

Extracting Knowledge Tokens from Text Streams

Eugene Alferov^{1,2} and Vadim Ermolayev¹

¹ Department of IT, Zaporozhye National University,
66 Zhukovskogo st., 69063, Zaporozhye, Ukraine

alferov.evgeniy@gmail.com, vadim@ermolayev.com

² Kherson State University, 27, 40 Rokiv Zhovnya ave., 73000, Ukraine

alferov_jk@ksu.ks.ua

Abstract. This problem analysis paper presents our position on how could the solution be sought to the problem of extracting semantically rich fragments from a stream of plain text posts. We first present our understanding of the problem context and explain the focus of our research. Further, in the problem setting section we elaborate the workflow for knowledge extraction from incoming information tokens. This workflow is then used as a key to structure our review of the literature on the relevant component techniques which may be exploited in a combination to achieve the desired outcome. We finally outline our plan for conducting the experiments with an aim to validate the workflow and find a proper combination of the component techniques for all steps which may solve our specific research problem.

Keywords. Workflow, knowledge extraction, text streams, processing, ontology learning, component techniques

Key terms. Data, Process, Knowledge, Approach, Methodology

1 Introduction

The dramatic growth of data volumes we face today is accelerated by the increase of social networking applications that allow non-specialist users create a huge amount of content easily and freely. Equipped with rapidly evolving mobile devices, a user is becoming a nomadic gateway boosting the generation of additional real-time sensor data. The emerging Internet of Things makes each and every thing a data or content, adding billions of additional artificial and autonomic sources of data to the overall landscape. Smart spaces, where people, devices, and their infrastructures are all loosely connected, also generate data of unprecedented volumes and with velocities rarely observed before. Noticeably, the major part of the new data comes in streams.

An expectation is that valuable information will be extracted out of all these data to help improve the quality of life and making our world a better place – for humans.

Humans are however left bewildered about how to use, analyze, understand all these data, giving a proper account to its dynamics. A topical recent estimate of the need for data-savvy managers in the United States is 1.5 million [1]. This manpower is needed to extract and use valuable information and knowledge for further decision making. The critical steps in this work are (i) extracting information and knowledge; and (ii) bringing the descriptions of the reflections of the world or domain into a refined state – accounting for the changes brought in by new data, at scale.

In this paper we focus on the step (i) extraction. In Section 2 we present the problem statement by giving basic definitions and providing our view on how could a processing workflow look like. The plethora of approaches, techniques, technologies, and software tools already exist for solving different parts of the overall problem. Hence we analyze the related work and structure this analysis using the workflow as the key in Section 3. Finally we conclude the paper and present our plans for the future proof of concept experimental work in Section 4.

2 Problem Statement

Ontology is a complex artifact that comprises structural components of several types. Further the structural denotation of an ontology used in Description Logics [2] is exploited: an ontology O comprises its schema S and the set of individuals I : $O = (S, I)$. Ontology schema is also referred to as a terminological component (TBox). It contains the statements describing the concepts of O , the properties of those concepts, and the axioms over the schema constituents.

If a finer grained look at an ontology schema is taken, one may consider S comprising the following interrelated constituents: $S = \{S^C, S^O, S^D, S^A\}$, where S^C is the set of statements describing concepts, S^O is the set of statements describing object properties, S^D is the set of statements describing datatype properties, and S^A is the set of axioms specifying constraints over S^C , S^O , and S^D (c.f. [3]). One may notice that these constituents correspond to the types of the schema specification statements of an ontology representation language L which is used for specifying O .

The set of individuals, also referred to as assertional component (ABox), is the set of the ground statements about the individuals and their attribution to the constituents of the schema.

Ontology Learning is the process of extracting the abovementioned constituents of O from a text stream source. More specifically, the problem which is approached in this research work is twofold:

For every individual plain text document (further referred to as information token) arriving in the stream window DO:

- (i) Extract ontological fragment (further referred to as knowledge token) specifying the semantics of the information token.
- (ii) Refine the ontology O incorporating the changes brought in by the knowledge token.

