Automatic Keywords Extraction – a Basis for Content Recommendation

Ivana Bosnić[1], Katrien Verbert[2], Erik Duval[2]

[1] Faculty of Electrical Engineering and Computing, University of Zagreb,
Unska 3, HR-10000 Zagreb, Croatia
[2] Dept. Computerwetenschappen, Katholieke Universiteit Leuven,
Celestijnenlaan 200A, B-3001 Leuven, Belgium
ivana.bosnic@fer.hr, {katrien.verbert, erik.duval}@cs.kuleuven.be

Abstract. This paper describes a use case for an application that recommends learning objects for reuse and is integrated in the authoring environment. The recommendations are based on the automatic detection of content being authored and the context in which this resource is authored or used. The focus of the paper is automatic keyword extraction, evaluated as a starting point for content analysis. The evaluations explore whether automatic keyword extraction from content being authored is a sound basis for recommending relevant learning objects. The results show that automatically extracted keywords are suitable for this purpose, if some observed issues are appropriately addressed.

Keywords: content, reuse, recommendations, keywords, keyword extraction

1 Introduction

Content reuse today – although somewhat increased by new technologies and interfaces to aggregate and remix the content – is still not straightforward for mainstream authors of educational content. Barriers limiting content reuse include the immaturity or absence of support for discovering and reusing learning content in authoring tools and difficulties associated with combining and referencing reused learning materials [1]. The goal of our research is to analyze the reuse potential of learning objects and to support their discovery, recommendation and reuse within available authoring tools. Recommendation is based on both the content being authored and the context in which the content is authored or used. This paper analyzes whether the results of automatic keyword extraction from the content being authored can be a basis for recommending resources relevant to the author. These keywords are generated based on both the on-the-fly analysis of content the author is editing, and context data that is available in an authoring or learning environment. Our research, presented in this paper, focuses primarily on the results of keyword extraction analysis, and on describing the process of content reuse which is based on this topic analysis and integrated in the authoring environments.

The paper is organized as follows: The application use case is presented in section 2. Automatic keyword extraction services are presented in section 3. Section 4 describes the comparison between two keyword generation services, while section 5 describes the keyword evaluations in the application prototype. The paper wraps up with conclusions and future work in section 6.

2 Application Use Case

The application purpose is to help authors of educational content, by:

- recommending relevant content during authoring, without manual searching by the author;
- enabling easier content reuse and remix, particularly of small fragments, by referencing or using advanced copy-paste functionalities;
- integrating these functionalities in the authoring or learning environments through extensions of applications such as wikis, blogs, or presentation software.

One of the application use cases can be described with the following steps:

1. The user authors the content in his authoring environment (e.g. Wiki);
2. The application collects the content being authored, together with context data available (e.g. age range, difficulty level) and proposes the recommendations;
3. The user views the recommendations to decide whether they are relevant to him;
4. If the content is useful for either copying partly or just for getting ideas, then the user chooses to reference this content. The reference is automatically inserted in the content being authored, in the appropriate format (e.g. WikiMarkup, HTML markup or plain text);
5. As the user continues to edit the content, the changes are incorporated and new recommendations are presented.

In order to discover the resources, the application, integrated in the authoring environment, analyzes the content being authored. An automatic keyword extraction service extracts keywords from the text. Additional context is obtained from the authoring or learning environment (the purpose of the course, the preferred format of resources to be reused, etc.). Together with the keywords, this context data is used to search and retrieve relevant resources from content providers, including large learning object repository networks and social bookmarking websites.

3 Keyword-Based Content Discovery

The usual way of querying content providers is by using keywords as search terms. In the case of repositories containing learning object metadata, search terms can be used to query fields such as title, description or keywords and further refined by using additional metadata fields that capture the context in which the learning content is used. In this section, automatic keyword extraction services that can be used as a basis for generating search terms are presented.

Keyword extraction services can be divided in two groups, based on the usage of algorithms for constructing the semantic context:

- **term extraction services** – this group of services extracts the keywords from a text. Examples include Yahoo Term Extraction Web Service [2] and Fivefilters [3].

