ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-901

CAKES: Cross-lingual Wikipedia Knowledge Enrichment and Summarization

Valeria Fionda[1] and **Giuseppe Pirró** [1]

Abstract. Wikipedia is a huge source of multilingual knowledge curated by human contributors. Wiki articles are independently written in the various languages and may cover different perspectives about a given subject. The aim of this paper is to exploit Wikipedia multilingual information for knowledge enrichment and summarization. Investigating the link structure of a Wiki article in a source language and comparing it with the structure of articles about the same subject written in other languages gives insights about the body of knowledge shared among languages. This investigation is also useful to identify knowledge perspectives not covered in the source language but covered in other languages. We implemented these ideas in CAKES, which: i) exploits Wikipedia information on the fly without requiring any data preprocessing; ii) enables to specify the set of languages to be considered and; iii) ranks subjects interesting for a given article on the basis of their popularity among languages.

1 INTRODUCTION

Wikipedia is a joint effort of volunteers that contribute to write in a cooperative way Wiki articles. A Wiki article (or page) focuses and describes in detail a single subject. Articles about the same subject are independently written in a variety of languages and therefore may have different length and cover different perspectives. For instance, the Wiki page about the *State of New York* in English contains a much richer description of the history of the state than its corresponding page in German. For instance, the *War of 1812* is mentioned in English but not in German. Wiki articles are internally structured like traditional Web pages and include links to other Wiki articles. For instance, when mentioning the *War of 1812* in the English page about *State of New York* there is a link to the Wiki page about the *War of 1812*. Interestingly, for a given Wiki article the links to the corresponding pages in other languages are also available.

An abstraction of a Wiki article is represented in Fig. 1 where together with the title of the article (e.g., New York (en)) there are reported the titles of other Wiki articles related by some hidden (i.e., embedded in the plain text) *semantic relations*. For instance, the link to the *War of 1812* in English is accompanied by the text "*... a War of 1812 era fort located in what is today Battery Park...*". The representation in Fig. 1 suggests that a source Wiki article can be *summarized* by a set of subjects identified by the outgoing links from the source toward other articles. On one hand, it can be noted that some subjects are shared among different languages. For instance, in both the English and German Wiki page about the *State of New York* there is a link pointing to the Wiki page of *Andrew Cuomo*, the governor of the state. On the other hand, aspects covered in a language may be missing in another. While the page in German has a link to the page about the *Seneca Lake*, the English one does not contain such a link.

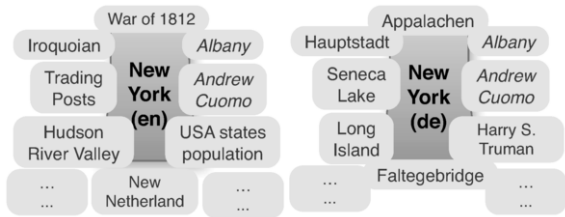

Figure 1. Wiki articles in different languages and some outgoing links.

The aim of this paper is to leverage the Wikipedia link structure and multilingual information for two main purpose. The first concerns knowledge enrichment of articles in a source language with suggestions about related subjects mentioned in other languages. In our previous example, the Wiki page in English could be enriched by putting information about the *Seneca Lake*. The second one is to identify *core* knowledge about a given subject shared among different languages. In our previous example, core knowledge about the *State of New York* includes *Albany* and *Andrew Cuomo* appearing both in the English and German pages. We implemented these ideas in the CAKES (Cross-lingual Wikipedia Knowledge Enrichment and Summarization) system. CAKES exploits information obtained from Wikipedia on the fly without any data preprocessing; enables to specify the set of languages to be considered and ranks subjects interesting for a given Wiki article on the basis of their popularity among languages. CAKES can support the editing of Wikipedia pages.

