What can NLP techniques do for eLearning?

Paola Monachesi*, **Eline Westerhout** *

* Utrecht University, Uil-OTS
Trans 10, 3512 JK Utrecht, The Netherlands
{Paola.Monachesi, Eline.Westerhout}@let.uu.nl

Abstract

The aim of the Language Technology for eLearning project is to show is to show that current results achieved in the area of Natural Language Processing and the Semantic Web, (i.e. ontologies) can play a relevant role in improving the functionality of existing Learning Management Systems (LMS). In this paper, we discuss how current NLP techniques have been employed for the development of a keyword extractor and of a glossary candidate detector by focussing on the results obtained for Dutch. We also briefly discuss the role that ontologies play in the project in the multilingual retrieval of the learning objects.

1. Introduction

The aim of the Language Technology for eLearning (LT4eL) project[1] is to show that current results achieved in the area of Natural Language Processing and the Semantic Web, (i.e. ontologies) can play a relevant role in improving the functionality of existing Learning Management Systems (LMS). LMSs allow tutors to set-up courses, to manage collections of learning materials and monitor students' progress, whilst providing students with a structured way to access data. However, given the huge amount of static and dynamic learning content created for eLearning tasks, it becomes necessary to enhance the management, distribution and retrieval of the learning material within Learning Management Systems and we claim that NLP techniques can be employed to this end.

More specifically, in the LT4eL project, we provide content creators with a keyword extractor which allows for semi-automatic metadata annotation of the learning objects. Keyword extraction has been widely explored in the NLP and IR community. In our approach, we adapt those results to the eLearning context by using statistical measures in combination with linguistic processing to detect salient words which are good keyword candidates.

In addition, we have developed a glossary candidate detector which allows for the creation of glossaries to be linked to learning objects. The glossaries are based on the definitions of the relevant terms which are attested in the learning objects. Definition extraction is the topic of much current research and techniques have been developed to this end within the Natural Language Processing and the Information Extraction communities mainly based on grammars that detect the relevant patterns and machine learning methods: in the LT4eL project, we adapt these techniques for eLearning purposes.

Standard retrieval systems tend to offer keyword-based searching, matching words present only in the query term. Yet most techniques do not take into account systematic relationships between concepts denoted in the query term and other related concepts which might be relevant for the user. In the LT4eL project, ontologies are instrumental in expressing such relations since they can be employed to query and to navigate through the learning material improving thus the learning process. Furthermore, by linking the ontology to language specific lexica, multilingual retrieval becomes possible. In particular, an ontology in the domain of *computing* has been developed based on the keywords extracted from the learning material.

In the LT4eL project, we have developed the various functionalities in the eight languages represented by our consortium that is Bulgarian, Czech, Dutch, English, German, Polish, Portuguese and Romanian. However, in this paper we will focus on the functionalities and the results obtained for Dutch.

[1] http://www.lt4el.eu/

2. Related work

The LT4eL project is quite unique in demonstrating the added value of integrating NLP techniques and results to enhance eLearning, it is thus difficult to compare our approach to others. It should be noticed, however, that the techniques employed in our project, have been widely explored in the natural language processing and information retrieval communities and in the LT4eL project, we take advantage of the results achieved in these areas and adapt them to the eLearning context.

Keyword extraction has been considered in combination with summarization ([24], [16], [30]). An additional use is to identify automatically relevant terms that can be employed in the construction of domain-specific dictionaries or more recently of domain ontologies ([23]). However, keyword and keyphrase extractors have been provided mainly for English, cf. [23], [6], [24],[30], [28], [22], [16], [7]. One innovative aspect of our project is that we provide this functionality for all the eight languages represented in our project, that is Bulgarian, Czech, Dutch, English, German, Polish, Portuguese and Romanian and we embed this significant result within the eLearning context. Another innovative feature is that keyphrases are extracted in addition to keywords. This responds to findings that users frequently use keyphrases to describe a document, cf. [8].

