Extraction and evaluation of keywords from Learning Objects – a multilingual approach

Lothar Lemnitzer and Paola Monachesi

Seminar für Sprachwissenschaft,
Universität Tübingen,
Germany,
lothar@sfs.uni-tuebingen.de
University of Utrecht
The Netherlands
paola.monachesi@let.uu.nl

Abstract

We report about a project which brings together Natural Language Processing and eLearning. One of the functionalities developed within this project is the possibility to annotate learning objects semi-automatically with keywords. To this end, a keyword extractor has been created which is able to handle documents in 8 languages. The approach employed is based on a linguistic processing step which is followed by a filtering step of candidate keywords and their subsequent ranking based on frequency criteria. Three tests have been carried out to provide a rough evaluation of the performance of the tool, to measure inter annotator agreement in order to determine the complexity of the task and to evaluate the acceptance of the proposed keywords by users.

1. Introduction

eLearning aims at replacing the traditional learning style in which content, time and place are predetermined with a more flexible, customized process of learning. While in traditional learning, the instructor plays an intermediate role between the learner and the learning material, this is not always the case within eLearning since learners are in a position to combine learning material and to create their own courses. However, a necessary condition is that content should be easy to find and metadata plays a crucial role to this end. It provides a common set of tags that can be viewed as data describing data. Metadata tagging enables organizations to describe, index, and search their resources and this is essential for reusing them.

In the eLearning community, various metadata standards have emerged to describe eLearning resources. The *Learning Object Metadata* standard launched by the IEEE[1] is the most widespread standard used for learning objects. Among other information, keywords can be provided as part of the LOM metadata set. Providing metadata, however, is a tedious activity and it is not widely accepted by content providers and authors as part of their work. This has, however, the highly undesirable consequence that content becomes less visible and more difficult to retrieve.

One of the goals of the LT4eL project[2], cf. (Monachesi et al., 2006), is to show that language technology can provide significant support for this task. The solution we offer is to provide a Language Technology based functionality, that is a keyword extractor which allows for semi-automatic metadata annotation of the learning objects within a Learning Management System (LMS). Keyword extraction is the process of extracting a few salient words or phrases from a given document and using these words as a surrogate of this document. Keyword extraction has been widely explored in the natural language processing and information retrieval communities and in our project we take advantage of the techniques and the results achieved in these areas and adapt them to the eLearning context. More specifically, our approach employs statistical measures in combination with linguistic processing to detect salient words which are good keyword candidates.

It should be noticed, however, that keyword and keyphrase extractors have been provided mainly for English, cf. (Sclano and Velardi, 2007), (Frank et al., 1999), (Wan et al., 2007),(Zha, 2002), (Witten et al., 1999), (Turney, 2000), (Mihalcea and Tarau, 2004), (Hulth, 2003). 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. This makes it necessary to address language specific aspect. The linguistic annotation for each language differs, and so do the parts of speech. With a language dependent lingistic model we intend to exclude all words which are assigned a part of speech that is not relevant for keywords (e.g. adverb, conjunction). Second, and we embed this tool in an eLearning context. Another salient feature is that keyphrases are extracted in addition to keywords. This responds to findings that users frequently use keyphrases to describe a document, cf. (Jones and Paynter, 2006).

More generally, the main objective of the LT4eL project is to show that the integration of Language Technology based functionalities and Semantic Web techniques will enhance the management, distribution and retrieval of the learning material within Learning Management Systems.

The rest of the paper is organized as follows. In section 2., we outline the architecture of the keyword extractor, in-

---

[1]cf. `http://ltsc.ieee.org/doc/wg12/LOM3.6.html`.

[2]The project is funded by the EC under the IST programme. We are grateful for the support. For further details, visit `www.lt4el.eu`.

cluding the methods we are using for ranking keywords and we point out the innovative features of our tool. The quantitative evaluation of the tool is discussed in section 3. and results obtained are analyzed. The keyword extractor has been integrated into the learning management system IL-IAS. We show the result in section 4.. Finally, section 5. contains our conclusions and plans about future work.

