Toward a Theoretical Framework of Terminological Saturation for Ontology Learning from Texts

Victoria Kosa and Vadim Ermolayev

Department of Computer Science, Zaporizhzhia National University,
Zaporizhzhia, Ukraine
victoriya1402.kosa@gmail.com, vadim@ermolayev.com

Abstract. In this position paper, we propose a detailed technical outline of what needs to be done, for example in a Master project, to bridge the research gap for the problem of the existence of terminological saturation. The problem is studied regarding a sequence of incrementally growing sub-collections of documents describing an arbitrary subject domain, using the OntoElect approach. After reviewing the related work, we present the formal basics of the approach and experimental evidence of the existence of terminological saturation. Consequently, we formulate the research hypotheses, and outline the methodology and plan for further research elaborating on this position.

Keywords: terminological saturation, theoretical framework, distance metric, envelope function, saturation conditions, saturation existence theorem

1 Introduction

Extracting a set of terms from a document collection, describing a subject domain, is an important initial step in figuring out a complete set of requirements for building an ontology for the domain [1, 2]. The result of this step will only be of value if a source collection of documents is sufficiently complete. Otherwise, important terms might have been missed. A straightforward way to assemble a complete collection is to retrieve all the available documents. Unfortunately, this is not realistic due to the varying availabilities and huge quantities of the sources in realistic domains. A way to reduce the size of the collection to be processed, while keeping the completeness of the term set dissolved in it, is to extract a terminologically saturated subset of documents – a sub-collection termed as a terminological core.

In frame of OntoElect methodology [2], we have proposed the domain-independent technique for that, based on detecting terminological saturation in the sequence of incrementally growing sub-collections. In our approach, the relevant [3] documents are iteratively added to the sub-collection, terms are extracted from the previous and current snapshots, and terminological difference [4] between the snapshots is measured. After terminological difference had gone below the individual term significance threshold, the current sub-collection could have been considered terminologically saturated and could have been regarded as a terminological core. Our prior work experimentally

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 40–51

proved that, following this approach, it is possible to decrease substantially the quantity of documents for term extraction and make the bags of significant terms more compact [5], while preserving the representativeness of a terminological core for the domain. Several important aspects [6, 7, 8] influencing the emergence of saturation have been observed as well.

The following research question has been left, however, without a proper attention in our prior research: Is there a way to prove formally the existence of terminological saturation for an arbitrary collection of textual documents? This question is important, as terminology extraction from texts is computationally hard, even if optimized [9]. Hence, it might be of value to know if terminological saturation is achievable, having the set of documents at hand, before starting iterative computations using incrementally growing textual datasets.

In this position paper, we aim at blueprinting the formal framework to solve the outlined problem. For that, the hypothesis about the conditions for the existence of terminological saturation need to be formulated. However, we leave this proof for a separate research project leading to a degree in Computer Science. Therefore, this paper is the proposal of a potential master project.

The remainder of the paper is structured as follows. In Section 2, we review the existing related work on terminological saturation. In Sect. 3 we present our background experimental evidence of the phenomenon of terminological saturation and deliberate on relevant research questions. In Sect. 4, we present the structure of the formal theoretical framework and sketch out some of its basic components, including the hypothesis about the formal conditions for the existence of terminological saturation. In Sect. 5 we present a vision of the plan of the future research work towards developing this theoretical framework. Finally, we make conclusive remarks in Sect. 6.

2 Related Work

An ontology as an artifact [10], by definition, is “a formal, explicit specification of a shared conceptualization” [11]. In Ontology Engineering, the mainstream interpretation of this term for a domain ontology is that it is a formal descriptive theory of the subject domain. A particular property of a domain ontology, which we focus on in this work, is that it has to be a shared specification. A commonly accepted way of the assessment of being shared is the degree to which an ontology supports the mental pictures, interpretations of, or views on the subject domain by the domain professionals – the knowledge stakeholders. The more views, further termed as requirements, are supported by the ontology, the higher is the acceptance of this ontology by the knowledge stakeholders. Therefore, it is better shared by them.