- **semantic entity extraction services** - this group of services not only extracts the keywords, but also detects the concepts related to the text, which are not present in the text itself. These services often have semantic linking features, i.e. they include additional encyclopedia links, images, articles, etc. Examples of such services are Zemanta [4], OpenCalais [5], Evri [6] and AlchemyAPI [7].

Most services provide interfaces for online use, mainly REST or SOAP. The usual result outputs are represented in RDF, XML, JSON or plain text. The services mostly use keyword classification schemes, such as the DBpedia ontology [8], Wordnet [9] or dmoz Open Directory Project [10]. Some services have their own entity databases.

Several comparisons of keyword extractors and semantic APIs exist. Zemanta and OpenCalais are recommended in [11], AlchemyAPI and Evri in [12], while [13] focuses on the characteristics of services for semantic tagging, without specific recommendations. Services from both groups were evaluated to compare and contrast their efficiency and potential use within our application:

- **Yahoo Term Extraction Web service** (Yahoo in the following text) is a popular keyword extractor with a RESTful interface, which returns up to 20 keywords that are found in the text. The keywords are not ranked internally. This service is successfully used in automatic metadata generation frameworks like SAmgI [14]. As SAmgI generates metadata for a subset of objects in the GLOBE network of repositories [15] that is used in our research, this was an additional reason to evaluate it for our purpose.
- **Zemanta** is a semantic entity extraction service with both RESTful and JavaScript interface. It returns up to 8 ranked keywords. Additionally, it recommends images, links to ~20 Web sites (Wikipedia, Youtube, IMDB, etc.) and blog/news articles from ~10000 sites. Optionally, Zemanta provides the keywords according to the *dmoz* keyword classification. Moreover, its extraction process can be influenced by emphasizing selected words.

The following section describes the comparison of these two services and the evaluation of their potential for automatic content discovery. In this evaluation, Zemanta and Yahoo were used to extract the keywords from several already existing presentations. These keywords were graded by users. In addition, the users were asked to manually provide keywords for the presentations and the keywords extracted by Yahoo and Zemanta were compared with these, user-generated keywords.

4 Evaluation of Keyword Extraction Services

4.1 Evaluation Methodology

The goals of this evaluation were to test the keyword extraction services with the examples of existing educational content, to compare the keywords extracted by Zemanta and Yahoo, and also to compare those to the user-generated keywords.

In the evaluation, 9 presentations were used – 3 for each topic (open source, databases and gravity force), different in their characteristics, which is expected to influence the quality of extracted keywords. A topic of open source mostly uses general words, descriptions and a smaller number of specific terms; a topic of databases is a more specific one, while an explanation of a gravity force contains formulas and lots of specific physics-related terms.

The presentations were gathered from Google's first page result on queries for "what is open source", "what is database" and "what is gravity", with file type filtering for Microsoft PowerPoint presentations. The excerpts chosen were text-only contents of 3 adjacent slides of each presentation, to better describe the context. Some slides had examples from other fields to help illustrate the concepts. Some texts were written as sentences, while others had only a few words per bullet. An assumption is made that the extraction services will have less success with shorter texts, partial sentences and the examples from different fields. However, these are often found in presentations, thus it should be tested whether keyword extraction gives satisfying results in those cases, too.

Six users were involved in the evaluation, which consisted of two parts:

1. The users were asked to read 9 text excerpts, and write the queries which they would use in search engines. They could type as many queries as they wanted.
2. For each of the 9 presentations, the users were presented with 8 keywords from Zemanta and the first 10 keywords from Yahoo. They were asked to grade the relevancy of each keyword, which, of course, could consist of one or more words.

4.2 Automatically Extracted Keywords

Two keyword extraction services were compared by the following criteria:

User keyword relevancy grading. Fig. 1 shows the average of relevancy grades per presentation. Zemanta is graded higher in 7 of 9 presentations.

If the same average is calculated for 3 **presentation topics**, it shows that the keywords from both services are graded higher as the topic specificity increases (Fig. 2). In all three topics, users have graded the keywords from Zemanta higher.

