Journal of Universal Computer Science, vol. 8, no. 10 (2002), 985-990
submitted: 24/4/02, accepted: 15/10/02, appeared: 28/10/02 © J.UCS

Software for a Multimedia Encyclopedia

Helmut Mülner
(JOANNEUM RESEARCH, Graz, Austria
helmut.muelner@joanneum.at)

Abstract: This paper discusses some experiences with the development of features based on natural language processing for a multimedia encyclopedia.

Keywords: information retrieval, software development, natural language processing
Category: H.3

1 Introduction

There is a large set of requirements for software for a modern multimedia encyclopedia (ME) to be successful in the market:

- It should offer a great looking user interface with a big re-recognition value that attracts a big share of the market.
- There should be an interface to other applications that makes it possible to e.g. look up a word in the encyclopedia while working on a text document.
- The program should react so fast to user request that it is possible for users to look up questions from a TV quiz show faster than the live candidates can answer them.
- The ME should not only provide fast and efficient information retrieval but also invite to casual browsing of the material by providing interesting context.
- Program internal web access makes it possible to access current and extended information and to keep the content up-to-date by providing regular updates.

In a cooperation with the Bibliographisches Institut Mannheim, a team of software developers at the Institute for Hypermedia Systems from JOANNEUM RESEARCH Graz has for several years conducted research to approach this ideal ME step by step. The result of these efforts is a product called "Brockhaus Multimedial" (BMM).

2 Challenges

Besides the usual problems of software development and the special problems of an efficient implementation of a rich set of full-text retrieval functions the German language offers additional challenges: many information retrieval systems cover the problem of inflection suffixes by guessing the inflection by using a rule system and reducing words probabilistically to stems (stemming) and by using only reduced words in indexing and retrieval. In contrast to the English language (Porter algorithm) there are no sufficient satisfactory algorithms for German that treat e.g. the words *Häuser*, *Haus, Staus* and *Laus* in the correct way. A bigger problem are the countless combined words (Komposita) that are written without blanks in between making the German a language of almost unlimited vocabulary. If an interested person wants to look up in a German full-text retrieval system (i.e. a ME) with how many elephants the Carthaginian Hannibal crossed the Alps, she may be unsuccessful because this information may be written in the database this way: “Nach der Kriegserklärung durch Rom zog **Hannibal**, um den römischen Offensivplan zu durchkreuzen, durch die Pyrenäen und Südfrankreich, überschritt mit seinem Heer und den 37 **Kriegselefanten** die verschneiten Alpen und stand Mitte Oktober 218 mit 26,000 Mann in der Poebene.“ (After Rome declared war **Hannibal** tried to thwart the Roman offensive plans by going through the Pyrenees and southern France. He crossed the snow-covered Alps with his army and 37 **war elephants** and arrived on October 218 with 26,000 men in the Po plain.) In this example, traditional text database systems would fail to find the elephants (“Elefanten”) because the indexed term would be “Kriegselefanten”.

3 Solutions

In the version 2003 of BMM we solved such problems by using a so-called **lemmatized index.** This index stores all information necessary to find for every word in the encyclopedia the corresponding base form or the derived forms, if the word is syntactically unambiguous (otherwise all existing base forms are retrieved). This has been made possible by creating the index in cooperation with the IAI (Institute for Applied Informatics) Saarbrücken [IAI] that has a long tradition in analyzing the German language. IAI provides a tool named “**Mpro**“ to analyze the text of the encyclopedia morphologically and syntactically.

Some numbers may illustrate the size of the task:

- The “Premium“ version of the BMM 2002 contains approx. **14.4 million words**, where 11.3 million are in article texts (without media, picture subtitles and other additional materials) consisting of 544,036 distinct “words“ (normalized to the capital letters A-Z).
- There are 11 million alphabetic words i.e. words which are neither numbers nor special symbols consisting of 555,667 distinct words which can be reduced to 523,506 distinct words after normalizing to the capital letters A-Z.
- Using the Mpro-tool produces 466,989 words of the relevant categories (substantives, adjectives and adverbials). Normalizing yields **252,553 lemmas** if we do not consider the grammatical category.
- There are approx. **110,000 encyclopedia articles** with 190,000 key words (words in titles). Part of the BMM is also a version from the year 1906 that contains 82,000 articles.

