

Language Technology for eLearning – Implementing a Keyword Extractor

Lothar Lemnitzer*,
Łukasz Degórski**

* University of Tübingen
Wilhelmstr. 19, 72074 Tübingen, Germany
{lothar}@sfs.uni-tuebingen.de

** Institute of Computer Science, Linguistic Engineering Group, Polish Academy of
Sciences
Ordona 21, 1/1/237 Warsaw, Poland
{ry-ba}@wp.pl

Abstract. We will describe the design and implementation of a keyword extractor in the context of eLearning. This tool should extract lexical units from learning objects which are highly topical for the content of this learning object and therefore good keywords. Authors of learning objects will be supported in generating this type of metadata. We will describe how we formalise the notion of *keywordiness* and use distributional characteristics in two statistical approaches which yield an ordered list of keyword candidates. The keyword lists which are generated are compared to manually compiled lists in order to evaluate the recall and precision of the tool.

1 Introduction

Given the huge amount of static and dynamic content created for eLearning tasks, the major challenge for extending their use is to improve the effectiveness of retrieval and accessibility by making use of Learning Management Systems (LMS) and Learning Object Repositories. The aim of the EU funded project *Language Technology for eLearning (LT4EL)* is to tackle this problem by providing Language Technology based functionalities and by integrating semantic knowledge to facilitate the management, distribution and retrieval of the learning material. The project wants to improve on existing eLearning platforms and repositories by allowing for the construction of user specific course material through the semantic querying for topics of interests. In order to reach this objective, we apply relevant research which has been carried out in the area of Language Technology and the Semantic Web. Specifically, we improve the retrieval of static and dynamic content by employing linguistic resources and tools for the semi-automatic generation of descriptive metadata and glossaries. We develop new functionalities such as a keyword extractor and a glossary candidate detector which are tuned for all the languages represented in our consortium, i.e.

Bulgarian, Czech, Dutch, English, German, Maltese, Polish, Portuguese, and Romanian. The new functionalities could be integrated in any open source LMS or learning object repository, however, for validation purposes the ILIAS Learning Management Systems (LMS) (<http://www.ilias.de>) has been selected.

In this paper, we will focus on the keyword extraction tool which will support authors and distributors of learning objects in the task of keyword generation. In section 2 we will motivate our task of automatically extracting keywords and put this task into a larger context. Section 3 is about the characteristics of good keywords. These characteristics will be operationalised for the keyword extraction task. In section 4, we will describe the linguistically annotated input. The following section is dedicated to the methods we are going to use in order to measure keywordiness and to some additional weighting which we will introduce to the keyword ranking. In section 6 we will explain how we will handle multi-word terms. Evaluation of the results is the topic of section 7. We will finish with a look at related work and an outline of further work for the generation of metadata and glossaries.

2 Semi-automatic metadata generation for eLearning Contents

In eLearning, we deal with static and dynamic contents. The static contents comprise to the largest part of learning objects of varying granularity. A learning object should be accompanied by metadata which describe it in terms of its content, its context of use as well as its interrelation with other learning objects. *Learning Object Metadata* (LOM) [1] has become the most widespread and well-known standard for the encoding of the metadata. The aim of the LT4eL project is to improve the retrieval and accessibility of eLearning content through the identification of the learning material by means of descriptive metadata. To this end, we employ available Language Technology resources and tools to develop functionalities which facilitate the semi-automatic generation of content-related metadata. It is not yet an established practice to supply metadata along with the contents of eLearning platforms. Their generation is a tedious task and not widely accepted by authors as part of their work. This has, however, the highly undesirable consequence that these contents are invisible outside the narrow scope of a particular LMS and not easy to disseminate, e.g. through learning object repositories. Language technology can provide significant support for this task. In particular, within LOM, the element *General* allows for the representation of a document's content by topical keywords.

The content-oriented functionalities being developed within the project can play a crucial role in the semi-automatic generation of metadata. The result will be two modules:

- semi-automatic selection of keywords;
- selection of definitory contexts, to be used for glossary entries.

The term *semi-automatic*, in that context, implies that the final selection of metadata remains under the control of the author. These functionalities will be realised as software modules which can be used stand-alone, as web services, or integrated into Learning Management Systems.

In the following section, we will describe the keyword extraction task in more detail.

3 Characteristics of good keywords

The keyword extractor is based on quantitative and qualitative assumption about the characteristics of good keywords.