2. The Keyword Extractor

The task of a keyword extractor is to automatically identify a set of terms in a document that best describes it. Keyword extractors have been employed to identify appropriate entries for building an automatic index for a document collection and have been used to classify texts. Keyword extraction has also been considered in combination with summarization ((Wan et al., 2007), (Mihalcea and Tarau, 2004), (Zha, 2002)). 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 ((Sclano and Velardi, 2007)) .

In the LT4eL project, we have adapted current techniques for term extraction in order to develop a keyword extractor which is employed for the semi-automatic metadata annotation of the learning objects. We have 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.

This method was found to lead to poor results, as claimed in (Mihalcea and Tarau, 2004) and consequently alternative methods were explored in the literature. They are mainly based on supervised learning methods, where a system is trained to recognize keywords in a text, based on lexical and syntactic features. However, given the specific application which has been envisaged in our project, that is the extraction of relevant keywords for metadata generation, we have privileged an approach which could 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. It should be noticed that a machine learning approach didn't seem a possible option: given the small corpus of learning objects available for each language, we wouldn't have had enough training data at our disposal.

The keyword extractor accepts linguistically annotated input and outputs a list of suggested keywords, as can be seen in figure 1. 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. From this intermediary representation an XML format is generated which preserves basic layout features of the original texts. Linguistic information is added to this format. The linguistic processing chains which which yield the linguistica annotated docuements are slightly different per language. However, they all provide the same types of information: part of speech, base form and morphsyntactiv features for each word. The process yields a linguistically annotated document in an XML format which is derived from the XCESAna standard for linguistically annotated corpora and which is the same for all languages[3]. The linguistic an-

notation comprises: a) the base form of each word; b) the part of speech of this base form; c) further morphosyntactic features of the word form which is used in the text.

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**: they represent the combination of a lemma and a part of speech tag. They are the basic units on which statistics are calculated and they are returned as keyword candidates.

- **Word Form Types**: they represent the actual form of the lexical unit in the input file in combination with their morphological information. Only those forms that can occur as possible keywords are retained – mainly nouns, proper nouns and unknown words.

- **Documents**: they represent the documents which constitute the corpus including their names and domains.

Potentially interesting sequences of words are extracted using the suffix array data structure (Yamamoto and Church, 2001) but a condition is that they must appear at least twice in the document. Afterwards, filtering occurs on the basis of language specific information and sequences longer than a certain threshold are discarded. In general, sequences comprising up to 3 words are retained.

The list of candidate keywords is ranked by their saliency and to determine it an approach based on frequency has been adopted.

Keywords are those terms that best identify the text and represent what the text is about (i.e. the topics of a text). They tend to occur more often in that text than could be expected if all words were distributed randomly over a corpus.

A well-established way to measure the distribution of terms over a collection of documents is TFIDF, cf. equation 1.

$$ TFIDF \quad \text{where} \quad IDF = \log_2 \frac{N}{df} \qquad (1) $$

Church argued that Poisson distributions or mixtures of Poisson distributions of words in texts are quite useful statistics (cf. (Church and Gale, 1995) and equation 2).

$$ \pi(k; \theta) = \frac{e^{-\theta} \theta^k}{k!} \qquad (2) $$

While the distribution of e.g. function words like *of, the, it* is close to the expected distribution under the Poisson distribution model, good keyword candidates deviate significantly from the expectation. The score of this deviation can be used as a statistics by which the lexical units are ranked (Church and Gale, 1995a). The deviation of the observed distribution of a word from the expected distribution under the Poission model, i.e predicted IDF (cf. equation 3) is called Residual RIDF (short: RIDF, cf. equation 4).