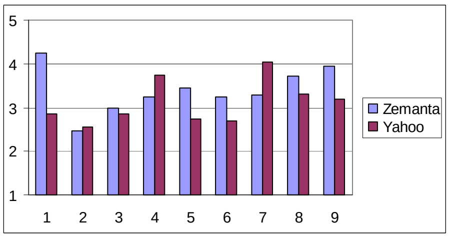

Fig. 1. The average of keyword relevancy grading per presentation. For each of the 9 presentations (X-axis), the users were grading the relevancy of 8 keywords from Zemanta and 10 keywords from Yahoo, with grades 1-5 (5 being the most relevant). The average of grades is calculated for two services separately (Y-axis). The grades for the same keywords were equally distributed among users.

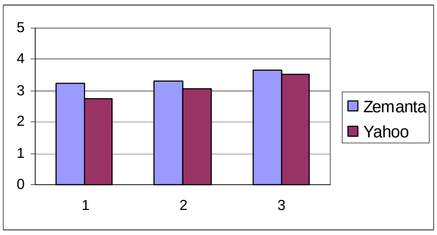

Fig. 2. The average of keywords relevancy grading (Y-axis) per presentation topic (1 – open source, 2 – databases, 3 – gravity) on the X-axis.

Fig. 3 shows the average of user grading for the keywords for each of the 8 Zemanta ranks. In general, the grading tends to drop as Zemanta ranking lowers, which justifies the decision to make queries by combining the highest Zemanta ranked keywords. Yahoo provides the keywords in order of appearance in the text, without any ranking mechanism, so this service could not be evaluated in this way.

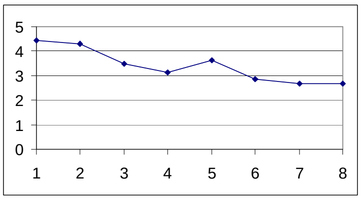

Fig. 3. The average user grading of keywords per particular Zemanta rank. The X-axis presents 8 Zemanta internal ranks. The Y-axis presents the average of user grades for the keywords in each Zemanta rank. In this diagram, the keywords from all 9 presentations were included.

4.3 User-Generated Keywords

To see how different the user keywords are from automatically extracted ones, the comparison of these two sets was made. This comparison is used to analyze how different are the results provided by keyword generation services from the user-proposed search queries - keywords. Only the keywords shared by at least two users were included, to provide more comprehensive and relevant results.

Two comparisons were made:

- exact match – checking whether the exact user-generated keyword was included in the list of extracted keywords. The difference in singular/plural form of nouns was counted as exact match, as most indexing services used can internally match these.
- similar match – checking whether a similar user-generated keyword was in the list of automatically extracted ones. The keywords as subsets of other keywords are considered similar (e.g. keyword “open source” is similar to “open source definition”), as well as the ones which could be matched with grammatical or syntax changes (e.g. keyword „gravity law“ is similar to „law of gravity“).

Fig. 4 shows the number of common user-generated keywords and the number of matches with automatically-generated keywords. The results show that the more important keywords – the ones which are common to more users – have a higher match rate. This is especially visible if similar matches are considered, which is an argument for use of advanced methods to find the keywords similar to automatically generated ones.

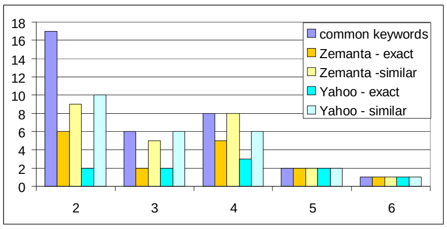

Fig. 4. The number of exact and similar matches between user-generated and automatically extracted keywords, in comparison to common keywords – the ones proposed by more than 2 users (Y-axis). The keywords are distributed by the number of users which proposed this keyword, shown in X-axis. With the exact match, Zemanta matches more words than Yahoo in 2 sets and in 3 is equal to Yahoo. With similar match, Yahoo matches more words than Zemanta in 2 sets, less words in 1 set and in 2 is equal to Zemanta.

The following section describes the initial keyword evaluations carried out in the application prototype environment, where the keywords had to be extracted during the presentation authoring. This approach poses additional challenges in text preparation and automatic keywords extraction, which are described in the following text. In these evaluations, the Zemanta extraction service was used.