To design a domain ontology to be well supporting the requirements of the relevant knowledge stakeholder community, it is imperative to be informed sufficiently fully about their views. This poses a challenge, as it is hard to elicit directly the interpretations of the domain from the knowledge stakeholders in an explicit form [4]. Different ontology engineering methodologies attack this challenge of requirements elicitation in slightly different manners, based on organizing systematic interviews or brainstorming sessions with the experts selected from the knowledge stakeholders ([12, 13, 14, 15, 16] to mention the few most frequently cited). However, there is always a risk, along this way, that the selected group and their requirements under-represent the sentiment of the specialist community. Furthermore, there is no guidance in Ontology Engineering literature on how to measure objectively the representativeness of the expert group and, therefore, the completeness of the requirements elicited from them. In fact, as the experts are expensive, the tradeoff between the completeness and the price is made in favor of lowering the price.

To overcome the abovementioned difficulty in ensuring representativeness, it has been proposed [17] to learn ontologies, or the requirements for ontology development (also termed as features), not from the group of experts, but from the artifacts developed by the knowledge stakeholders in the domain. The pragmatic reason was that extracting features from the artifacts is less laborious and could be automated, at least in part. Furthermore, collecting a representative sample of artifacts is more feasible than a representative group of experts.

One of the relevant types of these artifacts is professional textual documents. Ontology learning from texts is now a noticeable subfield in ontology learning with developed methodologies and processing pipelines [18, 1]. To learn the requirements for engineering a domain ontology, a representative set of textual documents needs to be collected. This document collection has to:

- Contain relevant texts of sufficiently good quality
- Be representative (sufficiently complete) in order to reflect community consensus

While there is a bunch of approaches to select relevant documents in the literature, the problem of checking if a document collection is sufficiently complete has not been adequately resolved. One reason for under-estimating the importance of ensuring the representativeness of a text collection is that text resources are abundant. Hence, one may always expect to be able to have enough if the domain of her interest is well circumscribed. For example, it is feasible to collect high-quality research papers on a particular topic or within a field, as we did for Knowledge Management (KM) [6]. This KM collection contains circa 9 000 journal articles in full texts, which might be considered as a sufficient volume due to the recommendations of linguistic corpora experts, e.g. [19]. However, even if one collects what she thinks enough, there has to be a way to measure the representativeness of this text corpus regarding ontology learning. It might also be worth knowing if continuing collecting more documents will finally result in a representative corpus.

Furthermore, it might happen that only a small part of a big document collection is sufficiently complete in terms of domain knowledge coverage. Therefore, having a method to find this terminological core within the entire collection would help substantially decrease the effort needed for ontology learning from these texts.

The only relevant Computer Science publication we found in the context of ensuring the completeness of the set of processed documents is [20] by Ferrari et al. This work proposes two completeness metrics that take into account the relevant terms and relationships among terms extracted from software system requirements specifications written in natural language. This approach helps assessing if the set of specified requirements is complete with respect to the available document or a small set of documents. However, it does not allow finding out if the used set of documents represents the sentiment of the specialist community satisfactorily fully. Despite that, the approach of Ferrari et al. resembles our work as both are based on terminology extraction.

The use of saturation phenomenon has received little attention in the Computer Science literature, in particular in Text Mining and Ontology Learning. Saturation of clauses was used in Theorem Proving [21]. Term saturation was also used in document clustering and query answering for building term proximity graphs [22]. One more example, related to clustering, is the use of hierarchical cluster analysis for building topic taxonomy for a properly sampled subset of documents [23]. Saturation measure (together with ceiling) is used for patent text clustering and topic classification [24].

The only broadly exploited analogy to terminological saturation, which is directly relevant for our purposes, we found in qualitative research methodologies for Sociology and Medical Sciences, which is theoretical (or data) saturation [25]. Qualitative research is applied in different domains for processing the interviews with the subjects and with an aim to build a (descriptive) theory that supports a research hypothesis in the given (social) context. The problem faced by qualitative analysts was that the interviews were expensive. So, it was desired to have an indicator of a representative subjects sample size, such that covers well the potential replies by the other subjects who were not interviewed.

In Qualitative Research, the phenomenon of data saturation finds its origin in the Grounded Theory method by Glaser and Strauss [26] for conducting interviews and processing the data collected in these interviews. They explained their method as “the discovery of theory from data systematically obtained from social research” [26]. In their proposal, a theory becomes grounded exactly due to its systematic discovery – i.e. every statement in the theory has to be supported by the data.