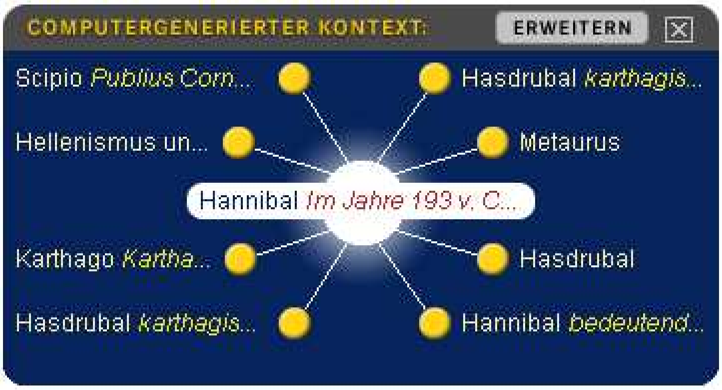

Figure 1: Small version of knowledge web with 8 links that can be viewed together with the article about Hannibal

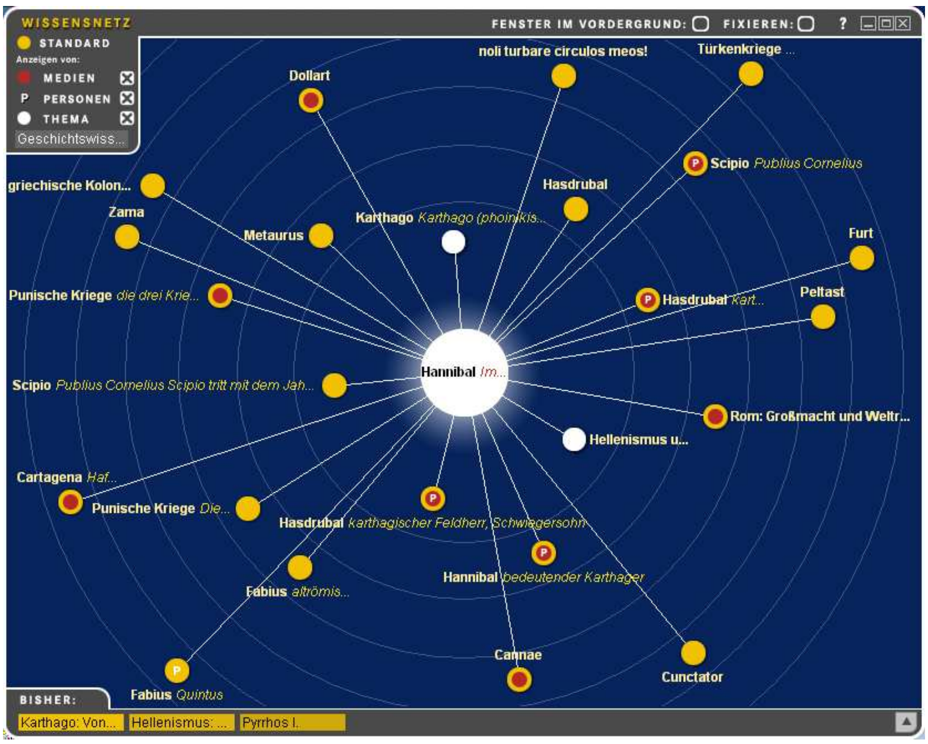

Figure 2: Large version of the same knowledge web with 24 links and some additional information that shows whether a link belongs to the same category and/or is an article about a person and/or contain multimedia information

Another use of the output from the morphological and syntactical analysis is the extraction of the most significant words from each article. These words are used when the user presses a special WWW search button and enables the program to search for additional material to each article by utilizing one of several search engines.

But the most important use for the base forms of words and separated composed words is the generation of the so-called visual knowledge web (VKW) which has been a part of the BMM since 2001. The user can display the VKW optionally directly below an encyclopedia article in a small form (max. 8 entries). A big version with up to 24 entries can be displayed in a pop-up window [Figures 1 and 2].