The focus of this paper is the first part of the problem – the extraction of knowledge tokens from information tokens of plain text in a particular professional domain coming in a stream. The texts of ICTERI paper abstracts have been chosen as the domain and source text corpus for our initial experiments – see also Section 4.

As an ontology is a complex artifact, the extraction of knowledge tokens from texts is also a complex process. It comprises several steps and, possibly, iterations for extracting different structural constituents of S and I . These steps produce several types of outputs in a particular sequence, sometimes referred to as the ontology learning layer cake (c.f. [4]). Those outputs are terms – concepts and their instances – datatype properties – taxonomic relationships and object properties – axioms. Based on [5] we present in Fig. 1 a workflow putting together extraction steps, inputs, outputs, and required component technology types.

The overall workflow contains two consecutive phases – Text Pre-processing and Ontology Extraction. Text Pre-processing phase gets the information token as a plain text input and produces its structured representation as a set of terms by applying several statistical and linguistic techniques. All the tasks of the Ontology Extraction Phase use the output of Phase 1 as their input and incrementally build up the knowledge token by adding different ABox and TBox constituents. For that statistical, linguistic, semantic, and logical techniques are employed in combinations. Fig. 1 lists all relevant component techniques per task. All of those are never used in implementations. Therefore our initial research objective is to find out which combination of component techniques works best of all for our specific data – i.e. copes well with (a) the texts of small size but belonging to a particular domain; and (b) limited processing time constrained by a stream window lifetime parameter. Further, after this constellation of component techniques is chosen, the objective would be to refine those which do not provide results of a satisfactory quality in our problem settings.

3 Related Research and Available Component Techniques

In this section we will describe the component techniques, outlined in Fig. 1, which we found relevant to our work. Those component techniques could overall be categorized as linguistic, statistic, semantic and logical (c.f. [5]). As pictured in Fig. 1 they could be applied at different steps and for different purposes. Though not explicitly shown in Fig. 1, the steps may undergo iterations for refining their results. Therefore, the workflow proposed in this paper could be considered as hybrid and iterative.

De-noising (statistical, linguistic). This is a method that extracts the de-noised text, comprising the content-rich sentences, from full texts [6]. Processing of noisy text becomes important because the quality of texts in the form of blogs, emails and chat logs can be extremely poor. The sentences in dirty texts are typically full of spelling errors, ad-hoc abbreviations and improper casing [7].

Tokenization. Tokenization is splitting the text into a set of tokens, usually words. This process is unsupervised and can be performed automatically by program-parser.

Part of speech detection/tagging (linguistic). Part of speech tagging (POST) is the process of assigning one of the parts of speech to the given word. POST provides the

syntactic structures and dependency information required for further linguistic analysis in order to uncover terms and relations. POST is a semi-supervised or even unsupervised process.