Research on the detection of definitions has been pursued in the context of automatic building of dictionaries from text, question-answering and recently also within ontology learning. In the area of automatic glossary creation, the DEFINDER system [19] combines shallow natural language processing with deep grammatical analysis to identify and extract definitions and the terms they define from on-line consumer health literature. Their approach relies entirely on manually crafted patterns. Research on definition extraction has been pursued very actively also in the area of Question-Answering. The answers to 'What is'-questions are usually definitions of concepts. A common approach in this area is to search the corpus for sentences consisting of a subject, a copular verb and a predicative phrase. If the concept matches the subject, the predicative phrase is returned as answer, also in this case research relied initially almost totally on pattern identification and extraction and only later, machine learning techniques have been employed. In [21], both the analysis of document structure as well as dependency parsing are explored while [5] propose an approach to definition extraction which operates on fully parsed text and machine learning techniques (cf. also [1], [15] for the use of machine learning methods in definition extraction). Research on definition extraction has been carried out also in the area of ontology learning. For example, within the German HyTex project [20], 19 verbs that typically appear in definitions were distinguished and search patterns have been specified based on the valency frames of these definitor verbs in order to extract definitions. Another work in this direction is that of [25] that proposes a rule-based method for extracting and analyzing definitions from parsed text on the basis of a semantically oriented parsing system. The results are then employed to improve the quality of text-based ontology learning. Also this approach relies on pattern extraction techniques to detect definitions and doesn't employ machine learning. A difference with respect to previous systems is its use of semantic information in the identification of patterns. In the LT4eL project, we adapt these approaches to the eLearning context in order to develop dictionaries based on definitions extracted from the learning objects.

The LT4eL project is also quite innovative in adopting ontologies for (multilingual) retrieval and reuse of learning objects. However, it should be noticed that some approaches in this direction are emerging as attested by the work of [9] and [31].

3. NLP for metadata generation: the keyword extractor

The task of a keyword extractor is to automatically identify a set of terms in a document that best describes its content. In our project, extracted keywords are employed for metadata generation, we have thus privileged a simple approach which is based on a frequency criterion to select the relevant keywords in a document which has been complemented with a linguistic processing step. An advantage of this approach is that it can be easily adapted to several languages.

In the LT4eL project, we use the same algorithm for all the languages under consideration while we encode the language specific differences in the language model. The keyword extractor accepts linguistically annotated input and outputs a list of suggested keywords to be included in the Learning Object Metadata (IEEE LOM). [2]

More specifically, the input for the keyword extractor is constituted by learning objects of various formats, e.g. PDF and DOC which are converted into HTML to retain layout information. A part-of-speech tagger, a lemmatizer and a morpho-syntactic analyser are used to provide documents with the necessary additional linguistic information which has been used in the extraction of keywords. A linguistically annotated document in an XML format which is derived from the XCESAna standard for linguistically annotated corpora is produced.

This linguistic information, which is extracted from the corpus of learning objects, is added to the language model for the specific language which consists of three parts: lexical units, word form types and documents. The list of candidate keywords is ranked by their saliency and to determine it an approach based on frequency has been adopted. In our project, we have implemented and evaluated the appropriateness of three statistical measures, that is TFIDF, Residual IDF and Adjusted RIDF, in ranking the most relevant keywords from our multilingual learning objects [3] and [4]. These statistical measures were tested on the various languages and results show that, for Dutch, TFIDF is the method that produces the best results. Similar results were obtained for the other languages.

In the development of the keyword extractor, special attention has been devoted to multiword terms. A first analysis of the manually selected keywords revealed, that for some languages a substantial amount of them is multi word. We noticed that results improved for all languages if multiword keywords up to a length of 3 words were included. Additionally, we used linguistic information to further restrict this set of multi-word key phrases, e.g. to exclude phrases which end in a preposition. We refer to [11] and [12] for more details on the keyword extractor and its ap-

[2]http://ltsc.ieee.org/doc/wg12/LOM3.6.html

plications.

3.1. Quantitative evaluation

It is not easy to establish an appropriate methodology to evaluate the keyword extractor given the eLearning application in which it is embedded. Term extraction for ontology building or for lexicon creation is usually evaluated in terms of precision which is measured by dividing the extracted terms which are appropriate for a given domain by the number of accepted terms. This approach doesn't seem acceptable in the case of our application in which appropriate keywords will be much more restricted in number than appropriate terms for a given domain. In addition, the choice of keywords for a given document is often determined by the context of its use and we thus expect there to be variation among annotators in determining which keywords are appropriate for a given document. Regardless of these limitations, two experiments have been carried out to evaluate our tool.

The first experiment has been designed to established how much variation there is among users in the assignment of keywords to a text, an evaluation of inter-anntotator agreement on the keyword selection task has been performed[2]. We have noticed that for all languages tested the keyword extractor is in better agreement with the human annotators than the human annotators among themselves. In the case of Dutch, the average human annotator agreement is 0.67 while the the keyword extractor agreement with human annotator is 0.72.

