an ontology. It is an effective way to detect and correct these incompatibilities between workflow partners. For this purpose, amelioration, optimization and automation are necessary for the semantic annotation phase of inter-organizational workflows.

In this position paper, we will present a critical review of the works that deal with the automation of semantic annotation phase, using techniques such as NLP (Natural Language Processing). Thus, we propose a solution to the problematic posed according to the results obtained.

2 LITERATURE REVIEW

In this section, we will present the works of the semantic annotation domain found in the literature that we judge interesting.

Authors in (Davis et al., 2009) use Controlled Natural Languages (CNLs) which are subsets of natural language whose grammars and dictionaries have been restricted in order to reduce or eliminate both ambiguity and complexity.

These languages prompt the novice user to annotate, while simultaneously creating, their respective documents in a user-friendly way, while protecting them from the formalisms of representation of the complex underlying knowledge. CNLs have already been applied successfully in the context of authoring ontology, but very little research has focused on CNLs for semantic annotation. They describe here a user-friendly semantic annotator, based on CLIE (Controlled Language for Information Extraction) tools.

However, this annotator only allows non-expert users to write and annotate minutes of meetings and semi-automatic status reports using controlled natural language.

Also, this other work (semantator) (Tao et al.,

Tarabet, C., Mouzai, M. and Abbassene, A.
Automatic Semantic Annotation: Towards Resolution of WFIO Incompatibilities.
DOI: 10.5220/0006377505560562
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 2*, pages 556-562
ISBN: 978-989-758-248-6

2013), which is an annotation tool that allows to annotate documents with semantic Web ontologies using a loaded free text document and an ontology, users can annotate document fragments with classes in the ontology to create instances and link instances created with ontology properties. Thus, it provides:

1) Basic manual annotation features: creation / deletion of ontology instance, creation / deletion of relationship, binding of equivalent instances and export / reload of existing annotations;

2) Automatic annotation by connecting to the annotator NCBO and cTAKES;

3) Basic reasoning support based on the underlying semantics of the preachers owl: disjointWith and owl: equivalentClass.

In contrast, this tool only allows semi-automatic annotation and focuses on clinical documents.

The main objective of the works (Kiyavitskaya et al., 2006) and (Kiyavitskaya et al., 2005) is to present a methodology supported by tools that semi-automates the semantic annotation process for a set of documents in relation to a semantic model (ontology or conceptual schema). It is proposed to address the problem using highly efficient and proven methods and tools in the field of software analysis for processing billions of source code lines of legacy software.

The semantic annotation method of documents uses generalized syntactic analysis and the TXL structural transformation system, the basis of the LS / 2000 (Kiyavitskaya et al., 2006) automated system. TXL is a programming language specially designed to allow, for example, rapid prototyping of language descriptions, tools and applications. The system accepts as input a grammar and a document, generates an analysis tree for the input document and applies transformation rules to generate an output in a target format (Kiyavitskaya et al., 2006), this transformation phase is not yet implemented in the work (Kiyavitskaya et al., 2005).

The disadvantage of these approaches is that it deals only with the semi-automatic annotation, so the results of the document (Kiyavitskaya et al., 2005) are adopted only in the tourism sector.

Antti Vehvilainen in this work (Vehvilainen, Hyvonen and Alm, 2008) dealt with applications of semantic Web technologies to help office services. They focus on QA (Questions / Answers) support services, where the service database is composed of answers to the previous questions, that is, QA pairs. They propose a semi-automatic semantic annotation of natural language text for the question-answer (QA) pairs annotation and case-based reasoning techniques to find similar questions. The methodology consists of using semantic Web technologies in content annotation, using the QA repository and integrating the information available online on the Web with the creation process and responses. They consider here the usefulness of CBR (Case-Based Reasonning) in the indexing and the extraction of informations since the similar pairs of QAs reproduce in the services of QA. Case-based reasoning (CBR) is a problem-solving paradigm in artificial intelligence where new problems are solved based on previously experienced similar problems. The CBR cycle consists of four phases:

1) Retrieve the most similar case (s),

2) Reuse the recovered cases to solve the problem,

3) Revise the proposed solution, and

4) Retain the solution as a new case in the base of cases.

Nevertheless, this approach is not generalized in all domains and only allows the semi-automatic annotation.

They treat here (Qacim and Salih, 213), the semantic similarity between sentences based on WordNet semantic dictionary. The proposed algorithm will be based on a number of resources, including Ontology and WordNet.

