Course Title: Automated Term Extraction and Ontology Learning from Texts

Level: MSci

Duration: 24 hrs

The Scope of the Course

This short MSci-level course covers one of the most vibrantly developed areas on the crossroads of Text Mining and Ontology Engineering, positioned as Ontology Learning from Texts. It gives a beginner's, though professional and up-to-date introduction into:

(i) What are ontologies and why one needs (to be at least knowledgeable about) ontologies for being successful in Data Science, in particular in Data Analytics, and related disciplines?

(ii) What are the knowledge sources for building ontologies and why a representative collection of high-quality professional texts is a right sublimation for these sources?

(iii) How to elicit the bits of the required knowledge from texts? Why Automated Term Extraction (ATE, or Recognition, ATR) is a relevant approach? What is the ATE processing stack? How would Linguistic and Statistical Processing be reasonably married together for increasing the quality of ATE?

(iv) What is Linguistic Processing (LP) in the context of ATE? How would a generic LP technology stack and workflow look like? Which Natural Language Processing techniques are relevant for LP in ATE?

(v) What is Statistical Processing (SP) in the context of ATE? How would a generic SP technology stack and workflow look like? Which Statistical Processing techniques are relevant for SP in ATE? How would LP and SP be rationally combined in ATE?

(vi) How would a statistically representative subset be identified in a document collection for ATE? What is Terminological Saturation (TS)? How can TS be measured? Which factors influence TS? How would a minimal subset of documents, representing the decisive minority sentiment, be identified?

(vii) Provided that the set of terms is extracted, what are the next steps in learning the ontology based on these terms? How would a generic ontology learning workflow look like? – This will be presented based on the example of the OntoElect methodology for ontology refinement.

(viii) What are the highlights and pitfalls in the field of Ontology Learning from Texts? Why do fully automated approaches fall short?

Didactics

The course is given in the form of tutorials with a hands-on practical component taking ~50 percent of teaching time. After each lecture, except the introductory one, the students are offered to:

- Use the instrumental software and the document collection(s) / dataset(s) provided by the tutor
- Refine the software in some advised way, e.g. by introducing a more sophisticated metric or an improvement in an algorithm
- Perform a cross-evaluation experiment to compare the initial revision of the software and their refined revision

These practical tasks are organized in a way to finally assemble a simple instrumental tool suite that helps performing a basic ATE workflow.

The final slot of the course is organized as a cross-evaluation contest for the solutions by students. They are offered to apply their tool suites to the same document collection and measure the quality of ATE results. The ranked list of the solutions is built based on the comparison of these results.

Course Topics

No	Topic	Dura-tion (hrs)
1	**Introduction: Data, Analytics, Ontologies, Text Mining, and ATE**: It is widely agreed that "a picture is worth a thousand words"[1]. When it comes to Data Science, in particular to Big Data analytics, a good picture may be worth of petabytes, saving weeks of data analysts' work. An adequate and timely representation of knowledge dissolved in and carried by the petabytes of data may be even more effective than a good picture. Indeed, data emerge reflecting a change in the world. Big Data is consequently a fine-grained reflection of the changes around us. Knowledge extracted from these data in an appropriate and timely way is an essence of understanding a change. So, adding a semantic layer to Big Data processing stack, by operationalizing this extracted knowledge, is a valid requirement and is timely. Ontologies – descriptive formal theories for a domain – are the main building blocks for such a semantic layer. There is a discipline in Computer Science (Artificial Intelligence), called Ontology Engineering, that provides the methods and instruments to develop ontologies. It is, however, a consensual opinion that building an ontology is a challenge – due to several valid reasons. This topic will overview these reasons and point out that the major challenge and pitfall is requirements elicitation. One way to bypass this pitfall is to mine the available domain data for the bits of knowledge. The extracted bits, to be trusted, have to be of high quality and represent the sentiment of the majority of the experts in the domain. Hence, what data to mine is also a very important question to answer. This part of the course suggests that using the collection of professional texts authored by the recognized experts in the domain would be a proper starting point. Having identified such a collection, the workflow for requirements elicitation may be started. The key step here is the extraction of a statistically representative set of terms that feature the domain in focus. For practically interesting domain, like Data Analytics, Financial Data Stream Processing, or alike a representative document collection may be too big for manual term recognition. Therefore, ATE, as a sub-discipline of Text Mining, has to be used as an enabler. The topic will be completed by a brief overview of the basic ideas in ATE. The tools using these ideas will also be mentioned.	4
2	**Linguistic and Statistical Methods and Metrics for ATE**: Despite being important for practice, ATE is still far from being reliable. New approaches to ATE are being proposed and still demonstrate their precision at the level below 80 percent. In the majority of approaches to ATE, processing is done in two consecutive phases: Linguistic Processing and Statistical Processing. Linguistic processors, like POS taggers or phrase chunkers, filter out stop words and restrict candidate terms to n-gram sequences: nouns or noun phrases, adjective-noun and noun-preposition-noun combinations. Statistical processing is then applied to measure the ranks of the candidate terms. These measures are either the measures of 'unithood', which focus on the collocation strength of units that comprise a single term; or the measures of 'termhood' which point to the association strength of a term to domain concepts. For 'unithood', the metrics are used such as: mutual information, log likelihood, t-test, or the notion of 'modifiability' and its variants. The metrics for 'termhood' are either term frequency-based (unsupervised approaches) or reference corpora-based (semi-supervised approaches). The most used frequency–based metrics are: TF/IDF; weirdness which compares the frequency of a term in the evaluated corpus with that in the reference corpus; domain pertinence. More recently, hybrid approaches were proposed, that combine 'unithood' and 'termhood' measurements in a	4 2 lec + 2 hands-on