$$ -log_2(1 - e^{-\theta}) \quad \text{where} \quad \theta = \frac{cf}{N} \qquad (3) $$

$$ (IDF - PredictedIDF) \qquad (4) $$

---

[3]The document grammar can be provided on demand, please contact us if you are interested in further details.
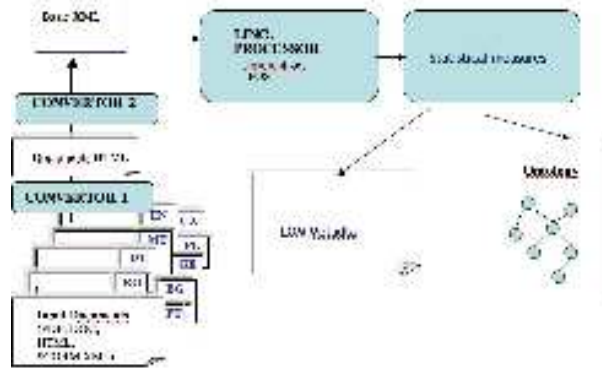
Figure 1: Architecture of the keyword extractor

During our experiments with these metrics we recognized that RIDF does not take the term frequency in the analysed document into account. Since this is the most important factor in our statistics, we added it and arrived at a statistics which we call Adjusted Residual IDF (short: ADRIDF. cf. equation 5).

$$ADRIDF = RIDF\sqrt{tf} \qquad (5)$$

In our project, we have implemented and evaluated the appropriateness of these statistical measures in ranking the most relevant keywords from our multilingual learning objects. In section 3., the results are discussed in detail.

The keyword extractor is built to deal with a wide range of languages: Bulgarian, Czech, Dutch, English, German, Polish, Portuguese and Romanian. This is a relevant result since techniques to extract keywords and keyphrases have been usually tested on English and never on such a variety of languages.

On the one hand, the management of such a wide range of languages makes it necessary: a) to build a common annotation format for all annotated corpora and b) to keep the language specific components of the tool as lean as possible. On the other hand, the multilingual aspect of the development gives us the chance to broadly evaluate the performance of the tool and its underlying information extraction methods, as discussed in detail in the next section.

It should be noticed that the great variety of languages we have dealt with, and the high number of parameter settings used to find the optimal results complicates the evaluation procedure and makes it necessary to perform several tests and experiments. The evaluation strategy must be both formative – i.e. inform the development of the tool, in particular the language-specific settings – and summative – i.e. assess the performance of the final tool. It has to be both intrinsic – i.e. assess the performance of the tool in isolation – and extrinsic – i.e. assess the performance of the tool as part of a learning environment.

As discussed more at length in section 3., the novelty of our application makes it difficult to adapt current evaluation tests for our purposes. On the other hand, we do need to assess the performance of the tool and verify that it achieves acceptable results before integrating it in the Learning Management System. Therefore we have to accept the limita-

language	Keyords (%)	Keyphrases (%)
Bulgarian	57	43
Czech	73	27
Dutch	75	25
English	38	62
German	90	10
Polish	33	67
Portuguese	86	14
Romanian	70	30

Table 1: Percentages of keywords and keyphrases per language

tion of these tests and work towards the development of new ones more fit to the purpose. It should be noticed that a non-optimal performance of the tool in the intrinsic evaluation might still lead to an appropriate behavior of the keyword extractor in the extrinsic evaluation. In particular, a scenario based evaluation of the tool in the context of the ILIAS system which takes into consideration the degree of satisfaction of the user and the impact on the learning process might be a more appropriate way to validate the keyword extractor.

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. E.g. for Polish we have 67 % keyphrases of two or more words. For other languages, e.g. German, multi word key phrases do not play a significant role, see table 1 for details.

We therefore put some effort to properly deal with these items and several tests have been carried out to detect the most appropriate length for multiword keywords and possible variation due to language. We followed the approach of Yamamoto and Church, cf. (Yamamoto and Church, 2001), to effectively identify and extract recurrent multi-word key phrases up to a predefined length. Additionally, we used linguistic information to further restrict this set of multiword key phrases, e.g. to exclude phrases which end in a preposition. Statistically, multi-word phrases are treated as single words.

Providing users with multiword keywords raise the issue of which should be the best way to represent them. We have

noticed that, at least for some languages such as Polish, a sequence of base forms looks quite unnatural therefore we have decided that the selected multi-word keywords are represented by their most frequent attested forms.

We refer to (Lemnitzer et al., 2007) for additional details on the use of the keyword extractor within the LT4eL project.

