TERMINOLOGICAL AND LEXICAL RESOURCES USED TO PROVIDE OPEN MULTILINGUAL EDUCATIONAL RESOURCES

BILJANA LAZIĆ
University of Belgrade, Faculty of Mining and Geology, biljana.lazic@rgf.bg.ac.rs

DANICA SENIČIĆ
Université catholique de Louvain, Faculté de philosophie, arts et lettres, danica.senicic@student.uclouvain.be

ALEKSANDRA TOMAŠEVIĆ
University of Belgrade, Faculty of Mining and Geology, aleksandra.tomasevic@rgf.bg.ac.rs

BOJAN ZLATIĆ
University of Belgrade, Faculty of Mining and Geology, bojan.zlatic@rgf.bg.ac.rs

Abstract: *Open educational resources (OER) within BAEKTEL (Blending Academic and Entrepreneurial Knowledge in Technology enhanced learning) network will be available in different languages, mostly in the languages of Western Balkans, Russian and English. University of Belgrade (UB) hosts a central repository based on: BAEKTEL Metadata Portal (BMP), terminological web application for management, browse and search of terminological resources, web services for linguistic support (query expansion, information retrieval, OER indexing, etc.), annotation of selected resources and OER repository on local edX platform. In order to successfully cope with multilingualism within the network, especially where terminology is concerned, a language support system is developed within the BAEKTEL metadata portal. In this paper we will describe the linguistic component of the system, the resources and tools used as an educational system as a whole and to improve the visibility of resources in the Internet. This component consists of morphological dictionaries, WordNet, domain specific terminological resources such as GeolISSterm, RudOnto, aligned texts in TMX format, corpora etc. Special attention will be given to Termi, newly developed application for terminology management.*

Keywords: *Open Educational Resources, Lexical resources, Natural Language Processing, Terminology*

1. INTRODUCTION

Natural Language Processing (NLP) has a two-faceted approach to education where one involves e-learning and computer-assisted learning and instruction and the other consists of NLP tools for analysis and use of language by machines [1].

The usage and application of the research done in the field of the NLP has been present in the domain of education from the 1960s. One of the first advances made in this direction was the pioneering work of Ellis Batten Page who is considered to be the father of automated essay scoring. With the increasing number of students attending universities and numerous possibilities provided by e-learning applications, the Technology Enabled Assessment (TEA) has shown significant growth as well. Further on, Intelligent Tutoring Systems (ITS) were developed and incorporated in the learning process, while later work also included spoken language technologies. The advances in these fields allowed for a more time effective assessment through TEA, which is a considerable advantage for both students and teachers, immediate constructive feedback for learners through ITS, with further enhancements with the development of spoken language technologies.

One of the major examples of applied NLP in e-learning are the open-access MOOC[1] platforms which are changing the face of distant learning and education altogether by erasing geographical and spatial constraints, leaving the traditional education model behind [2]. Interactive forums and teaching assistants rely greatly on various NLP tools to help them cater to a large number of students from all over the world. These tools may include assessment of text and speech, writing assistants, automatic generation of exercises, wrap up questions and online instructional environments [3]. The main goal of NLP tools in education is to automate time-consuming, laborious teachers' tasks such as curriculum creation or assignment assessment and do so in a timely manner. Time a teacher can spend with a student is usually very limited, with a detriment to students, resulting in insufficient interaction and feedback. These tools help overcome these hurdles at an advantage for both

[1] Massive Open Online Course - an online course aimed at unlimited participation and open access via the web

students and educators. The model of digital education also allows for a modern peer-to-peer education where students educate themselves and each other, exploring, developing and building skills without constant input from or intervention by teachers [2]

It is important to note that NLP requires multidisciplinary collaboration in all domains of its application. Other than indispensable intertwinement found on the crossroad between linguistics, psycholinguistics, computer science, engineering and statistics, as we go more in-depth, experts from more narrow fields are required. For example, NLP tools for language learning must connect to Second Language Acquisition (SLA) and Foreign Language and Teaching (FLTL) research with insights from Cognitive Psychology and Empirical Educational Science.

This paper will more thoroughly introduce how the terminology and ontologies are used in combination with NLP tools for the purpose of education. Exactly due to the tendency of global, digitized, education, it is of great importance that the terminology is acquired systematically in all languages involved, which would lead to equivalent opportunities and up to date education materials.