[1] This saying is attributed to Napoleon Bonaparte. John McCarthy supported a literally opposite opinion however. He mentioned: "As the Chinese say, 1001 words worth more than a picture." (www-formal.stanford.edu/jmc/sayings.html)

No	Topic	Dura-tion (hrs)
	single value. A representative metric is c/nc-value. **The lecture part** of this topic offers a comparative overview of the ATE metrics and methods associated with the use of these metrics. The tools based on the surveyed methods are also mentioned. **The hands-on part** of this topic offers a small plain text dataset in English and a set of alternative methods for ATE, with one example of implementation (c-value based method). The students are tasked to: (i) Develop an alternative method of their choice based on the description available in a suggested publication and lecture material (ii) Run a cross-evaluation experiment using the provided implementation of the c-value method and their own implementation of the alternative method on the offered dataset	
3	**ATE Workflow: Phases and Tools**: ATE, being the initial step in ontology learning, in turn requires substantial pre-processing. The workflow resulting in the ranked list of automatically extracted terms is explained in this course topic. It contains several phases – two for pre-processing and one for ATE. The pre-processing phases are: (i) collecting documents; and (ii) generating dataset(s) for ATE. Collecting documents may be done manually, if for example the document collection could have been easily identified for a domain, or the stakeholders offer their recommendations. Otherwise, a few prominent documents regarding the domain are selected and the rest of the collection is identified using the automated techniques like snowball sampling. One of the research prototypes for snowball sampling will be presented as a part of this topic. The result of this initial phase is the repository of the full texts of the documents and the catalogue of this repository. Phase (ii) pre-processes the documents provided quite frequently in PDF. The workflow does (ii.i) conversion to plain text; (ii.ii) some pattern-based cleaning; (ii.iii) dataset generation. This part of the lecture will present the details in each of the pre-processing steps that require attention. After the datasets are generated, the ATE phase is performed. This may be done using either your own term extractor software implementation, or one of the publicly available software tools like NaKTeM TerMine or UPM Term Extractor. The output of this step is the ranked list of multi-word terms. Among those terms, it is useful to distinguish: (i) valid and significant terms; (ii) insignificant terms; and (iii) noise. The latter two categories are often filtered out at the cut-off post processing step. Furthermore, several terms in the list, though spelled differently, may mean the same. So, term grouping needs to be done for the valid and significant terms. One of the techniques for that is the use of string similarity measures with appropriate thresholds. **The hands-on part** of this topic tasks students to develop software modules for: (i) **Document collection**. Given the collection of a few seed PDF documents in a directory explore the citation network of these documents, find and download additional documents using snowball sampling. Generate the catalogue of the collected documents using the template. (ii) **Document pre-processing**. Convert all the collected documents to plain texts. While converting correct typical PDF encoding errors (missing “fi”, “fl”, etc.). Output each individual sentence starting from a new line. (iii) **Dataset generation**. Develop the module that takes in the documents as plain texts and puts these together as one output plain text dataset file. Allow configuring different orders of adding documents and the total number of input files to be taken in. (iv) **Term grouping**. Develop the module that takes in the ranked list of terms and returns it with terms grouped based on the string similarity measured using one of the proposed measures. Students are also tasked to **test** their **developed software** using 5 most frequently cited seed documents from the TIME collection. TIME document collection is provided as a resource.	4 2 lec + 2 hands-on
4	**Terminological Saturation in Document Collections**: Document collections describing real world domains, like for example Knowledge Management, can be huge in volume. So, the resource to be spent for ATE and pre-	4 2 lec + 2 hands-on