3. Evaluation of the keyword extractor

The best way to validate the keyword extractor might be in the context of the Learning Management System, that is by authors or content providers which will employ it to annotate learning objects with LOM metadata semi-automatically. On the other hand, the keyword extractor which will be integrated into the LMS should be optimized for this task and thus a methodology should be developed to verify its appropriateness and to eventually improve its performance. Therefore, it is necessary to develop an evaluation methodology which should take into account the peculiarity of the application envisaged and the complexity of the task, that is the use of the tool for several languages.

There are certain parameters which have been taken into account in this process: a) the language(s) of the learning objects and the corresponding language models which influence the preselection of keyword candidates; b) the maximal length of keyphrases to be extracted; c) several distributional statistics; d) additional features to select and rank keywords such as the place where a word appears and layout features of a word.

Therefore, the verfication must be formative in a sense that it is repeated several times in the development cycle. It informs the optimization process for each language and verifies that certain changes or adjustments have a positive impact on the performance of the tool. The verification has also to be summative in the sense that at the end of the optimization process the overall performance for each language should be assessed. The optimized tool has been integrated into the Learning Management System, where it is currently validated in terms of the impact of this functionality for the learning process.

In the rest of the section, we describe three tests which we have foreseen to evaluate the keyword extractor.

Test 1 In order to have a rough idea of the performance of the tool, we have measured recall and precision of the keyword extractor for each language and each appropriate parameter setting. A gold standard has been established on the basis of manually annotated keywords (i.e. 1.000 keywords for each corpus of learning material). This part of the evaluation has been performed automatically.

Test 2 In order to assess the difficulty of the task that the keyword extractor has to perform, that is how much variation there is among users in the assignment of keywords to a text, an evaluation of inter-anntotator agreement (IAA) on the keyword selection task has been performed.

Test 3 In order to evaluate the appropriateness of the keyword extractor in the context of the semi-automatic metadata annotation of the learning objects, we confronted test participants for each language with a document and a limited set of keywords which have been extracted and ranked automatically 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.

3.1. Test 1: measuring performance of the keyword extractor

This evaluation of the keyword extractor is based on the keywords which have been selected and annotated manually. More specifically, for each language, at least 1000 keywords have been manually selected and marked in the corpus of learning objects by one annotator. Table 2 gives additional information on how many documents were annotated per language, how many keywords were annotated per document and average number of keywords per document. In this step of the evaluation, automatically extracted keywords have been matched with the manually selected ones. Thus, the manually selected keywords are used as gold standard. Recall and precision of the keyword extractor are measured against this standard in the following way:

- For each document d_i, let $WM = wm_1 \ldots wm_n$ be the set of manually selected keywords. Let N be the number of these keywords. For each i, j, if $i \neq j$, then $wm_i \neq wm_j$.

- Let $WA = wa_1 \ldots wa_m$ be the keywords selected by the keyword extractor and M the number of these keywords, such that $M = N$. For each i, j, if $i \neq j$, then $wa_i \neq wa_j$.

- Both WM and WA contain two subsets: WMS and WAS, the subsets of single word keywords, and WMM and WAM, the subsets of multi word keywords.

- For each element in WMM and WAM, the length is calculated as the number of words. If wm_k is a two word keyword, then $L_{wm_k} = 2$. For each single word keyword wm_l, $L_{wm_k} = 1$.

Recall and precision are calculated as follows:

- For each $i : 0\ M$, check whether wa_i matches any $wm element WM$. If this is the case and the match is exact, add a match value of one. All exact matches are summed up to a total value of EMV.

- If the match is partial, divide the length of the shorter keyword by the length of the longer keyword. If wa_l partially matches wm_k and $L_{wa_l} = 1$ and $L_{wm_k} = 3$, then the match value is $\frac{L_{wa_l}}{L_{wm_k}} = 1/3$.

- All exact matches and partial matches are summed up to a total value MV.

- Recall R: the recall of the keyword extractor is calculated as $\frac{MV}{N}$

- Precision P: the precision of the keyword extractor is calculated as $\frac{EMV}{M}$

- F2: the F2 measure is calculated as $frac2pr(p+r)$, i.e. no higher preference is given to either recall or precision.

