Semantic Saturation in Retrospective Text Document Collections

Victoria Kosa[1], Alyona Chugunenko[1], Eugene Yuschenko[2], Carlos Badenes[3], Vadim Ermolayev[1], and Aliaksandr Birukou[4]

[1] Department of Computer Science, Zaporizhzhya National University, Zhukovskogo st. 66, 69600, Zaporizhzhya, Ukraine
victoriya1402.kosa@gmail.com, aluonac@i.ua, vadim@ermolayev.com

[2] BWT Group, Mayakovskogo st. 11, 69035, Zaporizhzhya, Ukraine
admin@groupbwt.com

[3] Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
cbadenes@fi.upm.es

[4] Springer-Verlag GmbH, Tiergartenstrasse 17, 69121, Heidelberg, Germany
Aliaksandr.Birukou@springer.com

Abstract. This paper presents the motivation for, planning of, and very first results of the PhD project by the first author. The objective of the project is to experimentally assess the representativeness (completeness), for knowledge extraction, of a retrospective textual document collection. The collection is chosen to describe a single well circumscribed subject domain. The approach to assess completeness is based on measuring the saturation of the semantic (terminological) footprint of the collection. The goal of this experimental study is to check if the saturation-based approach is valid. The project is performed at the Dept. of Computer Science of Zaporizhzhya National University in cooperation with BWT Group, Universidad Politecnica de Madrid, and Springer-Verlag GmbH.

Keywords. Knowledge extraction from text, text mining, retrospective document collection, OntoElect, semantic saturation, completeness

Key Terms. KnowledgeEngineeringMethodology, KnowledgeEngineeringProcess, SubjectExpert, KnowledgeEvolution, KnowledgeRepresentation

1 Introduction

This short paper presents a PhD project aimed at developing the methodological and instrumental components for measuring the representativeness of high-quality collections of text documents. It is assumed that the documents in a collection cover a single and well circumscribed Domain of Discourse and have a timestamp associated with them. A typical example of such a collection is the set of the full text papers of a

professional journal or a conference proceedings series published from the first issue to date. The main hypothesis, put forward in this work, is that a collection can be considered as representative to describe the domain, in terms of its semantic (terminological) footprint, if any additions of extra relevant documents to the collection do not noticeably change this footprint. Such a collection could be further considered as complete and could be used for extracting domain semantic descriptions from it. In fact, the approach to assess the representativeness outlined above does so by evaluating the terminological saturation of a document collection.

It is well known that extracting knowledge from texts for developing domain ontologies is a complicated and laborious process which requires a substantial part of highly qualified human effort. So, knowing the smallest possible representative document collection for a domain is very important to efficiently develop ontologies with satisfactory domain coverage. Therefore, laying out a method to determine a saturated subset of documents within the collection is topical. It is also important to make this method as efficient and automated as possible to lower the overhead on the core knowledge engineering workflow.

Yet one more dimension of complexity in the context of knowledge extraction from texts is terminological temporal drift. Indeed, the semantic footprint of a retrospective collection could change in time. So, it is not clear how could the saturated subset of the collection be formed to account for this drift.

The objective of the presented project is to develop and evaluate in industrial settings an efficient and effective experimental method, supported by an instrumental toolset, to determine saturated subsets of high-quality domain-bounded retrospective textual document collections. As a theoretical background, the project uses the OntoElect approach [1]. Term extraction from text is done in cooperation with the Ontology Engineering Group of the Universidad Politécnica de Madrid[1]. The instrumental toolset is developed in cooperation with the BWT Group[2]. The industrial case study, focused on the Knowledge Management domain, is performed in cooperation with the internal LOD project of Springer-Verlag GmbH[3].

The remainder of the paper is structured as follows. Section 2 presents the motivation for this project based on the brief analysis of the related work. Section 3 briefly outlines the OntoElect approach. Section 4 describes our experimental setting in terms of objectives, instruments, datasets, and workflow. Section 5 presents our early results. Finally, the plans for the future work are discussed in Section 6.