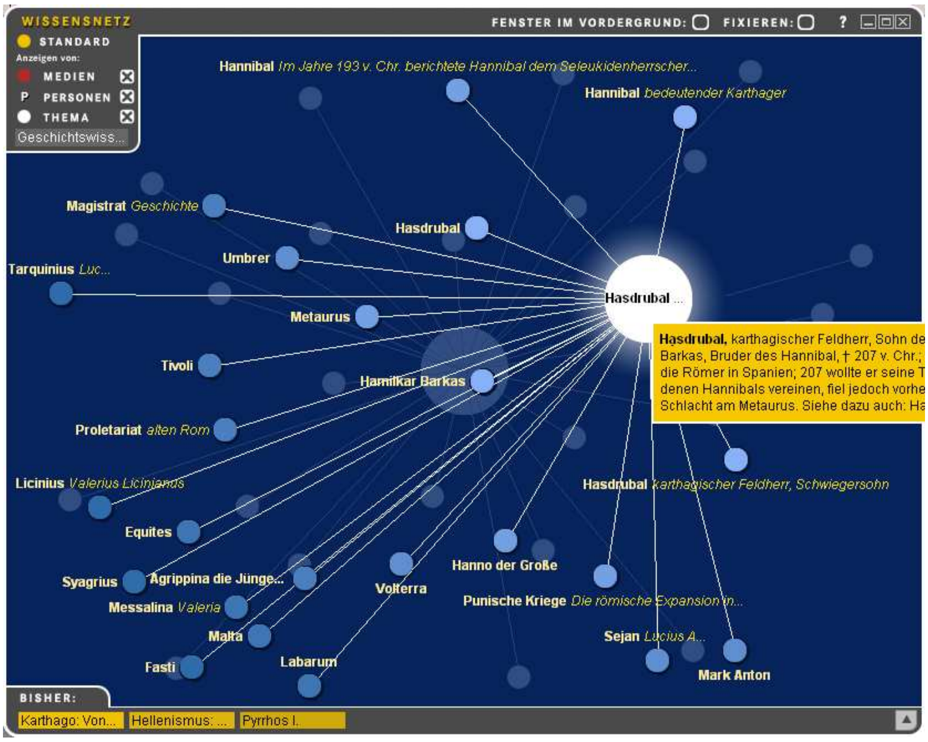

Figure 3: Hovering with the mouse over a link shows a preview of the article and a preview of the corresponding knowledge web

Such knowledge webs are pre-generated for each article and stored in a database that is included in the product. The following section will show some interesting aspects of the generation of these webs.

The similarity of two articles is computed as:
$(w + m + l + c) * T * S$
where

w = word similarity value
m = media value (common media)
l = link value (incoming and outgoing links)
c = classification value from editorial database
T = topic factor (there are 85 main topic areas)
S = source factor (special factor not explained here)

The main criteria for generating a link between two articles in the VKW is word similarity which is well know from information retrieval (e.g. [Frakes, Baeza-Yates 1992]). Our tool allows the selection of different coefficients for computing the similarity: simple Dice, Dice, Jaccard, Cosine (with or without increased weighting of title words). To reduce the large amount of data (see above) only substantives, adjectives and adverbs were used.

The distribution of words for the Brockhaus encyclopedia is different from the distribution of other German text corpora because of a very dense and low redundancy writing style. But also the high frequency words are unusual:

- Stadt (city) (>10,000)
- Land (country)
- Groß (big, large)
- Jahrhundert (century)
- Zeit (time, period)
- Einwohner (inhabitants)
- Lateinisch (Latin)
- Werk (work) (>7,300)

Other difficulties arose from the fact that the distribution of article length is unusual: there are some very long articles (e.g. about the history of Germany, the Austrian and the German constitutions) and a lot of very short articles (e.g. short explanations of figures of speech). There are also articles that consist only of a title and a link to another article (what we called blind link articles). The VKWs for these articles were generated by substituting the VKWs of the link targets (we had fun with circular links!). The low redundancy can be demonstrated by the fact that out of the 466,989 different words 261,143 are singles i.e. they only occur in one article. In other information retrieval projects the irregularities introduced by homonyms (false matches) and synonyms (missed matches) do not play an important role because of redundancy and context. These redundancies are missing in this kind of encyclopedia and therefore often lead to surprising connections. One of our standard examples is Kohl who was a well-known German politician but also is the name of a vegetable (cabbage). Such problems can be partly overcome by using the manually maintained categorization as an additional factor.

The following table shows some strange links we encountered during the history of the project. But experiences at fairs and other demonstrations have shown that most users don’t care too much about strange links. They accept the fact that these links were generated by a "dumb" computer and enjoy the opportunity to browse to unknown knowledge territory.

Article	Strange link in VKW	Reason for link
Mutterschaft	Schnupftabak	Common word: Aufziehen
Cú Chulainn (Nordic hero)	Zinnkies	Common word: Cu
Focusing	James Scott Monmouth	Commond word: Aufdeckung
Ace of base	Auferstehung Christi	Common word: Jona(s)
Graz	Fraueninsel	Wrong manual classification (number flip)

Table 1: Some strange links in knowledge web and why they were generated. (Some are from older versions and do not occur in the current edition

4 Conclusions

Despite a lot of obstacles most of the knowledge webs are a valuable tool for embedding lexicon articles into a context and invite the user to browse similar or related material.

It is also an interesting task for the software developer to think of and incorporate further improvements. To name just a few ideas: synonyms, ontology, semantic web etc.

References

[Frakes, Baeza-Yates 1992] William B. Frakes, Rocardo Baeza-Yates (ed.): "Information retrieval: data structures and algorithms", New Jersey 1992

[IAI] http://www.iai.uni-sb.de/home.html