KnowNow: A Serendipity-Based Educational Tool for Learning Time-Linked Knowledge

Luigi Di Caro[1], Livio Robaldo[1], and Nicoletta Bersia[2]

[1] University of Turin, Italy
{dicaro,robaldo}@di.unito.it
[2] Telecom Italia, Turin, Italy
nicoletta.bersia@telecomitalia.it

Abstract. In this paper we present the system *KnowNow*, a tool whose aim is to let the users navigate into text corpora through dynamic semantic information networks, created in real-time according to delimited time ranges. In educational scenarios, students are often asked to write short essays on different topics linked by temporal information. This usually involves a combination of several aspects to be evaluated, such as knowledge, imagination, structure and presentation. In the light of this, the introduction of Natural Language Understanding techniques together with cross-topic navigation and visualization tools and considerably help students to retrieve, link, and create well-structured and original contributions, as we demonstrate by using *KnowNow*.

Keywords: Natural Language Understanding, Semantic Search, Education.

1 Introduction

In educational scenarios, it is common to find teachers' requests for short essays that students must elaborate by picking different topics directly connected by temporal constraints. For instance, a student may present a work to combine history, geography, and physics by mentioning the military and political leader *Napoleon Bonaparte* (died in 1821), the *Eyjafjallajokull* volcano in Iceland (that it began to erupt in 1821), and the physician *Elizabeth Blackwell* (born in 1821, who was the first woman to receive a medical degree in the United States).

However, building such *knowledge graph* often results to be "boring" for the following reason: students are usually interested in topics that are likely to be temporally disconnected, so they have to select only one as starting point, and then attach quite unintentional facts that cover other domains.

Nowadays, there is a plenty of freely available resources that can be used for educational purposes, like Wikipedia[1]. Wikipedia is the largest free on-line encyclopedia that includes information of different areas and in different languages that has been already used in this context [3]. Since it contains several historical

[1] www.wikipedia.org

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 440–443, 2014.

facts (and so it is full of temporal information) but also hundreds of other topics, it perfectly fits the above-mentioned context.

In the next section we will illustrate the underlying technology of *KnowNow*, which is a combination of advanced Natural Language Techniques, Data Mining, Human-Interaction models, and Data Visualization schemes. *KnowNow* is the result of the project named *KnowYouAll*, that won a national competition for innovative ideas promoted by Telecom Italia[2].

2 KnowNow

KnowNow is made of different modules: a Time Extractor, a Named Entity Analyzer and Semantic Network builder, a Content Summarizer, and an interactive fish-eye visualization tool.

2.1 Data

As already mentioned in the introduction, we directly used Wikipedia as input corpus. For the demonstration, we randomly selected 10,000 Wikipedia pages, removing metadata information, html tags, links, and Wikipedia-specific texts that are not related to the content. This limit, however, does not reflect technical problems since our syntactic, semantic, and statistical analyses are only applied on small time-delimited document sets.

2.2 Time Extractor

After the cleansing of the input corpus, the system syntactically parses the text using TULE [2], a dependency parser for English and Italian. Since a single document may contain multiple temporal information (related to facts happened in different periods), the system has to extract them in order to build an inverse *temporal map* $< t_k, \{doc_{ids}\} >$ that links time frames[3] t_k with sets of documents $\{doc_{ids}\}$ that contain at least one fact happened in t_k. For recognizing temporal expressions, we used the rule-based techniques proposed in [5].

2.3 Semantic Network

While the syntactic analysis supports the extraction of temporal expressions, the system also includes a semantic analyzer that deals with the identification of *semantic units* for semantic search and access. We define a *semantic unit* as a named entity in the classical NLP task Named Entity Recognition (NER) [4]. A named entity is a type of class of objects, like people, organizations, places, and others. In *KnowNow*, we used the large ontology of semantic information of DBPedia[4], which is a structured version of Wikipedia. It contains several

[2] Working Capital (ed. 2012), `www.workingcapital.telecomitalia.it`

[3] Only timestamps having at least the year and the month are preserved.