The goal of this research (Qacim and Salih, 213) is to create an efficient automatic annotation platform that develops a way to automatically generate metadata to semantically annotate Web documents that improve information retrieval. The proposed system should be easily understood by non-technical users who may not be familiar with the technical language used to create ontologies. The proposed system provides ontological similarity to determine the relationships between words in sentences and concepts in ontology. It has been found that the meaning of the term similarity is ambiguous because of its use in many different contexts, such as biological, logical, statistical, taxonomic, psychological, semantic, and many other contexts, to resolve ambiguities, WordNet Must be used to provide a lexical ontology.

The semi-automated annotation of texts in natural language was approached in this work (Erdmann et al., 2000) by designing an information extraction-based approach for semi-automated annotation, which was implemented on SMES (Saarbrucken Message Extraction System), (Erdmann et al., 2000) which includes a tokenizer based on regular expressions. It is a generic component that respects several principles that are crucial to its objectives. (i) it is fast and robust, (ii) it maps terms to ontological concepts, (iii) produces dependency relationships between terms, and (iv) is easily adaptable to new domains. As in this approach, finite state technologies support lexical acquisition as well as semantic marking. The goal of the global process is the generation of so-called lexical networks that can be used to enable automatic and semi-automatic construction of texts on the Web. Incoming documents are processed using the SMES information retrieval system. SMES associates simple words or complex expressions with a concept of ontology, linked by the domain lexical link. Recognized concepts and relationships between concepts are underlined as suggested annotations. This mechanism has the advantage that all relevant informations in the ontology document are recognized and proposed to the annotator, but it's applied only on text in natural language and is not fully automated.

Aurélie Névéol in this article (Névéol, Doğan and Lu, 2011) dealt with the production of annotated data which is a necessary step for many natural language processing (NLP) or information processing tasks.

They have studied the semantic annotation of a large number of biomedical requests. This study shows that automatic pre-annotations are considered useful by most annotators to speed up the annotation rate and improve the consistency of the annotations while maintaining a high quality of the final annotations. The disadvantage of this work is that its field of application is restricted to the biomedical domain.

This work (Dingli, Ciravegna and Wilks, 2003) proposes a methodology to learn to automatically annotate specific domain information from large repositories with minimal user intervention. The methodology is based on a combination of information retrieval, information integration and automatic learning. Learning is sown by extracting information from structured sources. The retrieved information is then used to partially annotate documents. These annotated documents are used for learning bootstrap for simple information extraction (IE). It will be used to form more complex IE engines and the cycle will continue to repeat until the required information is obtained. User intervention is limited to providing an initial URL and correcting information if this is the case when the calculation is completed. The revised annotation can then be re-used to provide additional training and thus obtain more information and / or more precision.

The methodology was fully implemented in Armadillo, a system for extracting and integrating unsupervised data from large collections of documents.

This other work (Kiryakov et al., 2004) seeks to create an efficient, robust and scalable architecture for automatic semantic annotation, and to implement this architecture in a component-based platform for indexing and semantic search on large collections of documents. What is considered a primary innovative contribution, is the fact that it offers an end-to-end extensible system that processes the complete cycle of creating metadata, storing and semantic search, for online use that provide navigation semantically improved. Another approach presented in (Tulasi et al., 2017) deals with automatic semantic annotation based on ontologies, where all documents are collected from the Web and a database is created. The documents are then given as an input to make a semantic annotation on the ontology.

In this paper (Kiryakov et al., 2004) authors present a holistic architecture view for automatic semantic annotation with references to classes in ontology and instances, on the basis of these semantic annotations, it has indexed and retrieved documents. A system (called KIM), implementing this concept (Popov et al., 2003), provides a new infrastructure and knowledge and information management services for automated semantic annotation, indexing and document retrieval, it provides also a mature infrastructure for scalable and customizable information extraction (IE) and annotation and document management, based on GATE.

GATE (General Architecture for Text Engineering) (Ranganathan, Biletskiy and Kaltchenko, 2008), developed by the University of Sheffield, is an efficient tool used to perform some Natural Language Processing (NLP) operations. It has many features, such as manual annotation, automatic annotation using a variety of nomenclatures, information retrieval, ontology-based processing, and so on. GATE has played a major role in text annotation, which can be presented in different formats, such as: eml, mail, text, dhtm, pdf, xhtml, xml, rtf, txt, sgm, Sgml, htm, etc. Annotations are mainly performed to communicate semantics in electronic documents and / or underlining text that provides better human understanding.