Notably, a mainstream ontology engineering methodology could be very similarly termed as the discovery of a descriptive domain theory based on the data obtained from qualitative research – c.f. [11]. By analogy, an ontology could be regarded as grounded in evidence (data) if data saturation has been detected. For ontology learning from texts, a set of terms used in a domain could be regarded as such an evidence. Hence, if the set of terms becomes saturated, an ontology devisable from these terms (as the features pointing to the requirements) could be regarded as a grounded descriptive theory of the domain. Consequently, the subset of texts, from which the saturated set of terms has been extracted, could be regarded as the terminological core corpus for this domain.

Numerous attempts to operationalize the detection of data saturation has been mentioned in the Qualitative Analysis literature. However, it is still a “mysterious step” [25] in the Grounded Theory method. Sociologists identify the factors that pertain to or hinder data saturation and offer several methodological hints, informally. However, an objective and proven formal measure for detecting data saturation is still not available in the literature. To the best of our knowledge, the only reference, related to terminological saturation in textual data in the context of term extraction is our previous work [4].

3 Background and Experimental Evidence

We now briefly present our background knowledge in the context of detecting terminological saturation in document collections. This background has been developed in the OntoElect project[1]. We first outline the formal basics of the technique proposed for detecting terminological saturation in Sect. 3.1. In Sect 3.2, we then summarize our experimental evidence of the validity of this technique and emphasize the problem. This problem represents the research gap, which needs to be further studied and narrowed.

3.1 Formal Background in Detecting Terminological Saturation

In OntoElect [4, 2], we seek for a set of terms that statistically fully describe an arbitrary subject domain (Dom), for which a domain ontology needs to be developed or refined. Our supposition is that, if we have a sufficiently bounded Dom, the set of terms used to describe it is finite and not very large. These terms could be extracted from the documents (*Doc*), belonging to a documents collection ($DC = \{Doc\}$) that describes Dom. Hypothetically, one may collect all the existing documents that describe Dom – a complete document collection.

Definition 1: *A Complete Document Collection for* Dom. A DC containing all the documents describing Dom is a Complete DC (CDC).

For any realistic Dom, its CDC may be very big in volume. Therefore, extracting a representative set of terms from it would be a tedious and resource consuming task. This is why we look at the document sub-collections of a CDC for the Dom.

Definition 2: *A document sub-collection*. A document sub-collection (DSC) for the Dom is the subset of the CDC: $DSC \subset CDC$.

We are interested in finding a DSC of a minimal possible size that contains statistically the same set of terms as the CDC. The sets of terms are considered as statistically the same if the terminological difference, between the bags of terms extracted from these DSC and CDC, is negligible.

A Terminological Basis of a Domain. Perhaps, a good starting point is to clarify what it meant that a DC describes a Dom. We interpret such a description as containing a subset of valid and significant features (indicating requirements) that characterize the Dom. These features are labelled using the terms describing the domain.

Definition 3: *A Terminological Basis*. The finite set of terms t_i, identifying all the features that characterise Dom, form the terminological basis $TB = \{t_i\}, i = 1, \dots, dim$ of Dom.

The terms in a TB might not be equally important for describing the domain, as reflected in the documents of the CDC. Let a real positive value (*score*) be associated with every term used in the documents of the CDC. The more significant the term is for describing the domain, the higher is the *score*. Hence, a vector space model (VSM) [27] might be an appropriate formal representation of a document space for Dom. In this model, any document or DC, including CDC, is a point in a vector space with the basis

[1] https://www.researchgate.net/project/OntoElect-a-Methodology-for-Domain-Ontology-Refinement

$TB = \{t_i\}$ having dimension dim. It is not easy, however, to build a TB for any realistic domain. We have to have a technique to: extract t_i from the documents describing the Dom; make sure that an extracted $t_i \in TB$; and that all significant t_i have been extracted.

Extracted Terms. Let there be a mapping that transforms a DC into a bag of terms (B) extracted from the documents of this DC. Every element b in B is a pair $< t, score >$, where t is a candidate string and $score$ is the estimate of the likelihood that t is a relevant term for the Dom: the higher the $score$, the more likely $t \in TB$. The term candidates having high scores are denoted as **significant terms**.