2056

language	# annotated documents	# annotated KWs	KWs / doc
Bulgarian	42	3236	77
Czech	465	1640	3.5
Dutch	72	1706	23.6
English	45	1174	26
German	34	1344	40
Polish	25	1033	41
Portuguese	29	997	34.4
Romanian	41	2555	638

Table 2: Percentages of keywords and keyphrases per document

These calculations take into account that for the user it is better to be presented a part of a good multi word keyword than nothing at all. For precision, though, only the exact matches count.

Table 3 gives an overview of the performance of the keyword extractor for the various languages.

The various statistical measures employed, that is TFIDF, ADRIDF and RIDF, were tested on the various languages and results show that, in general, TFIDF and ADRIDF nearly produced the same results. ADRIDF performs better than TFIDF only in the case of Bulgarian and Polish, in all the other cases, performance is either the same of worst. RIDF performed worst for almost all settings; therefore, the statement of Church that residual inverse document frequency of a term is a good indicator of its keywordiness could not be proven. Simple frequency of occurrence of a term in a document plays a much more important role.

With respect to precision and recall, results varied significantly across languages. If we only consider TFIDF, the best result is 60 % for recall reached for Bulgarian and the worst is 18 % for German, while with respect to precision, the best result is again obtained for Bulgarian with 25 % while the worst is obtained for Romanian with 11 %. These values are influenced by two factors: a) the quality of the human judgment when selecting the keywords; b) the quality of the linguistic annotation of the corpora.

We also tested the impact of multiwords on results and we noticed that results improved for all languages if multiword keywords up to a length of 3 words were included. This is at least partially due to the fact that a higher proportion of multi word keywords increases the number of partial matches.

As already mentioned, this test was performed only to get a rough impression of the performance of the keyword extractor as well as to determine which statistical measure performed best and to determine the maximum length for multiword keywords. More generally, its major purpose lies in informing the developers by presenting those keywords which did not match with the manually annotated ones and by presenting those manually selected keywords which have not been extracted by the tool. Note that not all keyword candidates which do not match manually selected keywords are necessarily bad keywords.

In fact, we believe that there might be some variation among users in identifying keywords and it is for this reason that we have performed an experiment to measure inter annotator agreement which is described in detail in the following section.

3.2. Test 2: Inter annotator agreement

In order to assess the intrinsic difficulties of the keyword selection task and to verify the performance of the keyword extractor compared to human annotators, we have investigated the inter annotator agreement on at least one document for each language. More specifically, we wanted to investigate where the performance of our keyword extractor stands relative to the performance of a group of human annotators.

In the test described in the previous section, we have compared the output of the keyword extractor the the choices of one single human annotator. We have thus relied completely on the performance of this individual annotator, thus taking his choices as gold standard. The experiments which we describe in this section reveals how reliable the human judgement is and where the judgments of the keyword extractor stands relative to the human judgments.

A document of a manageable size – around 10 pages – has been chosen for manual keyword selection. The content of the learning object was chosen so that it would be easy to understand for the test persons: it was a document dealing with Multimedia belonging to chapter 3, part 7 of the Calimera Guidelines.[4] This material is available for all the languages under consideration.

A minimum of 12 test persons have been recruited for the experiment (with the exception of English and Czech). In the instructions, which have been written in English and have been translated into the eight languages we have tested, the annotators were asked to select not more than 15 keywords and to mark for each keyword how sure they were that this is a good keyword. A scale was given from 1 (very sure) to 3 (not so sure).

For German and Romanian, two experiments were performed. For German, the same text was given to two groups, a group of students who were not familiar with the topic and a group of experienced scientists. We wanted to investigate whether experienced scientists achieve a higher inter-annotator agreement than students who are not familiar with the topic. The Romanian group ran the experiment with two different texts to check whether characteristics of the text influence inter annotator agreement.