Firstly, a brief history and current state of the art of terminological resources are presented, followed by an overview of BAEKTEL (Blending Academic and Entrepreneurial Knowledge in Technology enhanced learning) resources, lexical resources, the process of terminology extraction and a presentation of TERMI, an application for terminology management.

2. TERMINOLOGICAL RESOURCES

Terminology is considered to be a young interdisciplinary scientific field. The interest in it arose in 1930s when an electrical engineer, Eugen Wüster, became engaged in publishing papers concerning terminology as an individual discipline. Its interdisciplinarity involves linguistics, more precisely, lexicography, cognitive and communication sciences, but also disciplines from different areas, e.g. mining or mathematics.

Definition of terminology varies from one dictionary to another. Macmillan Dictionary defines it as “the words and phrases used in a particular business, science, or profession”[2]. According to ISO 12620 terminology is “The set of designations belonging to the language of a given subject field”.[3] Terminological theory arose through practical experience and as such was supported by information sciences. There is widespread theory of four basic periods in development of terminology, the origins, the structuring of the field, the boom and the expansion [4].

Those stages are closely related with the development of computers. Consequently, we have different stages of computer usage, from input terminals through processing terminological data with personal computers, to the Internet expansion. The last one ensured the infrastructure for online terminological resources, such as electronic dictionaries and term bases, which can be monolingual, bilingual or multilingual. Additionally, it strengthened the need to standardize the one-profession-vocabulary, because of rapid development in scientific research which constantly produces new terms which need to be translated in other languages. There is a huge amount of texts available on the Internet which is growing daily and needs to be translated for different purposes, at the same time paying attention to terminology rules which regulate the choice of the most appropriate term. Inevitabely, this requires standardisation so more accurate translations are produced. To summarize above mentioned, terminology now constitues a very important field of Natural Language Processing whilethe work that has been done in the field of terminologyhas become to be an indespensible, widespread used resource.

The standards related to terminology management are often used by the localization and translation industry as well as public translation and terminology units and organizations.

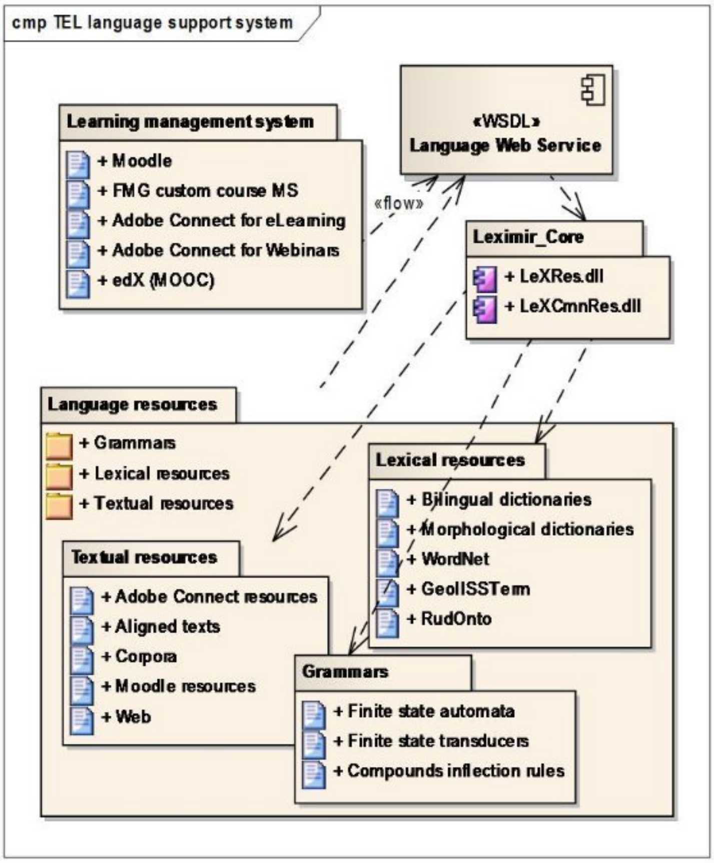

Image 1: BAEKTEL language support system [5]

3. BAEKTEL

To enable productive multilingual cooperation, open educational resources (OER) produced within BAEKTEL project will be available in different languages, mostly in languages of Western Balkans, Russian and English [6].