Retained Significant Terms. Let a term significance threshold (eps) be rationally chosen (or estimated) for the $score$s of individual terms in a B. It indicates a boundary above which the terms are regarded as significant, hence belong to the TB. After having built a B, let us retain the bs with the *score* > *eps* in the corresponding Bag of Significant Terms (T). In many ATE methods, eps is either chosen empirically, or selected based on common sense considerations. In OntoElect, we offer a rationale for the estimation of an *eps*.

A Simple Majority Vote on Terms. A subset of term candidates is considered the core, containing all significant terms, if the terms in this core reflect the sentiment of a simple majority of the knowledge stakeholders in the domain. Consequently, the *score* of a term might be interpreted as the sum of the **votes** for this term by the domain knowledge stakeholders.

Definition 4: *An Individual Term Significance Threshold*. Let B be sorted in the descending order of the *score*s. Then *eps* threshold for B is computed as follows (1):

$$eps = score_i : \sum_{j=1}^{i} score_j > 1/2 \sum_{j=1}^{\|B\|} score_j, \qquad (1)$$

where i is the minimal number such that the condition after the semicolon holds.

Definition 4 specifies that, for computing the eps for a B, the minimal subset of bs at the top of the ordered list is found, whose voters are the simple majority.

Definition 5: *A Bag of Retained Significant Terms*. Let $T \subset B$ such that the following condition (2) holds:

$$\forall i, score_i > eps. \qquad (2)$$

Then T denotes the bag of retained significant terms.

Successive Approximation for building a $\boldsymbol{TB}$. One (superficial) way for building a TB might be to extract and retain all the significant terms from the CDC. The terms in the bag of significant terms T_{CDC} will constitute the TB for the Dom. This is however hardly tractable for any realistic Dom because of at least two reasons: (i) one has to ensure that the collection of documents s/he possesses is a CDC, which is hard; and (ii) processing a DC, that qualifies to be a CDC for a realistic Dom, is very computationally expensive because of its volume. One feasible way might be to use a statistically representative DSC instead of CDC. Then, the relevant and valid terms extracted from this DSC will form a basis that is very similar to TB, with statistically negligible difference. Hence, we need a way to figure out if a DSC is statistically representative. In this regard, a successive approximation technique is plausible.

Terminological Saturation. Let: $DSC_1, DSC_2, \dots, DSC_i, \dots$ be the sequence of document sub-collections such that $DSC_1 \subset DSC_2 \subset \dots \subset DSC_i \subset \dots \subset CDC$; $T_1, T_2, \dots, T_i, \dots$ be the bags of retained significant terms extracted from $DSC_1, DSC_2, \dots, DSC_i, \dots$. Let thd be a function comparing the bags of significant terms T_i, T_{i+1} retained from the successive DSC_i, DSC_{i+1} and returning the difference as a real positive value. If, at some i: (i) thd goes below the threshold of the statistical error ε; and (ii) there is a convincing evidence that it will never go above this threshold; then the difference (distance) between T_i and T_{CDC} is not higher than ε. The set of terms in such a T_i could be used as an ε -approximation of TB. Such a T_i, labelled further as T_{sat}, is a **saturated term set**; B_{sat} is the bag of terms from which T_{sat} is retained; and DSC_{sat} is a saturated DSC for Dom. The difference (thd) between T_{sat} and any successive T, including T_{CDC}, is within the statistical error: $thd(T_{sat}, T_{CDC}) < \varepsilon$.

Terminological Saturation Threshold. The premise for a rational choice of the threshold (ε) for detecting terminological saturation is that a set of terms becomes saturated if it already contains all the terms from TB. Hence, whatever the terms are added in the subsequent Ts, these are not significant. Let us set $\varepsilon = eps_{B_{sat}}$ – the eps (1) computed for B_{sat} . Then T_{sat} (2) will contain statistically the same set of terms as TB.