From a technical point of view, the platform allows KIM-based applications to use it for automatic semantic annotation, retrieval of content based on semantic restrictions, and querying and modifying ontologies and underlying knowledge bases.

In this article (Leopold et al., 2015), they present an approach to automatically annotate process models with the concepts of a taxonomy. At this point, the focus is on business-based taxonomies, such as the Supply Chain Operations Reference Model (SCOR), the MIT Process Manual, and the Process Classification Framework (PCF). To do this, they propose an approach that combines the measurement of semantic similarity with probabilistic optimization. In particular, they use different types of similarity between the process model and the taxonomy and the distance between the concepts of taxonomy to guide the matching with a formalization of the Markov logic.

The semantic annotation is also used in the S-CREAM project. The approach is interesting for the strong implication of the automatic learning techniques for an automatic extraction of the relations between the annotated entities (Handschuh, Staab and Ciravegna, 2002).

A similar approach is also taken in the MnM project, where semantic annotations can be placed online in the content of the document and refer to an ontology and a KB (WebOnto) server, accessible via an API (Vargas-Vera et al., 2002).

The QuizRDF module, used to provide improvements to a standard full-text search with the metadata extracted from OntoBuilder. QuizRDF is an important source of inspiration for the design of KIM. An interesting indexing of the named entity and a system of questions / answers. Once indexed, the content is queried via NL questions, with the NE mark on the question used to determine the type of response expected (Davies, Fensel and Van Harmelen, 2003).

AeroDAML takes a similar approach to KIM, but implements it as a prototype of research on a much smaller scale (Kogut and Holmes, 2001).

SemTag is a platform closer in terms of objectives and architecture to KIM, which performs the semantic annotation on a large scale compared to the TAP ontology which is very similar in size and structure to the KIM and KB ontologies. SEMTAGS gradually creates a first-order markov model based on existing annotations and proposes a semantic annotation, new syntactic trees. It first performs a search phase, annotating all possible references to instances of the TAP ontology. In the second phase of disambiguation, SemTag uses a vector space model to assign the correct ontological class or to determine that this statement does not correspond to a class in TAP. Disambiguation is performed by comparing the context of the current statement (10 words on the left and 10 on the right) with the case contexts in TAP with compatible aliases (Dill et al., 2003) (Guha and McCool, 2001).

With regard to semi-automatic semantic annotation mechanisms, Pustejovsky describes the approach for semantic indexing and typed hyperlinks (Pustejovsky et al., 1997).

Another approach to semantic annotation of data has improved the retrieval of information and improved interoperability where a new approach based on NLP ontology has been proposed and applied on annual reports (Wang et al., 2009).

The results presented in this section are the result of a detailed study on the state of the art on the techniques and approaches used for the improvement, optimization and automation of the semantic annotation phase.

We have synthesized the results obtained in the following table:

Table 1: Comparison of the presented works.

	Semi-automatic	Automatic	Application	Approach used	Suitable for WFIO
(Davis et al., 2009)	yes	no	Administration (meeting minutes - status report)	CNL	no
(Tao et al., 2013)	yes	no	Clinic	Ontology / NLP	no
(Kiyavitskaya et al., 2006) (Kiyavitskaya et al., 2005)	yes	no	Tourism	Ontology / TXL	no
(Vehvilainen, Hyvonen and Alm, 2008)	yes	no	Help Desk Service	CBR	no
(Qacim and Salih, 213)	-	yes	Various fields	Ontology / WordNet	no
(Erdmann et al., 2000)	yes	no	Text in natural language	Regular Expressions / SMES	no
(Névéol, Doğan and Lu, 2011)	-	yes	Biomedical	NLP / pre-annotations	no

Table 1: Comparison of the presented works (cont.).

(Dingli, Ciravegna and Wilks, 2003)	-	yes	Various fields	IE / Automatic learning	no
(Kiryakov et al., 2004) (Popov et al., 2003)	-	yes	Any type of text (web page / regular document (non web))	Ontology / KIM / IE	no
(Leopold et al., 2015)	-	yes	Various fields	Markov Logic	no

Based on the results obtained, we consider that the works (Davis et al., 2009), (Tao et al., 2013), (Kiyavitskaya et al., 2006), (Kiyavitskaya et al., 2005), (Vehvilainen, Hyvonen and Alm, 2008) and (Erdmann et al., 2000) partially automate the semantic annotation phase while it is completely automatic in (Qacim and Salih, 213), (Névéol, Doğan and Lu, 2011), (Dingli, Ciravegna and Wilks, 2003) and (Kiryakov et al., 2004). Thus, we note that some works use common approaches including, NLP and ontology. As for the application, most of the works focus on a specific field. Finally, we consider adaptation to inter-organizational workflows an important criterion because they allow to answer the problem posed in section 1, they also allow the collaboration between organizations with a better exchange of data while being flexible and effective (Semar-Bitah and Boukhalfa, 2016). However, the approaches used in the cited works in this paper do not cover the notion of inter-organizational workflows.