Related Work. Filatova [2] considered multilingual information overlap for summarization. Overlap is assessed by computing relatedness between sentences appearing in the Wiki articles after translating them in English. CAKES has a different departure point; it exploits the link structure of an article and multilingual information to perform core knowledge identification and knowledge enrichment via link suggestion. Besides, CAKES does not need either sentence translation or relatedness computation. Finally, it exploits online data without any pre-computation. The problem of aligning Wikipedia infoboxes by exploiting multilingual information has been recently discussed [1, 3]. Note that Wikipedia infoboxes represent only a small number of facts about a given subject. CAKES deals with the multilingual link structure of whole articles to face different problems, that is, knowledge enrichment and summarization. When focusing only on infoboxes, CAKES can be used to enrich them by suggesting facts not included in a source language but present in infoboxes in other languages. In our previous example, the infobox in the page in German about the *State of New York* can be enriched with information about the *U.S. senators*. In fact while this piece of information is contained in the page in English it is missing in the page in German.

[1] Free University of Bolzano-Bozen, Piazza Domenicani 3, Bolzano, Italy, email: {fionda,pirro}@inf.unibz.it

2 CAKES: Approach and Evaluation

CAKES exploits the link structure of a Wiki article in a source language and the structure of its possible corresponding articles in a set of *k-1* target languages. We define a *k-partite* graph model where each partition corresponds to a Wiki page in one of the *k* languages as shown in Fig. 2. Here, direct arrows represent the fact that for the language from which the arrow originates there exists the corresponding article in the language where the arrow is directed. Core knowledge is identified by investigating links shared among all or some of the considered languages. Knowledge enrichment is achieved by identifying and suggesting missing links in the article in the source language.

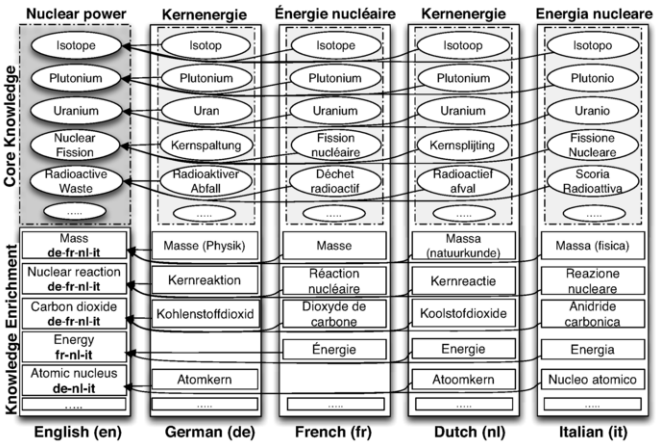

Figure 2. CAKES: an example.

We summarize the main functioning principles of CAKES through the example reported in Fig. 2 by considering English as the source language and *Nuclear power* as the subject. We considered the 4 most prolific languages in terms of Wiki articles. An excerpt of the core knowledge shared by the 5 languages is reported in the top left corner of Fig. 2. This includes subjects such as *Isotope*, *Plutonium* and *Nuclear Fission*, which represent a reasonable summary about *Nuclear power*. The bottom left part of Fig. 2 includes some links to subjects relevant for *Nuclear power*, which are not included in the English page although the corresponding Wiki pages exist in English. For instance, a link to the Wiki page of *Nuclear reaction* is present in the pages in German, French, Dutch and Italian. Interestingly, by looking within the content of the English page about *Nuclear power*, we noticed that the text *nuclear reaction* is present (see Section *Use in Space*). Therefore, an immediate improving would be to link this text with the corresponding Wiki page. The same reasoning applies for *Mass*, *Carbon dioxide* and *Energy*.

As for *Atomic nucleus*, it does not appear within the content of the English page but it can be still useful to include this subject in the English Wiki page about *Nuclear power* as it is considered in the pages in German, Dutch and Italian. Note that links suggested for knowledge enrichment are ranked according to the number of languages in which they are used. Besides, these links are suggested by finding their translations in the source language. As an example, while in the page in German about *Nuclear power* (i.e., *Kernenergie*) there is a link to *Atomker*, CAKES suggests directly the corresponding page in English, that is, *Atomic nucleus*. If the translation is not present then CAKES can be configured to suggest the missing subject either in its original language or by translating it with an online translator.