No	Topic	Dura-tion (hrs)
	processing, could be significant. Furthermore, the lists of terms extracted from such document collections can grow quite big to contain dozens of thousands of significant entries. Therefore, the valid task regarding these results is to find out what is the representative sub-collection of a minimal possible size that provides a statistically representative set of terms that describe the domain satisfactorily fully. A technique to extract such a sub-collection of a minimal size is based on measuring terminological saturation. The sub-collection is grown by adding the portions of several documents from the entire collection. The growth is stopped when the changes in the extracted sets of terms become neglectable – so the terminological footprint of the collection saturates. This topis will explain that terminological saturation depends on several factors like the size of the added document chunk, the order of adding documents, the impact of the chosen documents, etc. Recommendations will be given on how to form the datasets to achieve saturation in a quicker and smoother way. **The hands-on part** of this topic tasks students to: (i) Develop the module for measuring terminological difference, given the pair of files containing the ranked lists of terms (ii) Using the instrumental software and documents of Topic 4, generate the sequence of incrementally enlarged datasets, extract terms from these datasets, measure terminological differences in the consecutive pairs of the ranked lists of terms, evaluate terminological saturation, and identify the minimal representative sub-collection of documents. (iii) Compare saturation measurements with and without term grouping	
5	**Ontology Learning from Texts using OntoElect:** ATE is just the first step in learning an ontology from the requirements of domain knowledge stakeholders, dissolved in their texts. This topic of the course explains what should be done further to build an ontology, given the set of the ranked domain terms. It also presents how to evaluate if an ontology fits the requirements. This discussion is based on the use of the OntoElect methodology for ontology refinement. It demonstrates that a thorough routine for indirect elicitation, ensuring completeness, correctness of interpretation, using the requirements in ontology evaluation is a must for Ontology Engineering. Such a routine is also valid in the cases of the use of a very high-profile expert working groups elaborating requirements for ontology refinement. The workflow of OntoElect contains three phases: Feature Elicitation, Requirements Conceptualization, and Ontology Evaluation. It elicits the set of terms extracted from a saturated collection of documents in the domain. It further sublimates these terms to the set of required features using the information about term significance in the form of numeric scores. Further, it applies conceptualization and formalization activities to these features yielding their aggregations as ontological fragments interpreted as formalized requirements. Finally, the mappings are specified between the elements in the requirements and ontology elements. The scores are used in the mappings to indicate the strength of positive or negative votes regarding the evaluated ontology. The sum of the votes gives the overall numeric measure of the fitness of the ontology to the domain requirements. The paper presents the use of OntoElect in the use case of evaluating the W3C OWL-Time ontology against the requirements extracted from the proceedings of the TIME symposia series. **The hands-on part** of this topic offers students to develop a module for building the feature taxonomy with propagated significance scores from the ranked list of terms. It is suggested that taxonomic relationships are discovered using term grouping.	4 2 lec + 2 hands-on
6	**Students' ATE Solutions Cross-Evaluation Contest:** This is the final (exam) event in the course which is obligatory for those students who intend to receive credits for the course. It will be organized, either online or as a separate face to face meeting, in **2 weeks after** the course (Topics 1-5) is finished. Students are tasked (based on their work in hands-on parts of Topics 2-5) to assemble their own software suite for ATE and terminological saturation measurement. At the contest, the students will be offered to apply their software to the given collection of documents for pre-processing, term extraction, and terminological saturation measurement. The results of their	4

No	Topic	Dura-tion (hrs)
	runs will be evaluated by the tutors based on the set of quality metrics presented in the course. Final grades (and credits) will be given based on the results in the contest.	
	Total:	24

Textbooks and Other Reading

Pre-reading:

Pre-reading is not a requirement, but an advise of a practically helpful literature.