Keywords are supposed to represent the topic(s) of a text. They therefore tend to appear more often in that text than could be expected if all words were distributed randomly over a corpus. In other words: typical keywords tend to appear frequently in certain documents and will not appear at all in most other documents. A statistics which is used to model the expected distribution of words in texts is Poisson or, alternatively, a mixture of Poissons (cf. [4]). While the distribution of function words is close to the expected distribution under the Poisson distribution model(s), 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 ([3]).

A second distributional characteristic of keywords is what has been called their *burstiness* (cf. [8]). Good keywords tend to appear in clusters within documents. It can be observed that, once a keyword appeared in a text, it tends to appear in the following section of the text in much shorter intervals as would be the case if all the occurrences of this word were distributed evenly throughout the text. When the burst ends, the frequency comes back to normal, until the keyword appears - and probably bursts - again.

This distributional behaviour reflects the fact that good keywords represent topics, and reference to a certain topic is a local phenomenon: most texts deal with many (sub)topics and in some texts a topic is resumed after a while.

Additionally, good keyword candidates can be characterised qualitatively. They appear typically 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. Words occurring in these areas of the text will be weighed higher than the rest. Central terms might also be highlighted or emphasised by the author, e.g. by using bold font or underlining the word.

It has been shown that the results of the keyword selection, measured against the performance of human readers, improves significantly if the text is annotated linguistically (cf. [5] and [6]). Linguistic data and tools such as part of speech taggers and shallow parsers are needed for this task and are being supplied in this project for the various languages.

In the following section we will describe the way from the raw text of learning objects to full linguistic annotation.

4 Linguistically annotated input

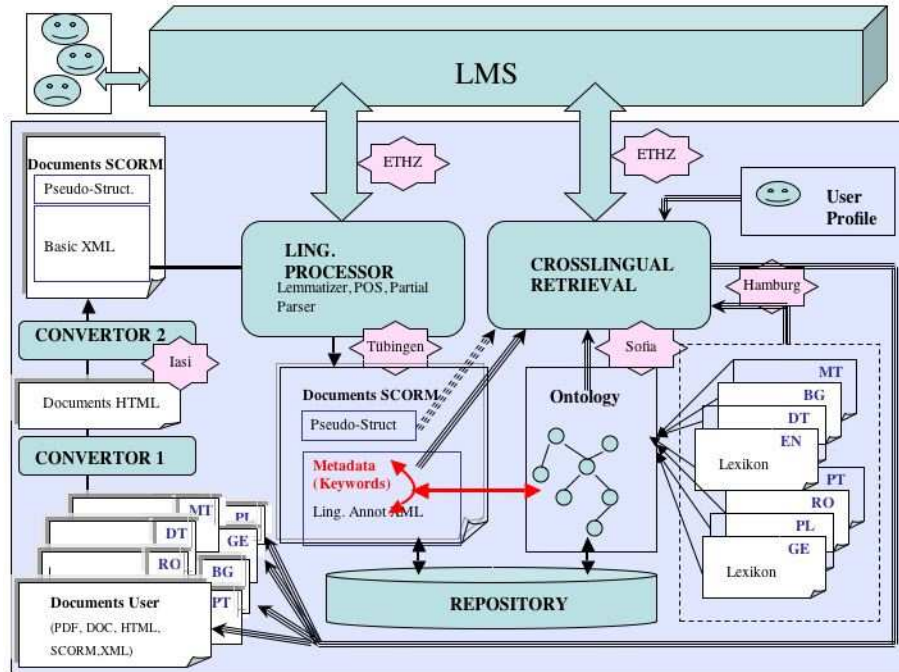


Fig. 1. LT4EL system architecture and workflow

Figure 1 describes the workflow which leads to linguistically annotated learning objects for all languages in the project. The keyword extractor draws on this linguistic information.

1. Learning objects of various formats, e.g. PDF and DOC, are converted into HTML, using standard tools.
2. The HTML files are then converted into a project specific XML format, called BaseXML. This is basically pure text with some structural and layout information that has been extracted from the HTML markup. Only the information relevant for metadata generation tasks is preserved.
3. Linguistic annotation will be added to the BaseXML files. Each language partner uses their own linguistic tools to provide the annotation.
4. The different output formats of the linguistic annotation are harmonised by converting them into one common document format which is determined by a DTD.