The second experiment has been developed in order to assess the appropriateness of the keyword extractor in the context of the semiautomatic metadata annotation of the learning objects. We have presented test persons with a document and a limited set set of keywords which have been extracted from this document. Each member of this set of keywords is assessed by the test person with respect to the adequacy to represent the text. The conclusion we have reached is that results are around 2.0 (good) with some variance between languages and presenting the best 10 keywords to the users would yield the highest acceptance for most languages. In particular, for Dutch, the average score for the first 20 keywords is 1.93 while for the first 10 is 1.68.

4. NLP and ML for glossary creation

Definition extraction is performed within different contexts, such as question answering, dictionary building and ontology learning. Within the LT4eL project, definitions are created in order to create glossaries semi-automatically. The challenges we have to tackle are related to the corpus we collected. First, our corpus contains a large variety of texts, which are characterized in the following way:

- different topics (e.g. MS Word, eLearning, Internet);
- different lengths (4 - 60 pages)
- written for diverse purposes (e.g. slides, papers, tutorials).

Whereas for short texts it is most important that we have a good recall, for the longer texts it is also very important that the precision is not too low. Therefore, it is crucial to find a good balance between precision and recall.

In our project, we address four types of definitions. The first category (i.e. *is_def*) are the definitory contexts in which a form of the verb *zijn* ('to be') is used as connector verb. These are the most straightforward definitions. The second group (i.e. *verb_def*) is formed by the definitory contexts in which other verbs are used as connector (e.g. *betekenen* ('to mean'), *wordt ... genoemd* ('is called'), *wordt gebruikt om* ('is used to')). Together with the first group, the second group comprises over 50 % of our definitions. The third type (i.e. *punct_def*) are the definitory contexts having specific punctuation features (e.g. *:*, *(..)*). The last category (i.e. *pron_def*) contains the definitory contexts in which relative and demonstrative pronouns (e.g. *dit* ('this'), *dat* ('that'), *deze* ('these')) and words like *hiermee* ('with this'), *hierdoor* ('because of this') are used to point back to a defined term that is mentioned in a preceding sentence. The definition of the term then follows after the pronoun, so these are often multisentence definitory contexts. Table 1 shows for each of the categories an example definition.

In our approach, we have combined NLP techniques and Machine Learning to extract definitions from our learning objects. In the first step, we wrote a grammar based on 300 manually selected definitions. Thereafter, we used machine learning to filter incorrect results in order to improve the precision.

4.1. Extraction of definitions

The creation of the grammar has been done on the basis of the patterns found in 21 files. The grammar was used for the extraction of definitions from the complete corpus. The precision (P), recall (R) and F-measure (F) obtained with the grammar were calculated taking the sentence as a unit. The following formulae were employed for calculating precision, recall and F-score:

$$\text{Precision} = \frac{\text{sentences marked manually and automatically}}{\text{sentences marked manually}}$$

$$\text{Recall} = \frac{\text{sentences marked manually and automatically}}{\text{sentences marked automatically}}$$

$$\text{F} - \text{score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

The results obtained with the grammar are shown in table 2. The manual annotation of definitions was used as gold standard against which precision and recall were measured.

We refer to [26] for more details on the performance of the grammar.

Type	Example sentence
is_def	Gnuplot is een programma om grafieken te maken *'Gnuplot is a program for drawing graphs'*
verb_def	E-learning omvat hulpmiddelen en toepassingen die via het internet beschikbaar zijn en creatieve mogelijkheden bieden om de leerervaring te verbeteren . *'eLearning comprises resources and application that are available via the internet and provide creative possibilities to improve the learning experience'*
punct_def	Passen: plastic kaarten voorzien van een magnetische strip, die door een gleuf gehaald worden, waardoor de gebruiker zich kan identificeren en toegang krijgt tot bepaalde faciliteiten. *'Passes: plastic cards equipped with a magnetic strip, that can be swiped through a card reader, by means of which the identity of the user can be verified and the user gets access to certain facilities. '*
pron_def	Dedicated readers. Dit zijn speciale apparaten, ontwikkeld met het exclusieve doel e-boeken te kunnen lezen. *'Dedicated readers. These are special devices, developed with the exclusive goal to make it possible to read e-books.'*