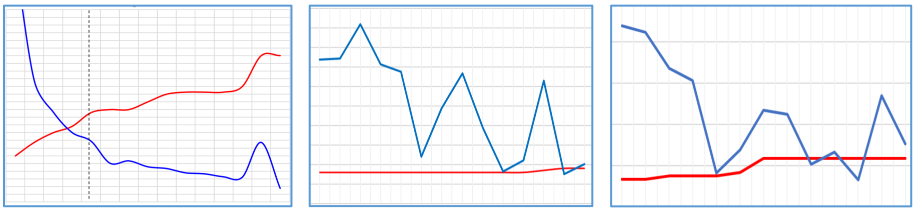
Terminological Difference Measure. Let us now define the function (thd) for measuring the distance between term sets. It takes in a pair of term sets and maps these arguments into a real positive value of the distance between them:

$$thd\colon \{< T_i, T_j >\} \longrightarrow \Re^+. \tag{3}$$

Due to the incremental nature of sub-collections DSC_i, not only the number of extracted terms will grow in B_{i+k} compared to B_i, but also the absolute values of the $score$s of the terms. Therefore, for making the $score$s comparable in the pairs, normalized $score$ values have to be used. Let $maxscore$ be the $score$ of the term, in a bag of terms B, having maximal value. In a B, sorted in the descending order of the $score$s, $maxscore$ is the $score$ of the first element. Then a *normalized score* (ns) could be computed as $ns = score/maxscore$. To measure thd we need to: (i) extract B_i from DSC_i and B_j from DSC_j; (ii) compute eps_i for B_i and eps_j for B_j using (1); (iii) retain significant terms in T_i and T_j using (2); (iv) compute ns for T_i and T_j; (v) compute $thd(T_i, T_j)$ based on (i)-(iv).

3.2 Experimental Evidence of Terminological Saturation

Experiments show that terminological saturation, measured using thd, exists in the collections, having appropriate volume, composed of carefully selected documents describing the same Dom (Fig. 1(a)). From the other hand, if the documents on arbitrary topics (different Doms) are randomly taken into DSC_i, then terminological saturation is not observed (Fig. 1(b)) and appears to be not reachable. Furthermore, in some collections thd measurements result in an unclear picture – as pictured in Fig. 1(c). In the latter case, thd values are volatile and oscillate quite sharply around the curve of eps, hence, cannot reliably indicate if terminological saturation is reachable. Fig. 1(c) pictures that there is often a chance that “might be” saturation observed in several measurement steps is further disproved by an additional measurement. In such cases, the collection might be not representative and, therefore, more documents have to be added to it. One more reason for high thd volatility might be that the collection is too noisy, as it appeared to be in the case of DAC[2] [7] (Fig. 1(c)).



(a) Saturation in DMKD-300[3] (b) Absence of saturation in RAW[4] (c) Saturation is not clear in DAC

Legend: 🟥 individual term significance threshold eps; 🟦 terminological difference $thd(T_i, T_{i+1})$

Fig. 1: Saturation measurements for different document collections, adapted from [28]

Hence, the open problem that needs to be further researched is finding the sufficient conditions for terminological saturation to exist after a necessary condition, its indication, have been observed in thd measurements.

Our position is that the problem needs to be solved formally and the theoretical framework for the solution has to be elaborated as a structured set of proven formal statements.

4 The Structure of the Formal Framework

Based on the OntoElect method and the experimental evidence of its validity, presented in Sect. 3, we now outline the open research questions and put forward the hypotheses that need to be proven. The hypotheses are formulated in a way to resolve the issues mentioned in the context of the volatility of thd measurements and, therefore, instability in terminological saturation. These statements form the logic of the sought theoretical framework.

4.1 The Formal Properties of thd

The thd function is the mapping of the vectors T_i, T_j, representing respective partial collections, into $\mathfrak{R}^+$. Several distance functions are known from the literature in this context – c.f. [8].

We put forward the following research hypotheses about this function.

Hypothesis 1: thd is Manhattan distance function.

[2] DAC is the collection containing 506 full text papers published in the proceedings of the Design Automation Conference between 2004 and 2010.

[3] DMKD-300 is the collection containing 300 full text articles published in the Journal of Data Mining and Knowledge Discovery between 1997 and 2010.

[4] RAW collection was synthetically formed of 80 randomly articles from English Wikipedia such that no two of them were about a similar topic and the size of an article was not too small.