The linguistic tools segment the text into linguistically relevant units and describe these units. The units are: sentence, chunk, and token. The most relevant

unit for keyword extraction is the token. Tokens are classified for the parts of speech (noun, verb, etc.) and their morphological features (number, person, etc.). In addition, the baseform of the token is specified. The following is an example from the Polish corpus:

```
<tok id="t226"  
base="szata"  
ctag="subst"  
msd="sg:nom:f">  
    Szata  
</tok>  
<tok id="t227"  
base="graficzny"  
ctag="adj"  
msd="sg:nom:f:pos">  
    graficzna  
</tok>
```

The keyword extractor makes the following uses of the linguistic annotation:

1. The software will work on *Lexical units*. Lexical units are abstractions from the inflected forms which are used in the texts (e.g. the tokens *system* and *systems* will both be counted as occurrences of the lexical unit /system/). Note that we do not try to disambiguate these lexical units semantically. Instead, we expect that a lexical unit which is a good keyword appears in only one sense throughout the text.
2. We use the part of speech information to classify lexical units as either relevant (mainly nouns) or irrelevant (mainly function words).

5 The statistical measures

5.1 Residual inverse document frequency

Two statistics which model the distribution of a term w in a collection of documents are the *document frequency* df_w and the *collection frequency* cf_w . Document frequency is the number of documents in which the term occurred at least once, while the collection frequency is the total number of occurrences of this term in the whole corpus. Inverse document frequency *IDF* for the word w is defined as $-\log_2(\frac{df_w}{D})$, where D is the total number of documents in the collection. We can also model *IDF* assuming that the appearances are independent and follow a Poisson distribution: ($IDF = -\log_2(1 - e^{-\frac{cf_w}{D}})$). It has been shown that function words follow largely this distribution, while interesting content words deviate from this distribution (cf. [3]). Thus, the difference between the predicted and real *IDF* is smaller for function words, and larger for content words. This difference, called residual inverse document frequency, will be used to measure the keywordiness of w .

5.2 Term burstiness

Term burstiness is a formal description of term (re)occurrence patterns in texts. It has been introduced by Slava Katz (cf. [8]) and developed further by Anne de Roeck and her team (cf. [9]). This approach is in contrast to those approaches, including RIDF, which view a text as a bag of words without taking its inner structure into account.

The concept of term burstiness, as described above, can be formalised by measuring the gaps between the individual occurrences of each term in a text. Sarkar et al. (cf. [9]) use a combination of two exponential functions to model term burstiness. The functions describe the reoccurrence rate of a term before and after it appeared in a text (the further being modeled after the reoccurrence rate of the term in other texts).

Intuitively, the reoccurrence rate of before and after the (re)appearance of the term vary significantly for good keywords. The authors show that the model is good in distinguishing a) frequently occurring common function words; b) frequent, but well spaced function words; c) infrequent and scattered function words and d) topical content word.

Since we are interested in the latter, we will experiment with term burstiness statistics.

5.3 Combining the measures

De Roeck et al. show that in the domain of topical content words their results match the results which have been yielded e.g. by Church (cf. [3]) very well. Nevertheless, we will compare the ranking of keyword candidates according to both measures. If the order of terms varies for our data, we will introduce a measure which combines the output of both methods.

Additional weight (a bonus) will be given to lexical units which:

- appear in salient regions of the text (header, first paragraph)
- are emphasised or highlighted

The exact measure of this bonus has to be determined experimentally.

6 Multi word lexical units

Special attention will be given to multiword terms. A first analysis of Polish learning objects that have been keyworded manually revealed that a large part of these keywords were multi-term units (e.g. *wirtualna szkoła*, *system zarządzania nauczaniem*).

We will, for all terms which are highly ranked as keyword candidates, use a statistics which measures the connectedness of these terms with their neighbours. We choose mutual information (MI, in short) for ranking pairs of keyword candidates and their neighbours for connectedness. MI measures the observed co-occurrence of two terms against the expected co-occurrence of these terms,

given their individual frequency in the corpus and the assumption that all words are distributed randomly. This statistics is well-known for identifying technical terms, and it has been shown by Yamamoto and Church that terms which rank high both in RIDF and MI score are both topical and terminologically interesting (cf. [10]).

Therefore we will inspect the neighbourhood of all high ranking keyword candidates to detect multi word units.