2 Related Work and Motivation

Perhaps one of the most comprehensive sources surveying the existing approaches and techniques for ontology learning from text is [6]. Another collection of research contributions in ontology learning and population, complementary to this review, is

[1] http://www.oeg-upm.net/
[2] http://www.groupbwt.com/
[3] http://www.springer.com/

[7]. This is the research area which often combines linguistic and statistical methods to process text corpora and extract knowledge fragments in different forms: ranging from key phrases and their importance / frequency values (e.g. [3]) to simple ontology modules e.g. specified in SKOS [8]. The dominant approach to assess the quality of extracted knowledge is comparing the resulting artifact to a Gold Standard [9] in the domain. Golden Standards are however quite rarely available. Another way to evaluate if the result fits the domain requirements well is to check it against the set of competency questions [10], provided by the knowledge stakeholders in the domain. Unfortunately these experts are also not readily available in the vast majority of cases. Therefore an objective indirect method to extract knowledge for producing ontologies from a representative document collection for the domain is on demand. An important question to answer in this context is: what is the minimal subset of a (potentially very big) document collection which is terminologically complete in statistical terms? The project presented in this paper aims at developing such an experimental method based on the OntoElect approach for ontology development and refinement. It also aims at a thorough experimental evaluation of this method.

3 OntoElect Saturation Metric and Measurement

OntoElect, as a methodology, seeks for maximizing the fitness of the developed ontology to what the domain knowledge stakeholders think about the domain. Fitness is measured as the stakeholders' votes – a metric that allows assessing the stakeholders' commitment to the ontology under development - reflecting how well their sentiment about the requirements is met. The more votes are collected – the higher the commitment is expected to be. If a critical mass of votes is acquired (say 50%+1), the ontology is considered to satisfactorily meet the requirements.

It is well known that direct acquisition of requirements from domain experts is not very realistic as they are expensive and not really willing to do the work falling out of their core activity. So, in this project, we are focused on the indirect collection of the stakeholders' votes by extracting these from high quality and reasonably high impact documents authored by the stakeholders.

An important feature to be ensured for knowledge extraction from text collections is that the dataset needs to be statistically representative to cover the opinions of the domain knowledge stakeholders satisfactorily fully. OntoElect suggests a method to measure the terminological completeness of the document collection by analyzing the *saturation* of terminological footprints of the incremental slices of the document collection - as e.g. reported in [2]. The full texts of the documents from the retrospective collection are grouped in datasets in the order of their timestamps. The first dataset contains the first portion of documents. The second dataset contains the first dataset plus the second portion of documents. Finally, the last dataset contains all the documents from the collection. At the next step of the OntoElect workflow the bags of multi-word terms are extracted from all the datasets using TerMine software [3] together with their *significance* (C-value) scores reflecting how often a term was met in the dataset. The workflow, presented below in Section 4 also suggests using an alter-

native way to extract the bags of multi-word terms [4] for comparing the quality of term extraction. Further the bags of terms of adjacent datasets (1st and 2nd, 2nd and 3d, ...) are compared and the *termhood difference* (*thd*) value is computed for each consecutive pair of the datasets. Terminological *saturation* is assessed by comparing the overall *thd* to individual *term significance*[4] threshold. A dataset for which stable saturation is observed is further considered as *complete* (and statistically representative) for knowledge extraction.

It is also worth noting that the outlined approach is domain independent as far as the used term extraction solutions are domain independent.

4 Experimental Settings and Workflow

The objective of the presented experimental research project is to check if the Onto-Elect approach to assess the representativeness of a subset within a document collection, based on measuring terminological saturation, is valid. The setting of the experiments should consider several parameters which may influence the measurements and, therefore the results of measuring saturation. These parameters are taken into account while answering the following research questions:

Q1: Which would be the proper direction in forming the datasets to check saturation: chronological, reverse-chronological, bi-directional, random selection? Which direction is the most appropriate to cope with potential terminological drift in time?