5 Keyword Evaluations in the Authoring Environment

5.1 Evaluation Methodology

Two keyword evaluations were carried out. The overall goal of these evaluations was to determine whether automatic keyword extraction from content being authored is a sound basis for recommending relevant learning objects to the author. More specifically, the relevancy and ranking of the extracted keywords were evaluated. The evaluations were done as a part of an overall evaluation according to the *discount usability engineering* principles [16]. Therefore, it should be noted that these are not the results of thorough evaluations, rather of basic, initial user tests.

The users were asked to create an informative presentation about a programming topic familiar to them. The time was limited to 15 minutes. Specifically, the users were given an empty presentation template in the MediaWiki service, enhanced by the WikiPres extension – a MediaWiki plugin for collaborative presentation authoring using WikiMarkup [17]. They were advised to make use of the recommendation application, and to properly attribute reused resources.

Once the presentation was finished, the users chose one of the more content rich slides they authored (not the title or introduction slide). They were presented with 8 keywords generated for that slide and asked to rank the 5 keywords they considered the most relevant. Fig. 5 presents the relation of the user ranking and Zemanta ranking. Fig. 6 shows the averages of user rankings for keywords in the same Zemanta rank.

5.2 Evaluation 1

Four users ranked the keywords extracted and ranked by Zemanta. Of course, the generated keywords were different for each user: the user ranking is compared with that of Zemanta.

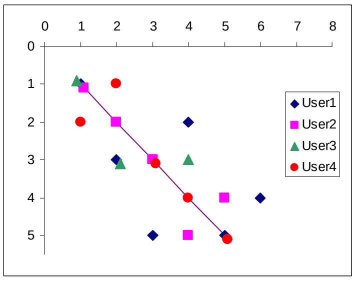

Fig. 5. The relation between the user and Zemanta ranking. The X-axis presents Zemanta ranks, from 1-8 (1 being the highest-ranked). The Y-axis presents user ranks from 1-5 (1 being the highest-ranked). The ranking itself is marked with a dot of a different type for each user. Ideally, the user and internal rankings would be identical, with all the dots on a diagonal line. Here, the dots are dispersed, but still near the diagonal line. The majority of dots are placed in the first five columns (Zemanta rank 1-5): this shows that users and Zemanta largely agree on what are the 5 most relevant keywords.

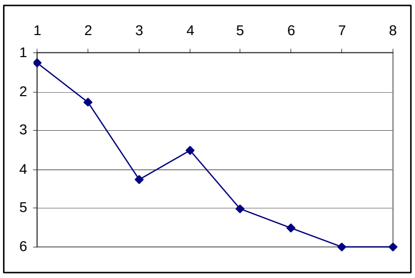

Fig. 6. The average user ranking. The X-axis presents Zemanta internal ranks. The Y-axis presents the average of user rankings for all keywords in a particular Zemanta rank. For instance, the highest-ranked keywords by Zemanta got 1, 1, 1 and 2 as user ranks, which gives an average of 1.25 out of 5. The diagram shows that the user ranking lowers together with Zemanta ranking; the keywords with the lowest Zemanta rankings are not among the most relevant to the users. For this calculation, the keywords not being among the 5 most relevant were given the rank 6.

Lessons learned. The interpretation of evaluation results shows that users mostly agree with Zemanta ranking, which is important for our purpose. Looking into the example of extracted keywords, it can be seen that there are also some irrelevant keywords. In addition, during the evaluation, the following issues were observed:

- **Content *cold start***. At the beginning of authoring, a number of words should be present for satisfactory results. Otherwise, irrelevant initial keywords are extracted.
- **Semantic relation of words**. Typically, users would test the application by typing a few words to start with, without making any sentence structure or phrases. As Zemanta tries to extract semantic relations from phrases, a text where the words do not make at least a phrase poses a problem for keyword extraction. The influence of this style of writing on keyword extraction should be further evaluated.
- **Unnecessary text markup**. The content submitted to the keyword extraction service contained XML tags, which were internally defining the layout. These were not removed automatically, and thus influenced the keyword extraction.
- **Ambiguity**. For small-size texts, keyword generation was sometimes biased by particular meanings of phrases, as the phrase context could not be determined.