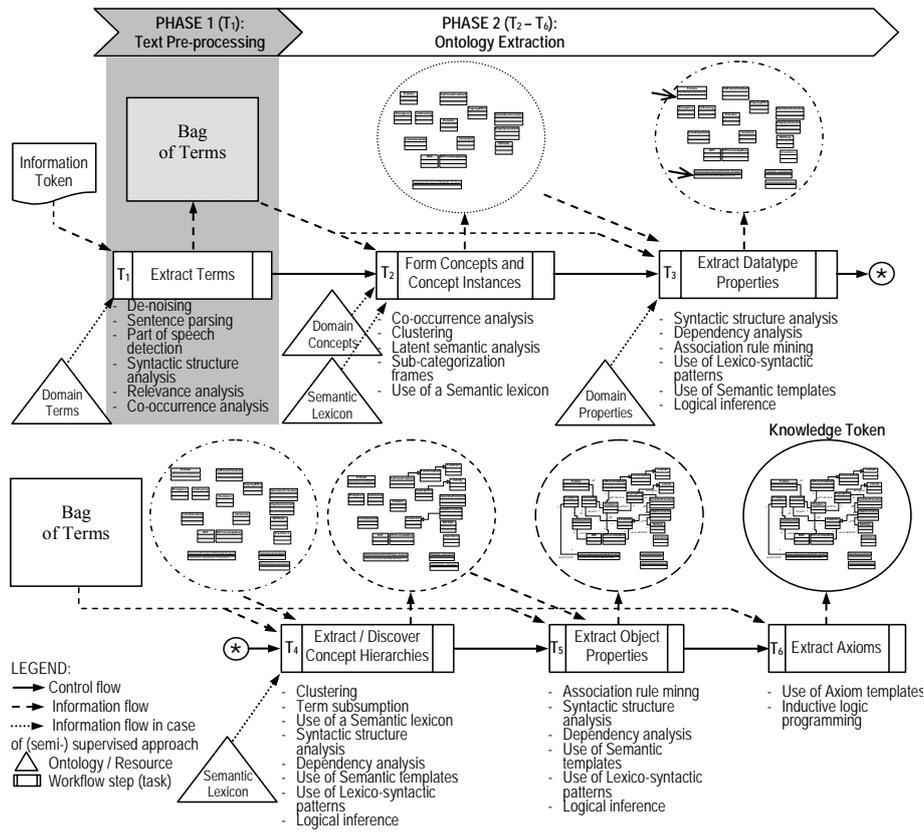


Fig. 1. A workflow for knowledge token extraction

Lemmatization (linguistic). Lemmatization is the reduction of morphological variants of the tokens to their base form that can be performed in unsupervised way. For achieving this word form must be known, i.e. the part of speech of every word has to be assigned in the text document. This process usually takes a time and may contain errors.

Chunking (linguistic). Chunking is unsupervised splitting a text in syntactically correlated parts.

Sentence parsing. Sentence parsing is identifying the syntactic structure of a sentence, for example in a form of a parse tree.

Syntactic structure analysis (linguistic). In syntactic structure analysis, words and modifiers in syntactic structures (e.g., noun phrases, verb phrases, and prepositional phrases) are analyzed to discover potential terms and relations. It can be done in unsupervised way.

Relevance Analysis (statistical). The extent of occurrence of terms in individual documents and in text corpora is employed for relevance analysis. This is semi-supervised or even unsupervised technique.

Co-occurrence analysis (statistical). Co-occurrence analysis identifies lexical units that tend to occur together for purposes ranging from extracting related terms to discovering implicit relations between concepts [5]. This technique is unsupervised.

Clustering (statistical). Grouping together variants of terms to form concepts and separating unrelated ones is known as terms clustering. It usually unsupervised technique. In this approach some measure of similarity is employed to assign terms into groups for discovering concepts or constructing hierarchy [8]. Some of the major issues in clustering are working with high-dimensional data and feature extraction and preparation for similarity measurement. This gave rise to a class of featureless similarity measures based solely on the co-occurrence of words in large text corpora. It is known that clustering results are of acceptable quality only if a statistically representative (i.e. large) text corpora is processed. This fact limits the applicability of this technique in our settings (texts of small size). However, used in the combination with other techniques, clustering may yield some valuable addition to the result – and thus needs to be tried.

Latent semantic analysis (statistical). Latent semantic analysis (LSA) is a theoretical approach and mathematical method for determining the meaning similarity of words and passages by analysis of large text corpora. The main idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other [9]. LSA can be useful in our investigation because it is a fully automatic mathematical and statistical technique for extracting and inferring meaningful relations from the contextual usage of words in text.

Sub-categorization (linguistic, semantic). Sub-categorization, or extracting sub-categorization frames, is an approach to extract one type of lexical information with particular importance for Natural Language Processing (NLP). Access to an accurate and comprehensive sub-categorization lexicon is vital for the development of successful parsing technology important for many NLP tasks (e.g. automatic verb classification) and useful for any application which can benefit from information about predicate-argument structure (e.g. Information Extraction) [10].

Using semantic lexicon (linguistic, semantic). A semantic lexicon is a dictionary or thesaurus of words/terms labeled with semantic classes (e.g., “ongoing effort” is an Activity) so associations can be drawn between words that have not previously been encountered [11]. Semantic lexicons are a popular resource in ontology learning and play an important role in many NLP tasks.