Q2: Would frequently cited documents form a minimal representative subset of documents? Do the most frequently cited documents indeed provide the biggest terminological contribution to the document collection?

Q3: Would the size of a dataset increment influence saturation measurements? Is there an optimal size of a data chunk for the purpose?

Q4: Which of the term extraction solutions (UPM Extractor [4] or Manchester TerMine [3]) yield more adequate and quality sets of terms?

Q5: Is the method for assessing completeness based on saturation measurements valid? Does it indeed provide a correct indication of statistical representativeness?

The answers to the outlined research questions are sought based on conducting experiments in real world industrial settings. For that the document collection has been formed in cooperation with Springer. Based on the expert advice of the partner, fifteen Springer journals[5] have been selected that are broadly relevant to the domain of Knowledge Management[6].

---

[4] An extracted term is *significant* if its score puts it in the upper part of the scored list. The upper part forms the prevailing sentiment of the domain knowledge stakeholders - the majority vote - as it accumulates 50%+1 stakeholder votes in the terms of the sum of the normalized scores (C-values) of the respective terms.

[5] The list of the selected journals is available at: https://github.com/bwtgroup/SSRTDC-PaperCatalogues/blob/master/ListOfJournals.xls

[6] Knowledge Management has been chosen as a target domain because: (i) the methodology developed in the presented experimental study is for knowledge engineering and management; (ii) the partners in the presented project possess extensive expertise in Knowledge Management

The chosen collection of journal papers appears to be well suited to attack the outlined research questions. Indeed, it is formed of the journals scoping into different subfields of Computer Science in broad. The journals in the selection are however mutually complementary in terms of providing terminology related to Knowledge Management. So there seems to be a balance between the broadness of the overall scope and the focus on the target domain. This balance needs to be checked experimentally by verifying if it contains a saturated terminological footprint on the domain. Furthermore, individual journal collections chronologically start at very different times and contain quite different numbers of volumes, issues and papers. So, these internal disbalances may really help reveal the complications like terminological temporal drift and different terminological contributions caused by varying data volumes coming from different journals.

The experimental workflow is based on the OntoElect workflow described in Section 3 and is outlined in Fig. 1. This workflow could be generically applied (using Configure Experiment step) to perform all the series described below.

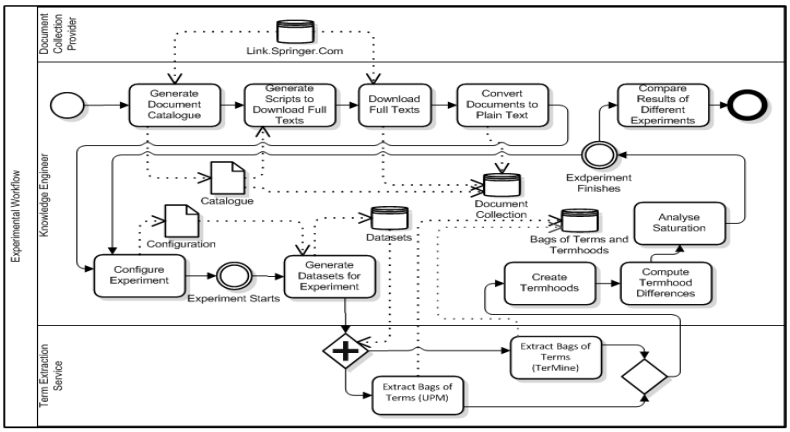

Fig. 1. Planned experimental workflow

Different kinds of experiments, using this workflow, are planned to be conducted in the presented study.