Practical tips on doing NLP in Python

- Steven Bird, Ewan Klein, Edward Loper: **Natural Language Processing with Python**. **Analyzing Text with the Natural Language Toolkit**. O'Reilly Media (2009) 1st Ed. available at http://www.nltk.org/book/
- Deepti Chopra, Nisheeth Joshi, Iti Mathur: **Mastering Natural Language Processing with Python**. Packit Publishing (2016)

(Text)books:

Diana Maynard, Kalina Bontcheva, and Isabelle Augenstein: **Natural Language Processing for the Semantic Web**. Morgan & Claypool (2017) ISBN paperback: 9781627059091; ISBN ebook 9781627056328; DOI: 10.2200/S00741ED1V01Y201611WBE015

> This book introduces core natural language processing (NLP) technologies to non-experts in an easily accessible way, as a series of building blocks that lead the user to understand key technologies, why they are required, and how to integrate them into Semantic Web applications. Natural language processing and Semantic Web technologies have different, but complementary roles in data management. Combining these two technologies enables structured and unstructured data to merge seamlessly. Semantic Web technologies aim to convert unstructured data to meaningful representations, which benefit enormously from the use of NLP technologies, thereby enabling applications such as connecting text to Linked Open Data, connecting texts to each other, semantic searching, information visualization, and modeling of user behavior in online networks.
>
> The first half of this book describes the basic NLP processing tools: tokenization, part-of-speech tagging, and morphological analysis, in addition to the main tools required for an information extraction system (named entity recognition and relation extraction) which build on these components. The second half of the book explains how Semantic Web and NLP technologies can enhance each other, for example via semantic annotation, ontology linking, and population. These chapters also discuss sentiment analysis, a key component in making sense of textual data, and the difficulties of performing NLP on social media, as well as some proposed solutions. The book finishes by investigating some applications of these tools, focusing on semantic search and visualization, modeling user behavior, and an outlook on the future.

Hendrik J. Kockaert, Frieda Steurs (Eds.): **Handbook of Terminology**. **Volume 1**. John Benjamins Publishing Co., Amsterdam/Philadelphia (2015)

> The Handbook of Terminology (HOT) aims at disseminating knowledge about terminology (management) and at providing easy access to a large range of topics, traditions, best practices, and methods to a broad audience: students, researchers, professionals and lecturers in Terminology, scholars and experts from other disciplines, such as linguistics, life sciences, metrology, chemistry, law studies, machine engineering, and any other expert domain. In addition, the HOT addresses experts in (multilingual) terminology, translation, interpreting, localization, editing, etc., such as communication specialists, translators, scientists, editors, public servants, brand managers, engineers, and (intercultural) organization specialists.

Christopher D. Manning, Hinrich Shuetze: **Foundations of Statistical Natural Language Processing**. The MIT Press (1999)

Statistical approaches to processing natural language text have become dominant in recent years. This foundational text is the first comprehensive introduction to statistical natural language processing (NLP) to appear. The book contains all the theory and algorithms needed for building NLP tools. It provides broad but rigorous coverage of mathematical and linguistic foundations, as well as detailed discussion of statistical methods, allowing students and researchers to construct their own implementations. The book covers collocation finding, word sense disambiguation, probabilistic parsing, information retrieval, and other applications.

Paul Buitelaar, Philipp Cimiano: **Ontology Learning and Population: Bridging the Gap between Text and Knowledge**. IOS Press (2008)

The promise of the Semantic Web is that future web pages will be annotated not only with bright colors and fancy fonts as they are now, but with annotation extracted from large domain ontologies that specify, to a computer in a way that it can exploit, what information is contained on the given web page. The presence of this information will allow software agents to examine pages and to make decisions about content as humans are able to do now. The classic method of building an ontology is to gather a committee of experts in the domain to be modeled by the ontology, and to have this committee agree on which concepts cover the domain, on which terms describe which concepts, on what relations exist between each concept and what the possible attributes of each concept are. All ontology learning systems begin with an ontology structure, which may just be an empty logical structure, and a collection of texts in the domain to be modeled. An ontology learning system can be seen as an interplay between three things: an existing ontology, a collection of texts, and lexical syntactic patterns. The Semantic Web will only be a reality if we can create structured, unambiguous ontologies that model domain knowledge that computers can handle. The creation of vast arrays of such ontologies, to be used to mark-up web pages for the Semantic Web, can only be accomplished by computer tools that can extract and build large parts of these ontologies automatically. This book provides the state-of-art of many automatic extraction and modeling techniques for ontology building. The maturation of these techniques will lead to the creation of the Semantic Web.