Manhattan (or often also called taxicab) distance [29] is:

$$dman(v_i, v_j) = \sum_{k=1}^{n} \left| v_i^k - v_j^k \right|, \tag{4}$$

where $v_i = (v_i^1, v_i^2, \dots, v_i^n)$, $v_j = (v_j^1, v_j^2, \dots, v_j^n)$ are the vectors in an n-dimensional $\Re^+$ vector space with fixed Cartesian coordinate system. Hence, it has to be proven that (4) is the formula for computing thd.

Hypothesis 2: Let CDS be a vector space formed of VSM representations of all possible document sub-collections of a CDC in Dom. Then $thd(T_i, T_j)$ is a metric function and CDS is a metric space.

For proving this statement, the metric conditions for $thd(T_i, T_j)$ need to be checked: (i) non-negativity; (ii) triangle inequality; (iii) symmetry; and (iv) identity of indiscernibles. Further, it has to be proven that a CDS is a metric space with $thd(T_i, T_j)$ as its distance metric.

4.2 Envelope Functions for thd and eps

For terminological saturation measurement, we are interested in computing thd not for arbitrary T_i, T_j, but for successive pairs T_i, T_{i+1}, $i = 1, 2, \dots$. Hence, the function $thds(i) = thd(T_i, T_{i+1})$ is of our practical interest. In particular, we are interested when $thds(i)$ goes below $eps_{B_{sat}}$. To find this out, we have to analyse:

- The values of individual term significance thresholds eps_i, used for retaining terms in T_i, which could be regarded as a function $eps(i) = eps_i, i = 1, 2, \dots : B_i \rightarrow \Re^+$
- The values of $thds(i), i = 1, 2, \dots : \{T_i, T_{i+1}\} \rightarrow \Re^+$

As it is revealed in our experiments (Sect. 3.2), $eps(i)$ is not necessarily a monotonically non-decreasing function. However, in general it might be possible to build its approximation, $eps_{min}(i)$, as a lower envelope function, that is a monotonically non-decreasing function.

Analogously to $eps(i)$, $thds(i)$ is not necessarily a monotonically non-increasing function. It might also be possible to build its approximation, $thds_{max}(i)$ as an upper envelope function, that is a monotonically non-increasing function.

Hypothesis 3: Terminological saturation exists if there exist: (i) a monotonically non-decreasing function $eps_{min}(i)$; a monotonically non-increasing function $thds_{max}(i)$; the intersection of these functions at some i.

4.3 The Existence of Terminological Saturation

Hypothesis 3 may be formulated as an existence theorem for terminological saturation in a sequence of incrementally growing document sub-collections as follows.

Theorem 1 (**sufficient conditions of terminological saturation**). Let: (i) $DSC_1 \subset DSC_2 \dots \subset DSC_i \dots$ be the sequence of document sub-collections, each describing an arbitrary domain Dom; (ii) $B_1, B_2, \dots, B_i, \dots$ be the sequence of the bags of terms extracted from $DSC_1, DSC_2, \dots, DSC_i \dots$ and $eps(i)$ is the function of individual term significance thresholds for $B_i, i = 1, 2, \dots$; (iii) $T_1, T_2, \dots, T_i, T_{i+1}, \dots$ be the sequence of the bags of retained significant terms for which pairwise successive terminological difference is computed using the $thds(i)$ function. Then, the sequence of DSC_i is terminologically saturated, i is the saturation point, and $TB_i = TB_{sat} = TB$, if:

(i) There exist a non-decreasing lower envelope function $eps_{min}(i)$ for $eps(i)$
(ii) There exist a non-increasing upper envelope function $thds_{max}(i)$ for $thds(i)$
(iii) $eps_{min}(i) \geq thds_{max}(i)$

4.4 Research Methodology and Plan for the Future Work

It is envisioned that the research outlined above will be done following a hybrid method. The statements of Hypotheses 1-3 have to be proven formally. Further, these proofs have to be verified in the experiments using the instruments and datasets available in the OntoElect Project [7] in the frame of the optimized processing pipeline [9]. In particular, the datasets generated from the DMKD-300 document collection will be used as our prior experiments demonstrated quick and stable terminological saturation in this collection.