Dependency analysis (linguistic). Syntactic structure consists of lexical items, linked by dependencies. They are binary asymmetric relations that are held between a head and its dependents. Dependency analysis examines dependency information to uncover relations at the sentence level. In this analysis, grammatical relations, such as subject, object, adjunct, and complement, are used for determining more complex relations. Dependency analysis is usually unsupervised approach.

Association rule mining (statistical). Association rule mining aims to extract correlations, frequent patterns, associations or casual structures among sets of items in data repositories [12]. It is an unsupervised component technique which works well for considerably big data corpora. Association rules highlight correlations between features in the texts, e.g. keywords. Association rules can be easily interpreted and are understandable for an analyst or even for a normal user.

Use of lexico-syntactic patterns (linguistic). Lexico-syntactic patterns (LSPs) are generalized linguistic structures or schemas that indicate semantic relationships among terms and can be applied to the identification of formalized concepts and conceptual relations in natural language text [13]. Lexico-syntactic patterns are suitable for automatic ontology building, since they model semantic relations. These display exactly the kind of relation between their parts that makes them easily translatable into an ontology representation.

Use of semantic templates (semantic, linguistic). Semantic templates are similar to lexico-syntactic patterns in terms of their purpose. However, semantic templates offer more detailed rules and conditions for extracting not only taxonomic relations but also complex non-taxonomic relations [5].

Logical inference (logical, semantic). In logical inference implicit relations are derived from existing ones using rules such as transitivity and inheritance [5]. However, the introduction of invalid or conflicting relations may also happen in case of an incomplete or underspecified inference rule set – for example because of improper account for the validity of transitivity or mutual disjointness axioms.

Term subsumption (statistical, semantic). In the subsumption method, a given term subsumes another term if the documents in which the latter term occurs are a subset of the documents in which the given term occurs [14]. A term subsumption measure is used to quantify the extent of a term x being more general than another term y . This technique is semi-supervised and unsupervised too. The term subsumption technique is easy to implement and it makes labeling concepts an easy task. However, with this method, it is difficult to classify terms that do not co-occur frequently and it requires a large data set to work reliably.

Use of axiom templates (semantic, linguistic). Axioms are useful for describing the relationships between the concepts of an ontology. They can be written in different ways depending on the relation that exist among the concepts.

Inductive logic programming (logical, semantic). Inductive logic programming (ILP) is a research area at the intersection of inductive machine learning and logic programming. ILP generalizes the inductive and the deductive approaches by aiming to develop theories, techniques and applications of inductive learning from observations and background knowledge represented in first order logical framework.

The overview of the applicability of the presented component techniques and their interrelationship with respect to the tasks in our workflow are presented in Table 1.

4 Summary and Future Work

Our literature search has revealed that extracting knowledge, or more specifically learning ontologies, from plain text corpora is a well developed research field that continues to produce new results. However, and to the best of our knowledge, extracting ontologies from text streams, with a constraint on the life time of an input information token, is a recently emerged research problem. The reasons for adding this specific problem to the research agenda are the phenomenon of Big Data, in particular its velocity dimension, as well as the need for better, more reliable, semantically rich solutions for automating Big Data analytics. One more complication introduced by our problem setting is the small size of an individual information token which hinders yielding good quality results using the majority of traditional statistical and linguistic techniques for ontology extraction from text corpora.

We argued in this paper that applying a combination of the relevant existing component techniques in a structured and iterative way may overall produce such a result – as an incremental collection of ontology elements in a knowledge token provided by individual techniques at different stages in our proposed workflow.

Table 1. Relevance of component techniques to the tasks within the workflow for extracting knowledge tokens from information tokens

Component technology	Task (Fig. 1.)					
	T ₁	T ₂	T ₃	T ₄	T ₅	T ₆
De-noising	st, li					
Part of speech detection/tagging	li					
Lemmatization	li					
Chunking	li					
Syntactic structure analysis	li		li	li	li	
Relevance Analysis	st					
Co-occurrence analysis	st	st				
Clustering		st		st		
Latent semantic analysis		st				
Sub-categorization		se, li				
Using semantic lexicon		se, li		se, li	se, li	
Dependency analysis			li	li	li	
Association rule mining			st		st	
Use of lexico-syntactic patterns			li	li	li	
Use of semantic templates			se, li	se, li		
Logical inference			lo, se	lo, se	lo, se	
Term subsumption				st, se		
Use of axiom templates						se, li
Inductive logic programming						lo, se