Edited Collections:

Chris Biemann, Alexander Mehler (eds.): **Text Mining. From Ontology Learning to Automated Text Processing Applications**. Springer International Publishing Switzerland (2014)

This book comprises a set of articles that specify the methodology of text mining, describe the creation of lexical resources in the framework of text mining and use text mining for various tasks in natural language processing (NLP). The analysis of large amounts of textual data is a prerequisite to build lexical resources such as dictionaries and ontologies and also has direct applications in automated text processing in fields such as history, healthcare and mobile applications, just to name a few. This volume gives an update in terms of the recent gains in text mining methods and reflects the most recent achievements with respect to the automatic build-up of large lexical resources. It addresses researchers that already perform text mining, and those who want to enrich their battery of methods. Selected articles can be used to support graduate-level teaching.

The book is suitable for all readers that completed undergraduate studies of computational linguistics, quantitative linguistics, computer science and computational humanities. It assumes basic knowledge of computer science and corpus processing as well as of statistics.

Theses:

Chetan Arora: **Automated Analysis of Natural Language Requirements using Natural Language Processing**. PhD Thesis. Faculty of Sciences, Technology and Communication, University of Luxembourg (2016)

Natural Language (NL) is arguably the most common vehicle for specifying requirements. This dissertation devises automated assistance for some important tasks that requirements engineers need to perform in order to structure, manage, and elaborate NL requirements in a sound and effective manner. The key enabling technology underlying the work in this dissertation is Natural Language Processing (NLP). All the solutions presented herein have been developed and empirically evaluated in close collaboration with industrial partners.

The dissertation addresses four different facets of requirements analysis:

• Checking conformance to templates. Requirements templates are an effective tool for improving the structure and quality of NL requirements statements. When templates are used for specifying the requirements, an

important quality assurance task is to ensure that the requirements conform to the intended templates. We develop an automated solution for checking the conformance of requirements to templates.

• Extraction of glossary terms. Requirements glossaries (dictionaries) improve the understandability of requirements, and mitigate vagueness and ambiguity. We develop an automated solution for supporting requirements analysts in the selection of glossary terms and their related terms.

• Extraction of domain models. By providing a precise representation of the main concepts in a software project and the relationships between these concepts, a domain model serves as an important artifact for systematic requirements elaboration. We propose an automated approach for domain model extraction from requirements. The extraction rules in our approach encompass both the rules already described in the literature as well as a number of important extensions developed in this dissertation.

• Identifying the impact of requirements changes. Uncontrolled change in requirements presents a major risk to the success of software projects. We address two different dimensions of requirements change analysis in this dissertation: First, we develop an automated approach for predicting how a change to one requirement impacts other requirements. Next, we consider the propagation of change from requirements to design. To this end, we develop an automated approach for predicting how the design of a system is impacted by changes made to the requirements.

Papers:

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)

Dobrovolskyi H., Keberle N., Todoriko O.: Probabilistic Topic Modelling for Controlled Snowball Sampling in Citation Network Collection. In: Różewski P., Lange C. (eds) Knowledge Engineering and Semantic Web. KESW 2017. CCIS, vol 786. Springer, Cham (2017)

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.) Revised Selected Papers of ICTERI 2017, CCIS, vol. 826, pp. 132--160 (2018)

Ermolayev, V., OntoElecting Requirements for Domain Ontologies: The Case of Time Domain. EMISA J (2018) - to appear

Chugunenko, A., Kosa, V., Popov, R., Chaves-Fraga, D., Ermolayev, V: Refining Terminological Saturation using String Similarity Measures. Submitted to: ICTERI 2018 (2018)

Technical Reports:

Kosa, V., Chaves-Fraga, D., Naumenko, D., Yuschenko, E., Badenes-Olmedo, C., Ermolayev, V., and Birukou, A.: Cross-Evaluation of Automated Term Extraction Tools. Technical Report TS-RTDC-TR-2017-1, 30.09.2017, Dept of Computer Science, Zaporizhzhia National University, Ukraine, 61 p.

Kosa, V., Chaves-Fraga, D., Naumenko, D., Yuschenko, E., Badenes-Olmedo, C., Ermolayev, V., and Birukou, A.: The Influence of the Order of Adding Documents to Datasets on Terminological Saturation. Technical Report TS-RTDC-TR-2018-1, 20.01.2018, Dept of Computer Science, Zaporizhzhia National University, Ukraine, 60 p.

Resources

To be announced later …

Pre-requisites

Practical proficiency in Python programming

Basic Knowledge of Information Retrieval

Basic knowledge of elementary probability theory and statistics