It is planned that the experimental part will be organized as follows:

(i) Envelope functions for $eps(i)$ and $thds(i)$ will be predicted based on the first 30 percent of measurements
(ii) Terminological saturation will be predicted based on these envelope functions and proven existence Theorem 1
(iii) Terminological saturation will be checked using the remaining 70 percent of measurements

Further, the same experiments will be done using the datasets of the RAW and DAC collections to verify if the prediction of terminological saturation works reliably in complex conditions and across domains.

5 Conclusive Remarks

The objective of this position paper is the proposal of a Master project aimed at developing a rigorous theoretical framework proving the existence of terminological saturation in the sequence of incrementally enlarged sub-collections of documents describing an arbitrary subject domain. The proposal is based on the background knowledge of the OntoElect project.

In the literature study focused on this topic, we found out that little attention has been paid, to date, to the problem of terminology saturation in textual corpora. In particular, the only measure of such a saturation is our own prior work [4]. Furthermore, a formal justification for the existence of terminological saturation in this context has not been provided and the conditions for saturation to exist have not been studied. Therefore, the development of a theoretical framework, proposed and outlined in this paper, is timely and important for research and practice.

The proposal of the framework is structured along the facets of: terminological distance measure and its metric properties; the envelope functions to cope with non-monotonicity of saturation measurement; and the existence theorem that states the required conditions.

Our plans for the future work are: (i) elaborate and validate experimentally the formal proofs of Hypotheses 1-3; (ii) modify our processing pipeline using the knowledge from the theoretical framework; and (iii) evaluate the instrumental software in the experiments on industrial-scale textual collections, like Springer KM [6].

References

1. 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 p. (2012). doi: 10.1145/2333112.2333115
2. Ermolayev, V.: OntoElecting requirements for domain ontologies. The case of time domain. EMISA Int J of Conceptual Modeling 13(Sp. Issue), 86–109 (2018). doi: 10.18417/emisa.si.hcm.9
3. Dobrovolskyi, H., Keberle, N.: Collecting seminal scientific abstracts with topic modelling, snowball sampling and citation analysis. In: Ermolayev, V. et al. (eds.): ICTERI 2018. Volume I: Main Conference, CEUR-WS, vol. 2105, pp. 179192 (2018)
4. Tatarintseva, O., Ermolayev, V., Keller, B., Matzke, W.-E.: Quantifying ontology fitness in OntoElect using saturation- and vote-based metrics. In: Ermolayev, V., et al. (eds.) Revised Selected Papers of ICTERI 2013, CCIS, vol. 412, pp. 136--162 (2013). doi: 10.1007/978-3-319-03998-5_8
5. Ermolayev, V., Batsakis, S., Keberle, N., Tatarintseva, O., Antoniou, G.: Ontologies of time: review and trends. Int J of Computer Science and Applications 11(3), 57–115 (2014)
6. Kosa, V., Chugunenko, A., Yuschenko, E., Badenes, C., Ermolayev, V., Birukou, A.: Semantic saturation in retrospective text document collections. In: Mallet, F., Zholtkevych, G. (eds.) ICTERI 2017 PhD Symposium, CEUR-WS, vol. 1851, pp. 1–8, (2017)
7. Kosa, V., Chaves-Fraga, D., Naumenko, D., Yuschenko, E., Moiseenko, S., Dobrovolskyi, H., Vasileyko, A., Badenes-Olmedo, C., Ermolayev, V., Corcho, O., Birukou, A.: The Influence of the Order of Adding Documents to Datasets on Terminological Saturation. Tech. Rep. TS-RTDC-TR-2018-2-v2, Zaporizhzhia National University, Ukraine, 72 p. (2018)
8. Kosa, V., Chaves-Fraga, D., Keberle, N., Birukou, A.: Similar terms grouping yields faster terminological saturation. In: Ermolayev, V. et al. (eds.) ICTERI 2018. Revised Selected Papers. CCIS, vol. 1007, pp. 43–70 (2019). doi: 10.1007/978-3-030-13929-2_3
9. Kosa, V., Chaves-Fraga, D., Dobrovolskiy, H., Fedorenko, E., Ermolayev. V.: Optimizing automated term extraction for terminological saturation measurement. In: Ermolayev. V. et al. (eds.) ICTERI 2019, Volume I: Main Conference, CEUR-WS, vol. 2387, pp. 1–16 (2019)
10. Guarino, N., Oberle, D., Staab, S.: What is an ontology? In: Staab S., Studer R. (eds.) Handbook on Ontologies, pp. 1–17, International Handbooks on Information Systems, Springer, Berlin, Heidelberg (2009)
11. Studer, R., Benjamins, R., Fensel, D.: Knowledge engineering: principles and methods. Data & Knowledge Engineering 25(1–2), 161–198 (1998)
12. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering. Springer, London (2004)