[4] `www.dbpedia.org`

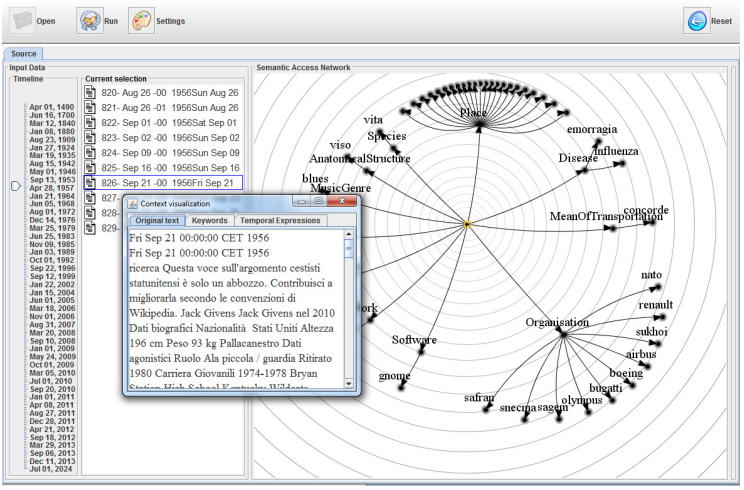

Fig. 1. A screenshot of the *KnowNow*'s main interface. The left panel contains a slider that allows the user to focus on different the time periods automatically extracted from the input corpus. On its right, the interface shows the list of documents that represent the *temporal window* around the selected date. The larger panel on the right shows a fish-eye semantic network calculated in real-time with respect the the left selection. Notice that in this example, an Italian Wikipedia corpus is used, and it is navigable through an English-based semantic network, relying on the Wikipedia interlingual links. Clicking on the button *Run* is then possible to see a hierarchical tag-cloud of the most dominant common terms used in the selected texts, as in [1]. The little window on the top shows the original content of one selected document in the list.

semantic units, organized in a multi-level taxonomy. For example, the instance *Pink Floyd* is associated to the node *Band*, which is a subclass of *Organization*, and so forth. By using these resources, *KnowNow* is able to let the users explore non-English texts with navigable semantic networks written in English. This is done by making use of the interlingual links of the Wikipedia pages, that provides the translation of specific entity names in different languages. This is a powerful feature, since it allows to explore the semantics of texts expressed in several languages by means of an English-based semantic network.

2.4 Content Summarizer

Texts are not only made of semantic units, but they also contain several *common words* that describe the content, and specifically, which named entities are involved in the events, and how. In this case, standard Data Mining techniques applied on texts are useful to allow the users navigate through the content by leveraging on words frequencies and co-occurrences. In particular, *KnowNow* relies on a technique that applies Latent Semantic Analysis on the input texts to construct a navigable tree of *dominant terms* [1].

2.5 Visualization Tool

In *KnowNow*, the information is displayed using different parts of the interface, shown in Figure 1. On the left, a slider permits to observe all the time frames extracted by the time extractor. The user is then able to focus on a particular *temporal window* around the selected date. This parameter, like all the ones mentioned (and not mentioned because of lack of space) in this paper are adjustable through the interface of the system. Once the user selected a temporal window W, *KnowNow* shows a fish-eye tree with all the semantic units found in the texts that have been associated to W, and so that contain facts happened in W. These documents are listed side by side with the slider, and the content can be visualized by clicking on them. The user can do *drag-and-drop* operations on the semantic network to put more visual emphasis on a specific subtree. Then, clicking on a node (or more than one node), *KnowNow* highlights those documents which are related to that relative semantics. Finally, the user may also want to explore the content expressed by common words. In this sense, the *Content Summarizer* extracts a hierarchical tag-cloud of the most dominant terms in the input texts by leveraging on a Latent Semantic Analysis of the term-document matrix. The user can click on the button *Run* to perform such process over the content associated to the current temporal window.

3 Demo Scenario

During the demonstration, we will allow the users to select different time ranges, showing how the semantic network is able to capture and visualize the main semantic information contained in the input texts, in real-time. Then, the tool allows for a number of further interactions, like the selection of specific semantic nodes, the classification and the ranking of the most relevant texts, fish-eye visualization of dominant terms, and the impact of parameters like size of time ranges, amount of data to be displayed, and several others.

References

1. Di Caro, L., Candan, K.S., Sapino, M.L.: Navigating within news collections using tag-flakes. Journal of Visual Languages & Computing 22(2), 120–139 (2011)
2. Lesmo, L.: The Turin University Parser at Evalita 2009. Proceedings of EVALITA 9 (2009)
3. Moy, C.L., Locke, J.R., Coppola, B.P., McNeil, A.J.: Improving science education and understanding through editing wikipedia. vol. 87, pp. 1159–1162. ACS Publications (2010)
4. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
5. Robaldo, L., Caselli, T., Russo, I., Grella, M.: From italian text to timeML document via dependency parsing. In: Gelbukh, A. (ed.) CICLing 2011, Part II. LNCS, vol. 6609, pp. 177–187. Springer, Heidelberg (2011)