Evaluation. We conducted some experiments by considering Wiki articles about different kinds of energy. Table 1 shows the result for this evaluation where **CS** is the number of pages in the core knowledge, **IL** the number of links for which there exists in the source article a piece of text identical to the subject (i.e., title of Wiki article) suggested for knowledge enrichment but not the hyperlink with the corresponding Wiki article. **E**(x) $x \in \{4, 3, 2, 1\}$ represents the number of subjects shared by x other languages but not present in the source language. Each row reports the results by considering in turn each of the 5 most prolific languages in terms of Wikipedia articles (i.e., en/de/fr/nl/it) as source language.

Table 1. Evaluating CAKES.

Topic	CS	IL	**E**(4)	**E**(3)	**E**(2)	**E**(1)
Nuclear power	9/9/ 9/9/9	27/14/ 10/2/8	4/0/ 0/9/1	8/10/ 24/32/9	50/55/ 66/76/40	258/319/ 350/293/275
Solar energy	5/5/ 5/5/5	14/9/ 4/0/5	2/6/ 1/2/1	6/10/ 5/7/7	18/33/ 14/37/31	177/280/ 220/247/269
Wind power	3/3/ 3/3/3	24/12/ 10/10/6	0/3/ 1/1/1	22/14/ 9/23/4	47/46/ 36/52/33	426/389/ 360/382/280

As it can be noted, the size of the core knowledge varies from 3 for *Wind power* to 9 for *Nuclear power*. This suggests that the *Nuclear power* subject has more aspects shared in the various languages than *Wind power*. As for the **IL** parameter, it tells us that 27 new links to the pages in English about subjects related to *Nuclear power* are ready to be included in the sense that the corresponding text already mentions the subject suggested by CAKES but the hyperlink is not present. For instance, the Wiki page about *Nuclear Reaction* suggested by CAKES can be linked to the following piece of text appearing in the *Use in Space* section in the English article about *Nuclear power*: *"...In addition, about 3% of it is fission products from* ***nuclear reactions****..."*. A similar reasoning applies for the page in German (i.e., *Kernenergie*), with 14 new links, in French with 10, in Dutch with 2 and in Italian with 8. A relatively high value for the parameter **IL** is related to the English page about *Wind power*, with 24 links.

Digging deep into link suggestion, we have that 4 languages mention 4 subjects that are not mentioned by the English page about *Nuclear power* (first digit in column **E**(4)). The highest number of suggestion is given for the page in Dutch with 9 new subjects (fourth digit in column **E**(4)). Similarly, 3 languages mention 8 subjects not mentioned in English, 10 subjects not mentioned in Deutch and so forth (see column **E**(3)). Obviously, the number of new subjects suggested increases as the number of languages in which these are used decreases, reaching the maximum of 426 new suggestions for the Wiki page in English about *Wind power* (first digit column **E**(1)). The average time to compute knowledge enrichment and summarization for each subject was of about 3 minutes.

3 Concluding Remarks

CAKES leverages multilingual information in Wikipedia for knowledge enrichment and summarization. Given an article a in a source language l (i.e., a^l) and other reference languages $\{l_1, ..l_{k-1}\}$ CAKES helps in enriching a^l by suggesting links to other Wiki articles $a^l_2, ...a^l_m$ included for the same subject in other languages but not in the source language. The set of subjects shared among a certain number of languages represents a meaningful body of knowledge about a particular topic. Integrating CAKES with Wikipedia for supporting users when modifying or creating Wiki articles is our main direction for future work.

REFERENCES

[1] E. Adar, M. Skinner, and D. S. Weld, 'Information arbitrage across multilingual Wikipedia', in *WSDM*, (2009).

[2] E. Filatova, 'Multilingual Wikipedia, Summarization, and Information Trustworthiness', in *Workshop on Inform. Access in a Multil. World*, (2009).

[3] T. H. Nguyen, V. Moreira, H. Nguyen, H. Nguyen, and J. Freire, 'Multilingual Schema Matching for Wikipedia Infoboxes.', *PVLDB*, **5**(2), 133–144, (2011).