7 Evaluation

We want to know how well the keywords which our programme extracts automatically will match those items which humans marked as relevant. In the project, learning objects for all the involved languages will be annotated manually for relevant terms. We will use the same learning objects to extract our lists of keyword candidates and set a threshold as a cutoff point for our keyword list. For those lists we will measure:

- How many of our keyword candidates are also in the list of manually selected keywords (this approximates the recall measure);
- How many of our keyword candidates are not in the list of manually selected keywords (this approximates the precision measure).

We are however aware that the notion of keyword is a vague one and that no two human annotators will come to the same lists of keywords. Human annotation is therefore not an ideal gold standard to measure our data against. We therefore decided to let two humans annotate the same Polish and German learning objects and to measure the inter-annotator agreement (cf. [2]). It seems to us a much fairer to assess the performance of the automatic keyword extractor against the performance and agreement of the human annotators.

8 Related work

Jelena Jovanović et al. (cf. [7]) provide, with their TANGRAM system, a solution for the automatic generation of metadata. They also work with the LOM metadata scheme. Their approach however, works top down. They start with a domain ontology and look in the learning objects for those terms which are derived, as lexical units, from the concepts of this ontology. We think that a top-down approach matches well with the bottom-up approach which we have outlined here. Indeed it is one of the goals of the LT4EL project to combine both approaches.

The LON-CAPA learning content management and assessment system also uses automatic keyword generation (www.lon-capa.org). In this system, however, the author is presented with all textwords which are not in a stopword list. If the user does nothing, the most frequently chosen keywords from a list are chosen (Gerd Kortemeyer, personal communication). So the system relies on an existing

community working in one domain and supplying keywords for learning objects in that domain. As far as we can see, there is no language technology used in this approach, which more resembles data mining. Relying on the collaborative work of a community of authors is an interesting approach which can be seen as complementary to our approach.

9 Conclusions

With LT4EL, we want to use Natural Language Processing techniques and resources to enhance the usability and effectiveness of eLearning systems and processes.

On the one hand, we want to support authors of learning materials in the tedious task of metadata generation. This part of the task should neither be avoided by the author, nor should it distract too much from the original task, the generation of high-quality learning material. On the other hand, we want to employ language resources, in particular ontologies, to help users find the best learning material they can get to satisfy their current information needs.

We consider the diversity of languages in the European context not as an obstacle to an international eLearning infrastructure, but as a challenge. By tackling multilingual issue we want to broaden the horizon of the learner beyond what is available in his native language and in English.

References

1. *Draft Standard for Learning Object Metadata*, IEEE, 2002 – P1484.12.1
2. Thorsten Brants. 2000. *Inter-Annotator Agreement for a German Newspaper Corpus*. In: Proc. of the Second International Conference on Language Resources and Evaluation (LREC-2000). Athens, Greece. <http://www.coli.uni-saarland.de/~thorsten/publications/>
3. K. Church, 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. <http://acl.ldc.upenn.edu/W/W95/W95-0110.pdf>
4. K. Church, W. Kenneth and W. Gale. 1995. *Poisson Mixtures* In: Natural Language Engineering 1(1995)2. pp. 163–190.
5. Anette Hulth. 2003. *Improved Automatic Keyword Extraction Given More Linguistic Knowledge*. In: Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing. pp 216–223.
6. Anette Hulth. 2003. *Analysing Term Selection Approaches and Features for Automatic Keyword Extraction*. In: NoDaLiDa '03, Proc. of the 14th Nordiske datalingsvistikkdager.
7. J. Jovanović, Dragan Gašević, and Vladan Devedžić. 2006. *Ontology-Based Automatic Annotation of Learning Content*. In: International Journal on Semantic Web and Information Systems, 2(2006), 2. pp 91–119.
8. S.M. Katz. 1996. *Distribution of content words and phrases in text and language modelling*. In: Natural Language Engineering 2(1996)1. pp. 15–59.

9. Avik Sarkar, Paul H. Garthwaite, and Anne De Roeck. 2005. *A Bayesian Mixture Model for Term Re-occurrence and Burstiness*. In: Proc. of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). Ann Arbor, Michigan. ACL. pp 48–55. <http://www.aclweb.org/anthology/W/W05/W05-0607>.
10. M. Yamamoto, and K. Church. 2001. *Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus* In: Computational Linguistics 27(2001),1. pp 1–30.