Legend : li – linguistic; lo – logical; se – semantic; st – statistical;

As this research is in an early phase, we do not yet have the proof for this hypothesis. However there is the plan in place for conducting the initial series of the “proof-of-concept” experiments in which the component technologies will be exploited in a semi-supervised or supervised fashion. For that we plan to use a small but well se-

manically annotated corpus of the abstracts (information tokens) and full texts of ICTERI papers collected in the ICTERiWiki portal¹. This document corpus is incrementally extended by adding the papers and their semantic annotations for each new ICTERI conference instance. The annotations are done using the ICTERI Scope Ontology by Tatarintseva et.al. [15]. These annotations will be used as a “Golden Standard” for evaluating the results of automated knowledge token extraction using the workflow proposed in this paper.

After the concept is proven and the constellation of the component techniques is circumscribed, we plan to test the approach on one of the professional news portals. Further, it is planned to extend the proposed knowledge extraction procedure to sensor stream data processing.

References

1. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A.: Big data: the Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute (2011), http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation
2. Nardi, D., Brachman, R.J.: An Introduction to Description Logics. In: Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., Patel-Schneider, P. F. (eds.) *The Description Logic Handbook*, Cambridge University Press New York, NY, USA (2007)
3. Davidovsky, M., Ermolayev, V., Tolok V.: Instance Migration between Ontologies Having Structural Differences. In: *Int. J. on Artificial Intelligence Tools*, vol. 20(6), pp. 1127–1156 (2011)
4. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: an Overview. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.) *Ontology Learning from Text: Methods, Evaluation and Applications*, IOS Press, Amsterdam (2005)
5. Wong, W., Liu, W., Bennamoun, M.: Ontology Learning from Text: a Look Back and into the Future. *ACM Comput. Surv.*, 44(4), Article 20, 36 pages. <http://doi.acm.org/10.1145/2333112.2333115> (2012)
6. Shams, R., Mercer, R. E.: Investigating Keyphrase Indexing with Text Denoising. In: *Proceedings of the 12th ACM/IEEE-CS Joint Conf. on Digital Libraries*, pp. 263–266, ACM (2012)
7. Wong, W., Liu, W., Bennamoun, M.: Enhanced Integrated Scoring for Cleaning Dirty Texts. arXiv preprint arXiv:0810.0332. (2008)
8. Cimiano, P., Hotho, A., Staab, S.: Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of Artificial Intelligence Research Archive*, 24(1), 305–339 (2005)
9. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to Latent Semantic Analysis. *Journal: Discourse Processes*, 25(2-3), 259–284 (1998)
10. Preiss, J., Briscoe, T., Korhonen, A.: A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In: *Annual Meeting. Association for Computational Linguistics*, 45(1), 912 (2007)
11. Thelen, M., Riloff, E.: A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In: *Proc. ACL-02 Conf. on Empirical Methods in Natural*

¹ <http://isrg.kit.znu.edu.ua/icteriwiki/>

- Language Processing, Association for Computational Linguistics, vol. 10, pp. 214–221 (2002)
12. Kotsiantis, S., Kanellopoulos, D.: Association Rules Mining: a Recent Overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1), 71–82 (2006)
 13. Summary on Requirements on Lexico-Syntactic Patterns (Synthesis by PC), http://www.w3.org/community/ontolex/wiki/Specification_of_Requirements/Lexico-Syntactic_Patterns
 14. De Knijff, J., Frasincar, F., Hogenboom, F.: Domain Taxonomy Learning from Text: the Subsumption Method versus Hierarchical Clustering. *Data & Knowledge Engineering*, (2012)
 15. Tatarintseva, O., Borue, Yu., Ermolayev, V.: Validating OntoElect Methodology in Refining ICTERI Scope Ontology. In: H.C. Mayr et al. (Eds.): *UNISCON 2012, LNBIP 137*, pp. 128–139 (2013)