---

[4]http://www.calimera.org/

Bulgarian			
Method	Recall	Precision	F-Measure
ADRIDF	0.60	0.30	0.40
RIDF	0.57	0.29	0.38
TFIDF	0.60	0.30	0.39
Czech			
Method	Recall	Precision	F-Measure
ADRIDF	0.22	0.17	0.18
RIDF	0.14	0.10	0.11
TFIDF	0.23	0.17	0.18
Dutch			
Method	Recall	Precision	F-Measure
ADRIDF	0.34	0.24	0.27
RIDF	0.25	0.19	0.21
TFIDF	0.36	0.25	0.29
English			
Method	Recall	Precision	F-Measure
ADRIDF	0.47	0.28	0.32
RIDF	0.33	0.18	0.22
TFIDF	0.48	0.26	0.32
German			
Method	Recall	Precision	F-Measure
ADRIDF	0.16	0.14	0.15
RIDF	0.15	0.12	0.13
TFIDF	0.18	0.15	0.16
Polish			
Method	Recall	Precision	F-Measure
ADRIDF	0.42	0.19	0.26
RIDF	0.29	0.15	0.19
TFIDF	0.42	0.19	0.25
Portuguese			
Method	Recall	Precision	F-Measure
ADRIDF	0.30	0.17	0.21
RIDF	0.21	0.12	0.15
TFIDF	0.31	0.18	0.22
Romanian			
Method	Recall	Precision	F-Measure
ADRIDF	0.26	0.12	0.15
RIDF	0.24	0.12	0.15
TFIDF	0.26	0.11	0.15

Table 3: Performance of the keyword extractor for the various languages

To measure pairwise inter-annotator agreement, both between human annotators and between KWE and human annotators, we used the so-called AC1 measure proposed by Kilem Gwet, cf. (Gwet, 2001) and elaborated by Debra Haley, cf. (Haley, 2007). Gwet and Haley investigate Cohen's kappa for inter-annotator agreement, which is normally applied to such tasks, and argue convincingly that under certain conditions this formula leads to unreliable results. In particular, κ is affected by skewed distributions of categories (the prevalence problem) and by the degree to which the coders disagree (the bias problem). In our experiments we deal with a skewed distribution. In particular, we have a large portion of elements (words) which are marked

Table 4: Contingency table for IAA

A	B	$B1 = A + B$
C	D	$B2 = C + D$
$A1 = A + C$	$A2 = B + D$	N

by neither annotator as a keyword. It is therefore appropriate to use Gwets AC1 statistics.

The AC1 formula, applied to our "two annotators and two categories" setting is as follows. Using the 2 times 2 contingency table below we arrive at the following equation for AC1

$$AC1 = \frac{\frac{A+D}{N} - 2\frac{\frac{(A1+B1)}{2}}{N}\left(1 - \frac{\frac{A1+B1}{2}}{N}\right)}{1 - 2\frac{\frac{(A1+B1)}{2}}{N}\left(1 - \frac{\frac{A1+B1}{2}}{N}\right)} \quad (6)$$

Using this metrics, we get the following results in terms of average inter-annotator agreement[5]:

Table 5: IAA per language

Language	Average human annotator agreement	KWE agreement with human annotators (using optimal settings)
Czech	0.71	0.78
Dutch	0.67	0.72
English	0.62	0.82
German	0.64	0.63
Polish	0.63	0.67
Portuguese	0.58	0.67
Romanian	0.59	0.61

Results of these experiments revealed that the inter-annotator agreement for this task is not very high for all languages, indicating that the task of selecting keywords cannot be defined in a clear way. We could not detect a significant difference between languages, nor between unexperienced and experienced annotators. It can be concluded that various sets of keywords can be identified which serve the purposes to represent the content of a document in similarly well. This would suggest that the choice of the key word extractor, as one way to identify keywords, yields results which will be acceptable to the user of the tool, even if the overlap with results of a particular human annotator are low. We ran a third experiment to test this assumption.

3.2.1. Assessing the adequacy of KWE-selected keywords

An evaluation which is fairly standard and which is e.g. performed by Velardi (cf. (Sclano and Velardi, 2007)) on a similar task is to expose user to a) a given document and b) a set of automatically extracted keywords and let them

---

[5]For Bulgarian, the collected data were too sparse to yield adequate results.

judge how adequate the selected words are as keywords for this document.