The first series of experiments is targeted at checking which direction of choosing papers for the datasets yields better saturated sets of terms and assesses terminological temporal drift. In this series the experimental workflow is applied to the datasets which are formed: (i) chronologically; (ii) reverse-chronologically; (iii) bi-directionally, i.e. including data increments containing the documents from both ends of the temporal span in turns (e.g. first issue, than last issue, than second issue, etc.); and (iv) including documents picked from the data collection uniformly randomly. Saturation measures and saturated sets of terms will be compared across these different choices. This series will allow answering **Q1**.

---

and therefore could be used as subject experts; (iii) there is a substantially big collection of high-quality full text documents broadly relevant to this domain available at Springer.

The second kind of experiment will base on the most appropriate selection direction choice, determined in the first series, and investigate the terminological impact of the frequently cited documents in the collection. For that, the impact of each document will be computed based on its citation frequency. The documents with impact equal to *n* will be replicated *n* times in the corresponding dataset. The experimental workflow will be repeated for these "impact" datasets and the results will be compared to the first series using "flat" datasets. The comparison will be done in terms of saturation measures and terminological contribution peaks [2]. This experiment may allow to answer **Q2** and extract the "decisive minority vote" subset of terms for Knowledge Management, contributed by the high-impact papers, as e.g. been done in [2] for Time Representation domain.

To answer **Q3**, the third series will focus on finding out what might be the optimal size of an increment to form experimental datasets. For this series, the datasets will be formed following the best selection direction discovered in the first series. The size of the increments will however be varying. Saturation measurements will be compared for different data increment sizes and the optimal value will be discovered if such an optimum does exist.

The fourth series is planned for experimental cross-evaluation of the available alternative software tools for multi-word term extraction from texts. Based on the datasets with the increments of optimal size determined in the series No 3, term extraction will be done separately using the UPM toolset and TerMine. The results will be compared in terms of saturation measures for flat datasets and decisive minority subsets of terms extracted from the impact datasets (series No 2). This may allow answering **Q4**. Perhaps, **Q5** is the most difficult question to answer and it still requires some thinking for offering a convincing method to assess the adequacy and validity of the experimental method investigated in the presented project. One possible way is to do that based on the cross-evaluation with another method for ontology learning, e.g. [5]. Another possible way is to select a much smaller subset of a document collection, e.g. only the papers with high terminological impact discovered in the series No 2. The set of terms extracted from this "decisive minority vote" subset could be manually checked by human experts.

5 Early Results

The project has been started in November 2016 and is in its initial phase. Since it has been started the following steps have been accomplished: (i) the document collection has been chosen; (ii) the catalogue of the papers in the document collection has been created; (iii) the full texts of the papers have been downloaded and converted to plain text format.

Overall the document collection contains more than 9 000 papers. The composition of the document collection is diagrammatically shown in Fig. 2. So, performing even those initial steps could not be done manually due to the volume and incurred manual effort. It has been therefore decided to develop some software instruments which help automate these routine steps.

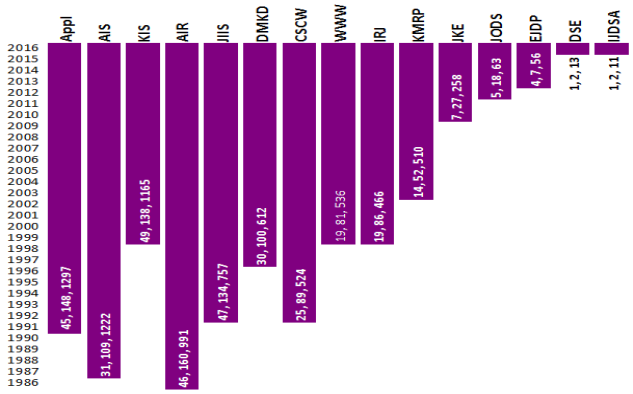

Fig. 2. Distribution of papers in the journals of the document collection. Y-axis shows the years of publication, X-axis corresponds to the journals. The numbers in the bars are: no of volumes, no of issues, the total no of papers in the journal.