Implementation modifications. Several modifications related to keyword extraction were implemented after the first evaluation:

- **Including the content from previous slides.** To address the cold start issue which occurs when a new slide is started, the content from two previous slides has been included in the keyword extraction, to provide a larger context. As even the completed slides can have a small number of words, this can be very useful. However, a problem can occur if there is a major topic change in adjacent slides.

- **Title emphasis.** To help solving semantic problems, the slide title was marked as *emphasized*, which is an additional Zemanta option to focus the extraction on particular words. Depending on the writing style of the author, this can improve the keyword extraction, but it can also degrade it (e.g. slide title "History", as the history of a technology, could bias the generator towards general human history).
- **Text cleaning.** The text submitted to the keyword extraction service was additionally cleansed of XML tags, as it was not done by Zemanta automatically.

5.3 Evaluation 2

The goal of the second evaluation was to analyze the influence of different text scenarios in presentation authoring: including an example, changing the sub-topic of the presentation and writing about a more general topic.

Four users were involved in the evaluation. The process was the same as in the first evaluation: authoring the introductory slides on a topic in the computer science field. To analyze the text scenarios, one user was asked to include a real-world example, while a second user was asked to focus on a specific subtopic in some slides. The third user was writing about a more general topic ("open source"). The fourth user was writing a presentation without a specific scenario. It was expected that the different text scenarios and one more general topic would lower the similarity between the user and Zemanta keyword ranking.

Fig. 7 and Fig. 8 present the evaluation results in the same way as the diagrams in the first initial evaluation. Fig. 7 shows the relation of the user ranking and Zemanta ranking. Fig. 8 shows the averages of user rankings for keywords in the same Zemanta rank. The highest-ranked keyword is ranked on average with 1.75, and the user relevancy ranking average drops as Zemanta ranking lowers, to an average of 5.5, for the fifth keyword.

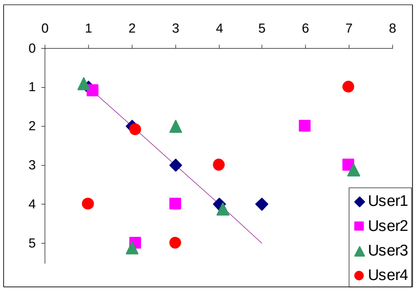

Fig. 7. The relation between user and internal ranking. The X-axis presents Zemanta internal ranks, from 1-8 (1 being the highest-ranked). The Y-axis presents user ranks from 1-5 (1 being the highest-ranked). The actual ranking is marked with a dot of a different type for each user.

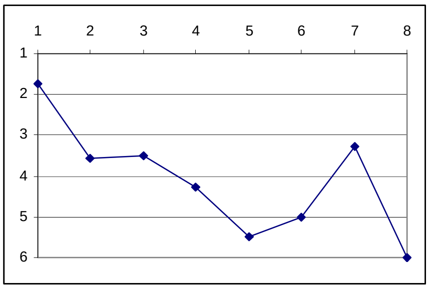

Fig. 8. The average user ranking. The X-axis presents Zemanta internal ranks. The Y-axis presents the average of user rankings for the keywords in a particular Zemanta rank. For this calculation, the keywords not being among the 5 most relevant were given the rank 6.

Some keywords most relevant to users occur in the lower Zemanta ranks (6-8):

- an example from banking for database systems was included, which caused the keywords related to the example (e.g. “bank”) to be extracted (User 2);
- in the presentation about a less specific topic (“open source”), a keyword which was relevant to the user was in the lower Zemanta ranking (User 3);
- in the presentation about HTML, the user was creating a slide specifically for dynamic HTML. As the previous slides were about HTML in general, the keywords were more related to HTML. The most important keyword – “*dynamic HTML*” – was ranked seventh by Zemanta (User 4).

One way to solve these problems is providing a larger context, from the content itself (additional slides) or from the external environment. Another solution is to give users the option not to include the context of previous slides (useful for changing topics) and not to emphasize the slide titles (useful for misleading titles), but this could reduce the application usability as the user needs to manually select these options. Detecting the change of topics can be done based on the slide layout changes, as some authors divide the presentations in subtopics with slides of a particular layout, or by heuristics based on the topic changes per each slide or per slide sets.