In this sections we report about the results and findings of this experiment which has been run for all languages. A document of moderate size – around 10 pages – was chosen and keywords for this diocument extracted. Six to twelve test persons per language took part in the experiment.

We presented test persons with the 20 highest ranked keywords and asked them to judge their adeqacy on a scale from 1 to 4:

- 1 = very relevant (would be a definite searching term)

- 2 = quite relevant (would be a secondary searching term)

- 3 = not relevant to the document

- 4 = not a valid term

Additionally, a value could be given in the case that the test person was not confident enough to decide. Test persons where also given the opportunity for adding keywords which they were missing in the list. Not all language group used this opportunity though.

The following table summarizes the average values for each language:

Table 6: Average acceptance rate for generated keywords per language

Language	Average score for first 20 KW	Number of additional keywords suggested
Bulgarian	2.21	21
Czech	2.22	none
Dutch	1.93	12
English	2.15	22
German	2.06	none
Polish	1.95	45
Portuguese	2.34	7
Romanian	2.14	none

From table 6 on can infer that the results are acceptable for all languages, with some room left for improvement. Second, we observed a tendency towards better scores for the first 10 keywords, in contrast to the score for the first 20 keywords and for some languages even compared to the score for the first 5 keywords. From these results we can infer that it is a good decision to present the user, in the real system, the first ten keywords, with the option to look at further keywords if they want to.

The lists of additional keywords are used by the developers and language groups as an advice to optimizing the tool. They also prove the usefulness of allowing, in the real system, users to add their own keywords in addition to those which are suggested by the system and approved by them.
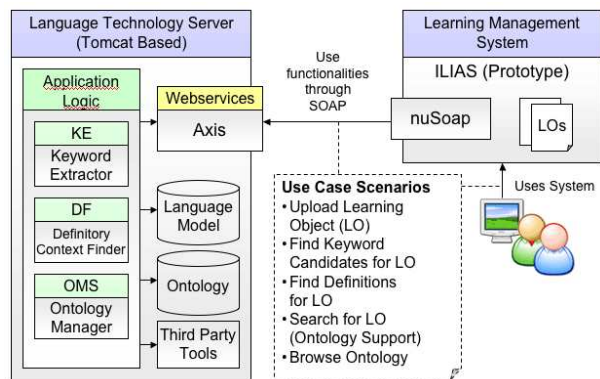

Figure 2: Architecture of the Language Technology enhanced LMS

4. Integration into ILIAS

The keyword extractor is a functionality which has been integrated in a learning management system to support the semi-automatic metadata annotation of the learning objects. It should assist authors of learning material to find and assign appropriate keywords to their learning objects. In the context of the LT4eL project, the tool has been integrated in the ILIAS learning management system even though it should be possible to enhance other LMS with it.

The tools and data reside on a dedicated server and are approached from inside the Learning Management System via Web Services. Figure 2 shows the major components of the integration setup. The language technology server on the left provides the keyword extractor and other NLP components (cf. (Lemnitzer et al., 2007) for more details). The functionalities can be accessed directly on the webserver for test purposes or they can be used by the learning management system through the web service interface. Figure 3
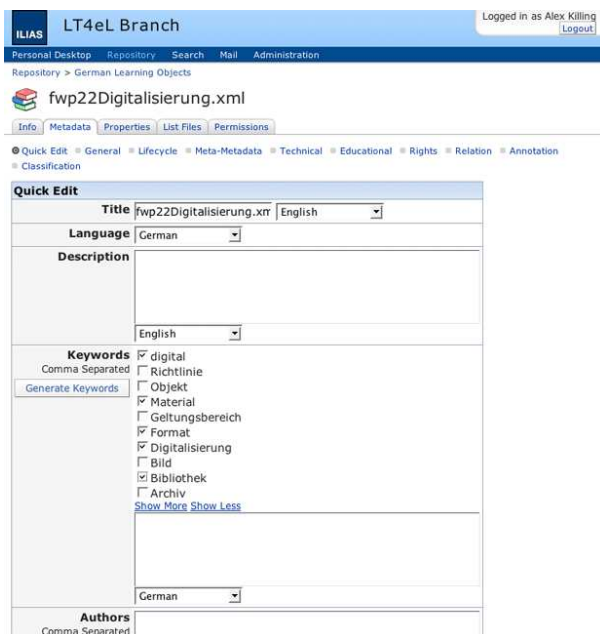