Table 1: Examples for each of the definition types

type	P	R	F
is_def	28.10	86.52	42.41
verb_def	44.64	75.76	56.18
punct_def	9.91	68.18	17.31
pron_def	9.18	41.30	15.02

Table 2: Performance of the grammar

4.2. The use of machine learning for filtering

As can be seen in table 2, the recall obtained with the grammar is acceptable (41.30 - 86.52). However, the precision of the results obtained with the grammar is low (9.18 - 44.64), especially for the punctuation and the pronoun definitions. In order to increase precision, we trained a Naive Bayes classifier using 10-fold cross validation and applied it on the results obtained with the grammar. We did this for two of the definition types, namely the *to be*-definitions and punctuation definitions. Weka, a collection of machine learning algorithms for data mining tasks, was used to perform the experiments [29].

The 10 attribute settings (table 3) we distinguish are based on the configurations used by [5]. The attributes focus on:

1. **Text properties**: bag-of-words, bigrams, and bigram preceding the definition.

2. **Syntactic properties**: type of determiner within the defined term (definite, indefinite, no determiner).

3. **Proper nouns**: presence of a proper noun in the defined term.

These 10 attribute settings were tested for the two definition types that occurred most frequently in our corpus, that is, the *to be*-patterns and the punctuation patterns. There were 274 *to be*-patterns extracted, of which 77 are real definitions (precision of 28.1 %). For the punctuation patterns, the precision was even lower. This set includes 454 sentences, of which 45 are correct definitions (precision is 9.9 %).

We refer to [27] for more details on the machine learning approach.

setting	description
1	using bag-of-words
2	using bigrams
3	combining bag-of-words and bigrams
4	setting 3 + bigram preceding definition
5	setting 3 + definiteness of article in marked term
6	setting 3 + presence of proper noun
7	setting 3 + bigram preceding definition + def. of article in marked term
8	setting 3 + bigram preceding definition + presence of proper noun
9	setting 3 + def. of article in marked term + presence of proper noun
10	using all attributes

Table 3: Configurations used for the Machine Learning experiment

4.3. Quantitative evaluation

For the *to be*-patterns, the results are best when all attributes are used (setting 10). For the punctuation patterns, precision is highest when configuration 7 is used for training the classifier whereas recall and F-score are best for setting 9.

Table 4 shows the final results we obtained after using the grammar for the extraction of patterns and filtering as most as possible incorrect sentences using the Naive Bayes classifier.

	to be-definitions			punctuation definitions		
setting	P	R	F	P	R	F
1	69.44	56.18	62.11	43.24	24.24	31.07
2	66.25	59.55	62.72	31.71	19.70	24.30
3	76.62	66.29	71.08	45.10	34.84	39.32
4	76.62	66.29	71.08	46.81	33.33	38.94
5	77.63	66.29	71.52	45.28	36.36	40.34
6	76.62	66.29	71.08	50.00	36.36	42.11
7	78.67	66.29	71.95	51.06	36.36	42.48
8	76.32	65.16	70.30	50.00	36.36	42.11
9	78.94	67.42	72.73	50.00	39.39	44.07
10	80.00	67.42	73.17	50.00	36.36	42.11

Table 4: Final results

5. Ontology for multilingual search

As already mentioned, the aim of the LT4eL project is to improve the retrieval and the usability of (multilingual) learning material within a Learning Management System. In order to achieve this objective, an ontology-based search functionality has been developed which is based on the following components:

- collection of (multilingual) learning objects;
- a (language independent) ontology that includes an upper ontology with about 50 concepts (i.e. DOLCE [14]), a domain ontology in the area of computing based on the keywords extracted from the learning objects which consists of about 750 domain concepts as well as appropriate definitions;
- a lexicon for each of the languages addressed which comprises words or phrases that are mapped to concepts attested in the ontology;
- a collection of learning objects annotated on the basis of the concepts attested in the domain ontology. To this end, annotation grammars have been developed while word sense disambiguation has been carried out manually.

The ontology-based search engine we have build is based on these various components and on the basis of several parameters including the search terms employed and the language(s) for which user wants to see available documents, the search engine returns a ranked list of documents which semantically match the search word(s), identified by their titles. We refer to [13] and [18] for more details in the use of the ontology in the LT4eL project.