University of Belgrade (UB) hosts a central repository based on:

[2] Available at: http://www.macmillandictionary.com/dictionary/british/terminology

[3] Available at: http://www.isocat.org/rest/dc/4024

- BAEKTEL Metadata Portal[4] (BMP)
- terminological web application for management, browse and search of terminological resources,
- web services for linguistic support (query expansion, information retrieval, OER indexing, etc.),
- annotation of selected resources,
- OER repository on local edX platform.

The BAEKTEL language support system consists of several software components administrating in the same time language resources: grammars, lexical and textual resources (Image 1).

4. LEXICAL RESOURCES

Morphological dictionaries are meant to be used by computers in the process of query expansion. Their usage is necessary because of the rich flexion of Serbian language and other similar languages of Western Balkans. Partners in BAEKTEL project produce materials in Serbian, Bosnian and Montenegrin language. When it comes to morphology, the aforementioned languages are quite similar, therefore it is possible to use the Serbian morphological dictionary. Serbian morphological dictionaries include semantic markers which allow the distinction between ijekavian, ekavian and ikavian pronunciation. Dictionaries cover both general lexica and proper names. Serbian morphological dictionaries are found in LADL (Laboratoire d'Automatique Documentaire et Linguistique) format. There are two types of dictionaries: dictionary of simple words and dictionary of compounds.

Two main components of dictionary of simple words are DELAS and DELAF. Here we have an entry found in Serbian dictionary of simple words: **učiteljica, N651+Hum+GM:fs4v**. The first part of entry is a lemma: učiteljica. N is a sign noun (part of speech), 651 is an inflectional class, +Hum is a marker for human entity and +GM is a marker for gender. After that, there is a part of entry for grammatical categories. F is gender feminine, s is sign for number - singular, 4 is code for accusative case and finally v is code that describes animacy, in this case animate.

The main components of dictionary of compounds are DELAC and DELACF. Entry the compound dictionary: **lekar(lekar. N2:ms1v) akušer (akušer. N2:ms1v),NC_NXN+Comp+Hum** where we can find descriptions of two words. Description given in brackets describes grammatical categories of simple words. Lekar is a noun, male, singular, nominative case and animate. Akušer is also male, singular, nominative case and animate noun. There are markers for a compound noun and human entity.

According to data from 2014, Serbian morphological dictionary of simple words consists of 133,361 lemmas. Their production is 4,581,657word forms. The number of units covered by Serbian morphological dictionary of compounds is 13,717, or 262,686 word forms [7].

RudOnto and GeolISSTerm are developed at the Faculty of Mining and Geology, University of Belgrade [8].

RudOnto is a terminological resource that covers domain of mining. It is organized as a taxonomy of terms. Each term is followed by a definition, its synonyms, and bibliographical reference to their source, as well as equivalent terms in other languages.

GeolISSTerm[5] is a thesaurus of geological terms with entries in Serbian and English, developed as a part of the GeolISS project. It contains more than 3000 dictionary entries.

Another important lexical resource is the Serbian WordNet. A WordNet network is composed of synsets, or sets of synonymous words representing a concept, with basic semantic relations between them forming a semantic network. Each synset word is denoted by a “literal string” followed by a “sense tag” which represents the specific sense of the literal string in that synset. Interlingual index (ILI) enables the connection of the same concepts in different languages. The Serbian WordNet covers about 20 000 synsets [9] from different specific domains (e. g. law, biomedicine, mythology, culinary etc.).

5. TERMINOLOGY EXTRACTION

Bearing in mind rapid changes in scientific domains and new terms production, automatic terminology recognition and extraction has become an important task. The extracted terms are then included in ontologies.

Even though attention has been raised to the fact that many authors from the research communities have different perceptions of what ontology in this sense encompasses [10], a definition brought by [11] roughly sums up that the ontology is a term used to refer to the shared understanding of some domain interest which may be used as a unifying framework to solve the problems of poor communication and surpass the difficulties in identifying requirements. Thus, an ontology can be perceived as a set of concepts, their definitions and inter-relationships. Applications of such ontologies, alongside with the automatic term extraction, which will be further discussed, can be found in machine translation, automatic indexing, building lexical knowledge bases and information retrieval [12]. Once they are extracted, completed ontologies represent an important education resource.