Figure 3: User interface to the Keyword Extractor

shows the first integration of the keyword extractor into the

ILIAS learning management system. The function is embedded into the existing LOM metadata handling of ILIAS to enable a semi-automatic generation of keywords. Users can: a) run the keyword extractor and get a list showing a predefined number of keywords for a document; b) select keywords from this list and c) and add their own keywords. The interactivity is an important feature of this tool. It will not be used to completely perform the task of keywording a documents, but to make informed suggestions which the user has to approve or reject.

The best way to evaluate the tool is in the context of its use within ILIAS. Therefore, a scenario based evaluation, which will take user satisfaction into account, might be the best way to assess its performance.

5. Conclusions and future work

One of the functionalities developed within the LT4eL project is the possibility to annotate learning objects semi-automatically with keywords that describe them, to this end a keyword extractor has been created. The approach employed is based on a linguistic processing step which is followed by a filtering step and keyword ranking based on frequency criteria.

Three tests have been carried out to provide a rough evaluation of the performance of the tool, to measure inter annotator agreement in order to determine the complexity of the task and to evaluate the acceptance of the extraced keywords by users.

The results are promising also considering that the task has been carried out for 8 different languages. However, there are possible ways in which the results of the keyword extractor could be improved.

Keyword candidates tend to appear in certain salient regions of a text. These are the headings and the first paragraphs after the headings as well as an abstract or summary. Salient terms might also be highlighted or emphasised by the author, e.g. by using italics. Investigations of the manually annotated keywords have shown that a word with a salient position or marked by layout features is at average twice as probable to be marked as keyword as those words which do not bear these features. Therefore, we will use layout information as additional filter in proposing salient keywords (cf. also (Sclano and Velardi, 2007)).

Currently, a user- and scenario-oriented evaluation is being performed in order to evaluate the influence the impact of the keyword extractor and its results on the learning process. The scenarios chosen are, in short: for tutors, to find relevant documents for a certains topic, probably in various languages, for an international audience. Keywords should be helpful in this task; for students: preepare a paper synopsis for a seminar presentation. Again, the keywords should give first hints about the relevance of a learning object for this task. Of course, the most important use of the tool for authors is to facilitate the assignment of keywords to their learning objects.

As part of the our dissemination activities, we maintain a user panel. Registered users will be informed about the status and availability of our tools and data. If you want to join the panel, please inform us.

6. References

Kenneth W. Church and W. Gale. 1995. Inverse document frequency (idf): A measure of deviations from poisson. In *Proc. Third Workshop on Very Large Corpora*.

K. Church and W. Gale. 1995a. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190.

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

Kilem Gwet. 2001. *Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement Between Two or Multiple Raters*.

Debra Trusso Haley. 2007. *Using a New Inter-rater Reliability Statistics*. Ph.D. thesis, The Open University, Milton Keynes.

A. Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of EMNLP2003*.

S. Jones and G. W. Paynter. 2006. An Evaluation of Document Keyphrase Sets. *Journal of Digital Information*, 4(1).

L. Lemnitzer, C. Vertan, A. Killing, K. Simov, D. Evans, 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.

R. Mihalcea 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.

Paola Monachesi, Lothar Lemnitzer, and Kiril Simov. 2006. Language Technology for eLearning. In *Proceedings of EC-TEL, Crete 2006*. Springer.

F. Sclano and P. Velardi. 2007. TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. In *Proc. of the 3rd International Conference on Interoperability for Enterprise Software and Applications I-ESA 2007*, Funchal, Madeira Island, Portugali.

P. D Turney. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2:303–336.

X. Wan, 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*, pages 552–559, Prague.

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

M. Yamamoto and Kenneth W. Church. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, 27(1):1–30.

H. Y. Zha. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of SIGIR2002*, pages 113–120.