5.4 Lessons Learned

The majority of best-ranked keywords in these two evaluations were in the first 5 of the keywords suggested by Zemanta. Due to the specifics of the scenarios, some keywords which users chose as most relevant were in the lower Zemanta ranks.

The users were creating presentation texts for evaluation purposes, not for real presentations. Therefore, some presentations contained very few words, which were not semantically connected. Although some authors prefer to create presentations without many words, the majority of authors still write at least a set of phrases on the slides, which is necessary for obtaining the relevant terms from keyword extraction services.

6 Conclusions and Future Work

The evaluations performed confirm Zemanta as a sound basis for the intended purpose, based on the results and available features such as proposing the keywords - mostly abstractions - which are not present in the text, emphasizing the words to influence the extraction and internal ranking. The five highest-ranked keywords extracted by Zemanta will be used, as the users graded these keywords on average with more than grade 3 (the average of grades 1-5).

Future improvements of keyword extraction include the use of keyword classification schemes to detect similar terms and exploring folksonomies as an additional way to find tags that are often used together. To address the problems observed in various text scenarios, two options will be implemented if the user wants to adapt the keyword list: removing a keyword from the list and simple user rating. If rating is used, Zemanta ranking will be combined with the user rating to form a more relevant keywords list.

Several questions remain: Will extracted keywords be found in metadata entries? Do more relevant keywords in the queries produce more relevant recommendations? What can be done not to omit the relevant content, while using this approach? These questions are certainly important and should be investigated.

Besides the keywords, other research segments not discussed in this paper, such as including context information from the environment, will influence the quality of final recommendations. Therefore, further research will focus on usability of content reuse workflows, extraction of context from the authoring environments or learning management systems and mapping such context to learning object metadata. The proposed solutions will be evaluated using the developed prototype application.

Acknowledgments. This work is supported in part by the Croatian Ministry of Science, Education and Sport, under the research project “Software Engineering in Ubiquitous Computing”. Katrien Verbert is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO).

References

1. Wirski, R., Brownfield, G., Oliver, R.: Exploring SCORM and the national flexible learning toolboxes. Proceedings of the 21st ASCILITE Conference, Perth. (2004).
2. Term Extraction Web Service - YDN, http://developer.yahoo.com/search/content/V1/termExtraction.html.
3. term extraction | fivefilters.org, http://fivefilters.org/term-extraction/.
4. Blog Smarter | Zemanta Ltd., http://www.zemanta.com.
5. Home | OpenCalais, http://www.opencalais.com/.
6. Developer Portal - News - Evri, http://www.evri.com/developer.
7. AlchemyAPI - Transforming Text Into Knowledge, http://www.alchemyapi.com/.
8. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web. 7, 154-165 (2009).
9. Fellbaum, C., others: WordNet: An electronic lexical database. MIT press Cambridge, MA (1998).
10. ODP - Open Directory Project, http://www.dmoz.org/.
11. Entity Extraction & Content API Evaluation « ViewChange Development Blog, http://blog.viewchange.org/2010/05/entity-extraction-content-api-evaluation/.
12. Puzzlepieces – Comparing NLP APIs for Entity Extraction, http://faganm.com/blog/2010/01/02/1009/.
13. Dotsika, F.: Semantic APIs: Scaling up towards the Semantic Web. International Journal of Information Management. 30, 335-342 (2010).
14. Meire, M., Ochoa, X., Duval, E.: Samgi: Automatic metadata generation v2. 0. Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications. p. 1195–1204 (2007).
15. GLOBE | Connecting the World and Unlocking the Deep Web, http://globe-info.org/.
16. Nielsen, J.: Usability engineering at a discount. Proceedings of the third international conference on human-computer interaction on Designing and using human-computer interfaces and knowledge based systems (2nd ed.). (1989).
17. Bosnić, I., Pošćić, A., Ačkar, I., Žibrat, Z., Žagar, M.: Online Collaborative Presentations. Proceedings of the 32nd International Conference on Information Technology Interfaces - ITI 2010. pp. 1-6 , Cavtat/Dubrovnik, Croatia (2010).