6. Qualitative evaluation of the functionalities

The functionalities we have developed have been integrated in the ILIAS Learning Management System and are being offered as web services allowing thus for their usability in other open source LMS.
In this paper, we have discussed a quantitative evaluation of the performance of the functionalities we have developed, in particular the keyword extractor and the glossary candidate detector. This evaluation has been crucial to verify that the tools we have developed produce state of the art results. However, we believe that the best way to evaluate these functionalities is in the context of their use within ILIAS.
Given the eLearning context in which we operate, it might be thus more relevant to evaluate to which extent the ILIAS system improved with the developed functionalities help users in the learning tasks. These are both the content providers/tutors who will exploit the functionalities in order to compile a course for a specific target group and who want to draw on existing texts, media etc. and learners who are looking for contents which suit their current needs, e.g for self-guided learning.
A suitable validation methodology has been developed which is based on User Scenarios, which focus on the role of teachers and learners. User Scenarios are defined as *a story focused on a user or group of users, which provides information on the nature of the users, the goals they wish to achieve and the context in which the activities will take place*.
The validation we have carried out for Dutch on the basis of user scenarios reveals that tutors appreciate the keyword extractor and they would use it to assign metadata to learning objects. However, there is no equal satisfaction with respect to the quality of the proposed keywords but tutors agree that it is easy to add new keywords manually to improve on results. The issue is whether it is really possible to improve the results given the subjectivity of the task and the relevance of the context in finding the appropriate keywords. Also in the case of the Glossary Candidate Detector, tutors agree that it is a useful tool but there is variation with respect to the quality of the results. Students are more positive than tutors on the use of the keyword extractor and they generally prefer a search based on keywords than one based on full text given that the former one gives more focussed results. As for the Glossary Candidate Detector, students find that having a glossary related to a learning object can be very useful, especially when preparing for a test.
As for having the possibility to search documents across languages with the help of the ontology, the majority of the students tested agrees that the ontology browser is useful, easy to use and it provides a useful visualization of the relation among concepts. In addition, all students agree

that the semantic search gives better results than full text search. They also all agree that it is useful to have different search methods which can be used, that is full text search, keyword search and semantic search by means of ontology browsing. Some students have suggested interesting ways in which the search methods could be combined, that is semantic search for a first orientation and to identify relevant relations while full text search can be useful to get additional details or to get deeper into the subject matter. All students appreciate the possibility of searching documents in various languages, they especially appreciate the fact that they can type queries in their own language and get documents in a different language. Similar results were attested for tutors.

7. Conclusion

The main objective of the LT4eL project is to show that NLP techniques and ontologies, which are a crucial component of the Semantic Web vision, can play a relevant role in enhancing the learning process. In particular, they can improve on the search and retrieval of (multilingual) learning material. The user centered evaluation we have carried out shows that there is variation on the quality of the results of the functionalities developed, i.e. the keyword extractor, the glossary candidate detector and the multilingual semantic search based on ontology browsing. However, the majority of users agree that the functionalities developed are useful and they would employ them if they would be available showing thus that NLP techniques are ripe for applications in the eLearning context.

References

[1] Blair-Goldensohn, B. and K. R. McKeown, and A. Hazen Schlaikjer. New Directions In Question Answering, chapter Answering Definitional Questions: A Hybrid Approach. AAAI Press, 2004.

[2] Bruce R. and J. Wiebe. 1999. *Recognizing subjectivity:A case study of manual tagging*. In: Natural Language Engineering. Vol. 5,No 2, pp 187–205.

[3] Church, K. and W. Kenneth and W. Gale. 1995. *Inverse Document Frequency (IDF): A Measure of Deviations from Poisson*. In: Proc. of Third Workshop on Very Large Corpora.

[4] K. Church and W. Gale. 1995. *Poisson mixtures*. In: Natural Language Engineering. Vol. 1,No 2, pp 163–190.

[5] Fahmi, I. and Bouma, G.(2006), *Learning to identify denitions using syntactic features*, in R. Basili and A. Moschitti (eds), Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications.

[6] Frank, E. and Paynter, G. and Witten, I. and Gutwin, C. and Nevill-Manning, C. 1999. *Domain-specific Keyphrase Extraction.* In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence 1999, pp. 668-673

[7] Hulth, A. 2003. *Improved automatic keyword extraction given more linguistic knowledge*. In: Proceedings of EMNLP2003.

[8] Jones S. and G. W. Paynter. 2006. *An Evaluation of Document Keyphrase Sets.* In: Journal of Digital Information, Vol.4, No1.