3 CONTRIBUTION

The aim of this study is to explore the different works related to the domain of automation of semantic annotation in order to define a solution whose objective is to detect and to correct the incompatibilities between the workflow partners.

To this end, so that organizations can collaborate with better compatibility, we propose an approach that aims to automate the semantic annotation phase for inter-organizational workflows (IOWFs) using the NLP approach that has proved to be successful. We recommend the adoption of methods to improve and optimize the IOWF semantic annotation, namely:

1) Hierarchy: It allows to improve the annotation by providing a formal framework that allows to argue on the consistency of the extracted information. In particular, semantic hierarchies have proved to be very useful in reducing the semantic gap. Three types of hierarchies for image annotation and classification have been recently explored:
   1) Hierarchies based on textual knowledge;
   2) Hierarchies based on visual (or perceptual) information, i.e. low-level characteristics of the image;
   3) Hierarchies that we call semantic based on both textual and visual information (Bannour and Céline, 2013).
2) Indexing: It makes it possible to document the knowledge represented by the semantic annotations, or even to keep them up to date when the reference texts evolve. In general, it is a question of using the semantic structure constructed by the annotations to identify elements in a document and to navigate semantic elements to fragments of text or vice versa. However, an indexing process consists of annotating text and gathering it by following a semantic organization (Lévy, Nazarenko and Guissé, 2010) (Guissé et al., 2010).
3) Learning: A learning process is characterized by an interaction between the learner and the environment by setting up a system capable of learning how to annotate a given corpus. The goal of this method is to acquire better or new knowledge and / or a mechanism or procedure (inference engine and knowledge) by deducing a set of rules. There are three techniques of learning in particular, learning patterns, digital learning and active learning (Bannour and Audibert, 2012).

4 CONCLUSION AND FUTURE WORKS

In this paper, we have presented the different approaches to semantic annotation found in the literature to remedy the problems of incompatibility of inter-organizational workflows. On the basis of the study carried out, we have recommended approaches that can contribute to the automation phase of semantic annotations while improving and optimizing the semantic annotations.

In the future, we want to concretise our vision by adopting the approaches cited in Section 3 to ensure better collaboration among workflow partners.

REFERENCES

Abbassene, A., Alimazighi, Z., and Aouachria, M. (2015). *Towards an open framework for inter-organizational workflow semantic annotation*. Algiers, Algeria: ISPS'15 - 12th International Symposium on Programming and Systems, USTHB.

Davis, B., Varma, P., Handschuh, S., Dragan, L., and Cunningham, H. (2009). *Controlled Natural Language for Semantic Annotation*: The Semantic Web: Research and Applications Volume 5554 of the series Lecture Notes in Computer Science, pp. 816-820.

Tao, C., Song, D., Sharma, D., and Chute, C. G. (2013). *Semantator: Semantic annotator for converting biomedical text to linked data*: Journal of biomedical informatics Volume 46, Issue 5, pp. 882-893.

Kiyavitskaya, N., Zeni, N., Mich, L., Cordy, G. R. and Mylopoulos, J. (2006). *Text Mining Through Semi Automatic Semantic Annotation*: Practical Aspects of Knowledge Management Volume 4333 of the series Lecture Notes in Computer Science, pp. 143-154.

Kiyavitskaya, N., Zeni, N., Cordy, J. R., Mich, L. and Mylopoulos, J. (2005). *Semi-Automatic Semantic Annotations for Web Documents*: Proc. SWAP 2005, 2nd Italian Semantic Web Workshop.

Vehvilainen, A., Hyvonen, E. and Alm, O. (2008). *A Semi-Automatic Semantic Annotation and Authoring Tool for a Library Help Desk Service*: Emerging Technologies for Semantic Work Environments: Techniques, Methods, and Applications.

Qacim, A. and Salih, M. (2013). *Toward From Manual to Automatic Semantic Annotation : Based on Ontology Elements and Relationships*: International Journal of Web & Semantic Technology; Vol. 4 Issue 2, p. 21.