In order for these ontologies to be as efficient as possible, it is of great importance that they are constantly reviewed and updated. In the today's world of rapid changes in technology and information exchange, taking up such enterprise manually is an incredibly laborious and time-consuming task. Traditionally, this kind of undertaking would be done by a terminologist who would list potential term candidates to include in the ontology and would then proceed by consulting a domain expert to arrive at a final list of validated terms [13]. Other problem that also arises is that such lists, based solely on human assessment, are often being questioned among experts and have the risk of being unsystematic and subjective [12], [14]. Automatic term extraction is a process that is meant to facilitate this painstaking task and identify terms less obvious to humans by using computer aided techniques. For now, the automatic extraction is used as a preliminary process, to

[4] Available at: http://meta.baektel.eu/

[5] Available at the address: http://geoliss.mprrpp.gov.rs/term/

identify term candidates, but is expected to replace manual term extraction completely.

Due to the rich morphology of Serbian language and the complexity of terms (they are the most often composed of two or more words called multi word units) it is not a simple process.

Members of Language Resources and Technologies Society developed semiautomatic approach for term recognition, extraction and lemmatization. Picture 1 illustrates steps in terminology extraction. Crucial resources are morphological dictionaries and grammars. They are combined with some statistical measures for term extraction. The first step is analysis of terms in existing term base mentioned before (RudOnto, GeolISSTerm). It was recognized 14 most productive patterns that represent structure of MWU terms. They are represented in form of transducers applied on domain corpus to extract terminology. Examples of patterns are presented in [15]. After applying these transducers on domain text extracted potential terms were evaluated.

Results presented in previous paper were satisfying enough to speed up the development of a terminological dictionary.

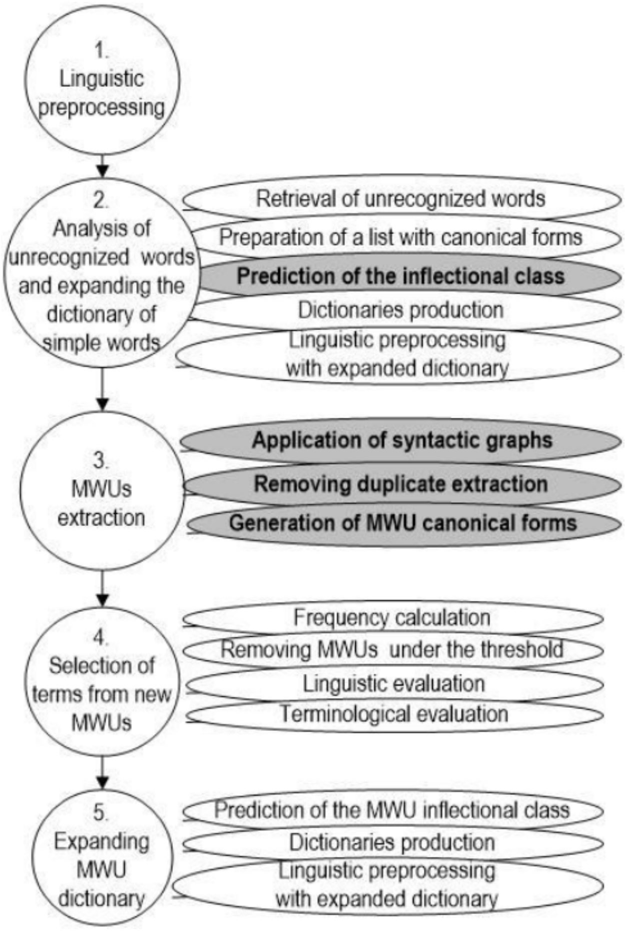
1. Linguistic preprocessing
2. Analysis of unrecognized words and expanding the dictionary of simple words
Retrieval of unrecognized words
Preparation of a list with canonical forms
Prediction of the inflectional class
Dictionaries production
Linguistic preprocessing with expanded dictionary
3. MWUs extraction
Application of syntactic graphs
Removing duplicate extraction
Generation of MWU canonical forms
4. Selection of terms from new MWUs
Frequency calculation
Removing MWUs under the threshold
Linguistic evaluation
Terminological evaluation
5. Expanding MWU dictionary
Prediction of the MWU inflectional class
Dictionaries production
Linguistic preprocessing with expanded dictionary

Image 2: Diagram of terminology extraction [15]

6. TERMI – AN APPLICATION FOR TERMINOLOGY MANAGEMENT

Termi application is developed at the University of Belgrade Faculty of Mining and Geology, with the support of BAEKTEL project. It is available at the following address: http://termi.rgf.bg.ac.rs/. It provides terminology management, regardless of term domain.