13. Pinto, H. S., Tempich, C., Staab, S., Sure, Y.: DILIGENT: towards a fine-grained methodology for distributed, loosely controlled and evolving engineering of ontologies. In: de Mántaras R.L., Saitta L. (eds.) 16th European Conf. on Artificial Intelligence, ECAI, pp. 393–397, IOS Press, (2004)
14. Schreiber, G. et. al.: Knowledge Engineering and Management: The CommonKADS Methodology. MIT Press, Cambridge, Massachusetts (1999)
15. Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E., Gangemi, A. (Eds.): Ontology Engineering in a Networked World. Springer (2012)
16. Sure, Y., Staab, S., Studer, R.: On-To-Knowledge methodology. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 117–132, Series on Handbooks in Information Systems, Springer, Berlin, Heidelberg (2003)
17. Maedche, A., Staab, S.: Ontology learning for the Semantic Web. IEEE Intell. Syst 16(2), 72 –79 (2001)
18. Buitelaar, P., Cimiano, P., and Magnini, B. (eds.): Ontology Learning from Text: Methods, Evaluation and Applications. Frontiers in Artificial Intelligence and Applications, vol. 123, IOS Press (2005)
19. Corpas Pastor, G., Seghiri Domínguez, M.: Size matters: a quantitative approach to corpus representativeness. In: Rabadán, R., Fernández López, M., Guzmán González, T. (eds.) Lengua, traducción, recepción en honor de Julio César Santoyo, pp. 111–145, Universidad de León Área de Publicaciones, León (2010)
20. Ferrari, A., dell'Orletta, F., Spagnolo G.O., Gnesi, S.: Measuring and improving the completeness of natural language requirements. In: Salinesi C., van de Weerd I. (eds.) REFSQ 2014, LNCS, vol. 8396, Springer, Cham (2014)
21. Riazanov, A., Voronkov, A.: Adaptive saturation-based reasoning. In: Dines Bjorner, D., Broy, M., Zamulin, A. (eds.) Andrei Ershov 4th Int Conf on Perspectives of System Informatics (PSI'01), pp. 55–61 (2001)
22. Chernyak, L., Berenstein, A.: Method and apparatus for informational processing based on creation of term-proximity graphs and their embeddings into informational units. US Patent Application Publication, No US 2006/0031219 A1, Feb. 9 (2006)
23. Doerre, J., Gerstl, P., Goeser, S., Mueller, A., Seiffert, R.: Taxonomy generation for document collections. US Patent, No US 6 446 061 B1, Sep. 3 (2002)
24. Han, H., Xu, S., Zhu, L.: Mining technical topic networks from chinese patents. In: Jung, H., Mandl, T., Womser-Hacker, C., Xu, S. (eds.) 1st Int W-shop on Patent Mining and Its Applications (IPAMIN 2014), CEUR-WS, vol. 1292 (2014)
25. Aldiabat, K. M.: Data saturation: the mysterious step in Grounded Theory methodology. The Qualitative Report 23(1), 245–261 (2018)
26. Glaser, B. G., Strauss, A.: The Discovery of Grounded Theory: Strategies for Oualitative Research. Aldine, Chicago, IL (1967)
27. Salton, G., Wong, A., Yang, C. S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
28. Kosa, V., Chaves-Fraga, D., Naumenko, D., Yuschenko, E., Badenes-Olmedo, C., Ermolayev, V., Birukou, A.: Cross-evaluation of automated term extraction tools by measuring terminological saturation. In: Bassiliades, N., et al. (eds.) ICTERI 2017. Revised Selected Papers. CCIS, vol. 826, pp. 135–163 (2018)
29. Gomaa, W. H., Fahmy. A. A.: A survey of text similarity approaches. Int J Comp Appl 68(13), 13–18 (2013)