Erdmann, M., Maedche, A., Schnurr, H. P. and Staab, S. (2000). *From manual to semi-automatic semantic annotation: about ontology-based text annotation tools*: Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content, pp. 79-85.

Névéol, A., Doğan, R. I. and Lu, Z. (2011). *Semi-automatic semantic annotation of PubMed queries: A study on quality, efficiency, satisfaction*: Journal of Biomedical Informatics 44, pp. 310-318.

Dingli, A., Ciravegna, F. and Wilks, Y. (2003). *Automatic Semantic Annotation using Unsupervised Information Extraction and Integration*: ceur-ws.

Kiryakov, A., Popov, B., Terziev, I., Manov, D. and Ognyanoff, D. (2004). *Semantic annotation, indexing, and retrieval*: Web Semantics: Science, Services and Agents on the World Wide Web 2, pp. 49-79.

Tulasi, R. L., Rao, M. S., Ankita, K. and Hgoudar, R. (2017). *Ontology-Based Automatic Annotation: An Approach for Efficient Retrieval of Semantic Results of Web Documents*: Proceedings of the First International Conference on Computational Intelligence and Informatics Volume 507 of the series Advances in Intelligent Systems and Computing, pp 331-339.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and Goranov, M. (2003). *KIM – Semantic Annotation Platform*: The Semantic Web - ISWC 2003 Volume 2870 of the series Lecture Notes in Computer Science, pp. 834-849.

Ranganathan, G. R., Biletskiy, Y. and Kaltchenko, A. (2008). *Semantic annotation of semi-structured documents*: Electrical and Computer Engineering: CCECE.

Leopold, H., Meilicke, C., Fellmann, M., Pittke, F., Stuckenschmidt, H. and Mendling, J. (2015, June). *Towards the automated annotation of process models* : In International Conference on Advanced Information Systems Engineering : Springer International Publishing, pp. 401-416.

Handschuh, S., Staab, S. and Ciravegna, F. (2002, October). *S-CREAM—semi-automatic creation of metadata:* In International Conference on Knowledge Engineering and Knowledge Management. Springer Berlin Heidelberg, pp. 358-372.

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A. and Ciravegna, F. (2002, October). *MnM: Ontology driven semi-automatic and automatic support for semantic markup:* In International Conference on Knowledge Engineering and Knowledge Management. Springer Berlin Heidelberg, pp. 379-391.

Davies, J., Fensel, D., & Van Harmelen, F. (Eds.). (2003). *Towards the semantic web: ontology-driven knowledge management*. John Wiley & Sons.

Kogut, P. A. and Holmes III, W. S. (2001, October). *AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages*: In Semannot@ K-CAP 2001.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A. and Zien, J. Y. (2003, May). *SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation*: In Proceedings of the 12th international conference on World Wide Web. ACM, pp. 178-186.

Guha, R. and McCool, R. (2001). *Tap: Towards a web of data.*

Pustejovsky, J., Boguraev, B., Verhagen, M., Buitelaar, P. and Johnston, M. (1997, March). *Semantic indexing and typed hyperlinking*: In Proceedings of AAAI Spring Symposium, NLP for WWW.

Wang, B., Huang, H., Wang, X., & Chen, W. (2009, December). *An Ontology-Based NLP Approach to Semantic Annotation of Annual Report*: In Computational Intelligence and Security, 2009. CIS'09. International Conference : IEEE, Vol. 1, pp. 180-183.

Semar-Bitah, K. and Boukhalfa, K. (2016). *Towards an Inter-organizational Collaboration Network Characterization. In Modelling and Implementation of Complex Systems*: Springer International Publishing, pp. 233-248.

Bannour, H. and Céline, H. (2013). *Construction de hiérarchies sémantiques pour l'annotation d'images :* Revue des Sciences et Technologies de l'Information-Série RIA: Revue d'Intelligence Artificielle, 27(1), 11-37.

Lévy, F., Nazarenko, A. and Guissé, A. (2010). *Annotation, indexation et parcours de documents numériques : Document numérique*, 13(3), 121-152.

Guissé, A., Lévy, F., Nazarenko, A., & Szulman, S. (2009). *Annotation sémantique pour l'indexation de règles métiers :* In L'Homme, M.-C. and Szulman, S., editors. Conférence Internationale sur la Terminologie et l'Intelligence Artificielle (TIA 2009), page (electronic medium). Université Paul Sabatier-Toulouse.

Bannour, S., & Audibert, L. (2012, June). *Vers une approche interactive pour l'annotation sémantique* : In IC2012.