[9] Jovanovic, J. and D. Gasevic and C. Brooks and V. Devedzic and M. Hatala (2007) *LOCO-Analyst: A Tool for Raising Teachers Awareness in Online Learning Environments*. In:Proceedings of EC-TEL 2007, Springer LNCS.

[10] Katz, S.M. 1996. *Distribution of content words and phrases in text and language modelling*. In: Natural Language Engineering 2(1996)1. pp. 1559.

[11] Lemnitzer L. and C. Vertan and A. Killing and K. Simov and D. Evans and D. Cristea and P. Monachesi. 2007. *Improving the search for learning objects with keywords and ontologies*. In: Proceedings of the ECTEL 2007 conference. Springer Verlag.

[12] Lemnitzer L. and Monachesi P. 2008 *Extraction and evaluation of keywords from Learning Objects a multilingual approach*. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2008).

[13] Lemnitzer L. and Simov K. and Osenova P. and Mossel E. and Monachesi P. (2007) *Using a domain ontology and semantic search in an eLearning environment*. In: Proceedings of The Third International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering. (CISSE 2007). Springer-Verlag. Berlin Heidelberg.

[14] Masolo C. and S. Borgo and A. Gangemi and N. Guarino and A. Oltramari and L. Schneider. 2002. *The WonderWeb Library of Foundational Ontologies*. WonderWeb Deliverable D17, August 2002.

[15] Miliaraki S. and I. Androutsopoulos. *Learning to identify single-snippet answers to definition questions*. In: Proceedings of COLING 2004, pages 13601366, 2004.

[16] Mihalcea R. and P. Tarau. 2004. *TextRank: Bringing Order into Texts*, In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, July 2004.

[17] Monachesi, P. and L. Lemnitzer and K. Simov. *Language Technology for eLearning*. Proceedings of EC-TEL 2006, in Innovative Approaches for Learning and Knowledge Sharing, LNCS 0302-9743, pp. 667-672. 2006

[18] Monachesi, P. and K. Simov and E. Mossel and P. Osenova and L. Lemnitzer. *What ontologies can do for eLeanring*. Proceedings of IMCL 2008 conference. 2008

[19] S. Muresan and J. Klavans. *A method for auto- matically building and evaluating dictionary resources*. In Proceedings of the Language Resources and Evaluation Conference (LREC 2002), 2002.

[20] A. Storrer and S. Wellinghof. *Automated detec- tion and annotation of term definitions in German text corpora*. In: Proceedings of LREC 2006, 2006.

[21] Tjong Kim Sang, E. and G. Bouma, and M. de Rijke. *Developing offline strategies for answering medical questions*. In D. Molla and J. L. Vicedo, editors, Proceedings AAAI 2005 Workshop on Question Answering in Restricted Domains, 2005.

[22] Turney, P. D. 2000. *Learning algorithms for keyphrase extraction*. Information Retrieval, 2:303-336.

[23] Sclano, F. and P. Velardi, TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. To appear in Proc. of the 3rd International Conference on Interoperability for Enterprise Software and Applications I-ESA 2007, Funchal, Madeira Island, Portugal, March 28-30th, 2007.

[24] Wan X. and J. Yang and J. Xiao. 2007. *Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction.* In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, 2007, pp. 552–559.

[25] Walter S. and M. Pinkal. *Automatic extraction of definitions from German court decisions*. In: Proceedings of the workshop on information extraction beyond the document, pages 2028, 2006.

[26] Westerhout E. and P.Monachesi. *Extraction of Dutch definitory contexts for elearning purposes*. In: Proceedings of CLIN 2006, 2007.

[27] Westerhout E. and P.Monachesi. *Combining pattern-based and machine learning methods to detect definitions for eLearning purposes*. In: Proceedings of LREC 2008, 2008.

[28] Witten,I.H. and G. W. Paynter and E. Frank and C. Gutwin, and C. G. Nevill-Manning. 1999. *KEA: Practical automatic keyphrase extraction*. In: Proceedings of Digital Libraries 99 (DL'99), pp. 254-256.

[29] Witten I. and E. Frank. *Data mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufman Publishers, 2005.

[30] Zha, H. Y. 2002. *Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering*. In: Proceedings of SIGIR2002, pp. 113-120.

[31] Zoaq, A. and R. Nkambou and C. Frasson (2007) *Building Domain Ontologies from Text for Educational Purposes*. In: Proceedings of EC-TEL 2007, Springer LNCS.