The application consists of three basic web pages which manage terminology: browse, search and update. Additionally, there are pages which manage profiles, bibliography and a login page.

Each term comes with the name, definition, synonyms, abbreviations and a bibliographic source. Each term, except the top term in dictionary tree, has a hyperonym term, while each term can have an arbitrary number of hyponym terms.

Term name is also a link that leads to a page that presents a complete overview of the term with information about it (translations, descriptions, synonyms, acronyms, hypernym concept, hyponym concepts, bibliography).

Important preference for OER-s, is the possibility to embed link to specific term. The result is tooltip with a definition and traslation of term with link to the term in Termi.

5. CONCLUSION

Lexical and terminological resources offer priceless aid for better understanding of the available OER contents. Presented resources are also helpful in a sense of appropriate translation option. Successful methods used in automatic term extraction can be applied to units that belong to the general lexica, as well. The potential expansion of such resources would inevitably lead to a more fruitful information retrieval and extraction, providing an invaluable education resource, applicable in all of its domains. In the further work bilingual terminology extraction will be considered.

REFERENCES

[1] I. Gurevych, D. Bernhard and A. Burchardt, “Educational Natural Language Processing,” Notes for ENLP tutorial held at AIED 2009 in Brighton, Jul, 2009.

[2] J. M. Balkin and J. Sonnevend, “Digital Transformation of Education (April 4, 2016),” in Education and Social Media: Toward a Digital Future, 2016, Forthcoming, C. Greenhow, J. Sonnevend and C. Agur, Ed. Cambridge, MA: MIT Press, 2016, pp. 22.

[3] D. Litman, “Natural language processing for enhancing teaching and learning,” in Proc. Natural language processing for enhancing teaching and learning, 2016, pp. 4170–4176.

[4] T. M. Cabré Castellví, Terminology: Theory, Methods, and Applications. Amsterdam: John Benjamins, 1999, pp. 115.

[5] I. Obradović, R. Stanković, J. Prodanović and O. Kitanović, “A TEL platform blending academic and entrepreneurial knowledge,” in Proc. 4th Conference on e-Learning, 2013, pp. 65–70.

[6] R. Stanković, D. Carlucci, O. Kitanović, N. Vulović and B. Zlatić, “LRMI markup of OER content within the BAEKTEL project,” in Proc. 6th International Conference on e-Learning (eLearning-2015), 2015, pp. 98–103.

[7] C. Krstev, Processing of Serbian. Belgrade, Serbia: Faculty of Philology, 2008, pp. 40–47.

[8] R. Stanković, I. Obradović, O. Kitanović and Lj. Kolonja, “Building terminological resources in an e-learning envinronment,” in Proc. 3rd International Conference on e-Learning (eLearning-2012), 2012, pp. 114–119.

[9] S. Vujičić Stanković, C. Krstev and D. Vitas, “Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain,” in Proc. 7th Global WordNet Conference, 2014, pp. 127–132.

[10] M. Hepp, “Ontologies: State of the art, business potential, and grand challenges,” in Ontology Management, 1st ed., M. Hepp, P. DeLeenheer, A. De Moor and Y.Sure, Ed. Springer US, 2008, pp. 3-22.

[11] M. Uschhold and M. Gruninger, “Ontologies: Principles, methods and applications,” Knowledge engineering review, vol. 11, No. 2, pp. 93–136, June. 1996.

[12] P. Pantel and L. Dekang, “A statistical corpus-based term extractor,” in Advances in Artificial Intelligence, 1st ed., E. Stroulia and S. Matwin, Ed. Springer Berlin Heidelberg, 2001, pp. 36-46.

[13] K. Heylen and D. De Hertog “Automatic Term Extraction,” in Handbook of Terminology, 2nd ed., vol. 1, H. J. Kockaert and F. Steurs, Ed. Amsterdam: John Benjamins, 1964, pp. 203–221.

[14] A. Oliver and M. Vàzquez, “TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction,” in Proc. Recent Advances in Natural Language Processing, 2015, pp. 473–479.

[15] C. Krstev, R. Stanković, I. Obradović and B. Lazić, “Terminology Acquisition and Description Using Lexical Resources and Local Grammars,” in Proc. 11th Conference on Terminology and Artificial Intelligence, 2015, pp. 81– 89.