For creating the catalogue of the papers which will further be used for generating datasets, a tailored parser[7] has been developed. The parser receives a Springer journal web page URL as its input and stores the list of all the papers of this journal in the specified .csv file[8]. The information about a paper contains all its reference information, the abstract, and the no of citations acquired from Google Scholar. For downloading the full texts of the papers another software module has been developed[7]. It receives a .csv list of papers to be downloaded and generates a script to download the full texts of the papers based on their DOI information taken from the catalogue. The papers in PDF are stored in a folder specified as a parameter. One more software module has been developed[7] for batch conversions of paper full texts in PDF to plain text. It gets a path to the directory where PDF articles are stored, as a parameter. It produces the outputs for each input file in plain text format in which hyphenations are removed and each sentence occupies a separate line for better term extraction.

Our next step is generating the incremental datasets using the plain texts of the papers in the directions and with increment sizes as specified in Section 4 for the first and third series of our experiments. The software for generating these datasets is currently under development.

6 Conclusive Remarks and Future Work

The PhD project presented in this paper is at an early stage. It currently focuses on the

[7] All the developed instrumental software modules are available at: https://github.com/bwtgroup/SSRTDC-Springer-article-parser, https://github.com/bwtgroup/SSRTDC-Collections-Springer-PDF-Downloader, https://github.com/bwtgroup/SSRTDC-PDF2TXT

[8] The catalogues of the acquired journal papers in .XLSX format are available at: https://github.com/bwtgroup/SSRTDC-PaperCatalogues/. The data has been collected on December 3-4, 2016.

detailed planning of experiments, developing software for the instrumental support of the experimental workflow, and preparing the data collection. The early results have been reported in Sections 4 and 5.

The short-term plans for the future work include: (i) further development of the instrumental software to support all the steps in the experimental workflow; (ii) the performance of the experimental study as outlined in the presented experimental setup; (iii) the assessment of the efficiency of the developed software. The analysis of the short-term results may further lead to a better understanding of a model and metric for the completeness of a document collection for knowledge extraction. So, in the mid-term, based on this refined understanding, the objectives of the study may be specified in a more detailed manner unfolding into a refined experimental setup and possibly leading to new kinds of experiments.

Acknowledgements

The research leading to this paper has been done in part in frame of FP7 Marie Curie IRSES SemData project (http://www.semdata-project.eu/), grant agreement No PIRSES-GA-2013-612551.

References

1. 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 412, pp. 136--162 (2013)
2. Ermolayev, V., Batsakis, S., Keberle, N., Tatarintseva, O., Antoniou, G.: Ontologies of Time: Review and Trends. Int. J. of Computer Science & Applications. 11(3), 57--115 (2014)
3. Frantzi, K., Ananiadou, S. and Mima, H.: Automatic Recognition of Multi-Word Terms. Int. J. of Digital Libraries 3(2), pp.117-132 (2000)
4. Corcho, O., Gonzalez, R., Badenes, C., and Dong, F.: Repository of indexed ROs. Deliverable No. 5.4. Dr Inventor project (2015)
5. Osborne, F., Salatino, A., Birukou, A., Motta, E.: Automatic Classification of Springer Nature Proceedings with Smart Topic Miner. In: Groth, P. et al. (Eds.) ISWC 2016, LNCS 9982, pp. 383-399 (2016)
6. 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 (2012)
7. Buitelaar, P, Cimiano, P. (eds.): Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press (2008)
8. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System reference. Technical report, W3C (2009)
9. Zavitsanos, E., Vouros, G. A., and Paliouras, G.: Gold Standard Evaluation of Ontology Learning Methods through Ontology Transformation and Alignment. IEEE Trans. on Knowledge & Data Engineering, 23(11), 1635--1648 (2011)
10. Ren, Y., Parvizi, A., Mellish, C., Pan, J. Z., van Deemter, K., and Stevens, R.: Towards Competency Question-driven Ontology Authoring. In: Presutti, V. et al. (Eds.) ESWC 2014, LNCS 8465, pp. 752--767 (2014)