Facilitating text understanding for e-learning users

Marco Alfano, Biagio Lenzitti, Giosuè Lo Bosco, Valerio Perticone

Abstract: *Web information on different disciplines (mainly technical ones) is mainly created by the different experts (engineers, scientists, physicians, lawyers, etc.) who use their own 'technical' language. On the other hand, this information is often read by general users who do not have the same skills and vocabularies of the experts and have difficulties to understand and learn it. In order to allow e-learners to use any document available on the web and understand it, it is desirable to have a system that takes a text written with technical terms and automatically translates it in a plain language and provides additional information with the same kind of language. In this work we present the methodology and implementation detalis of such a system and describe the prototype we have developed for the health/medical field.*

Key words *E-learning, Plain Language, Thesaurus, Vocabulary, Dictionary, Health data.*

INTRODUCTION

The World Wide Web can be seen as an enormous container of information that is mostly freely available. This information often presents a didactic structure [5] and can be used by anybody to learn about a specific subject. However, it is not always easy for e-learners to find the information of interest in this chaotic and non-homogeneous world where anybody can freely introduce contents using different languages and communication styles.

Web information on different disciplines (mainly technical ones) is mainly created by the different experts of those disciplines (engineers, scientists, physicians, lawyers, etc.) who use their own language, often very technical. In fact, knowledge, seen as an expanding universe, has over the years moved away the different fields creating huge differences on the languages and communication styles of the produced info. On the other hand, this info is often read by general users who do not have the same skills and vocabularies of the experts. Thus, e-learners have difficulties to find on the web what they are looking for because of the amount of information received and its unfamiliar language.

In order to reach the widest audience, a language that can be understood by anyone should be used in writing web content. This 'plain' language is a kind of language that audience can understand the first time they read it [10]. However, language that is plain to one set of readers may not be plain to others. Simple English Wikipedia is an example of web content developed in plain language [11]. Also, in aerospace industry, a simplified language has been developed to facilitate the readability of technical manuals [12].

From a didactic point of view, the best repositories of e-learning objects, such as MERLOT [6], and OER Commons [9], give the e-learners the possibility to choose the proper modules according to their educational levels. Of course, the teacher has to prepare such modules beforehand by choosing the contents and terms that can be easily understood by the targeted audience and this can be very time consuming. Moreover, the user will only be able to choose among a limited set of those pre-formatted learning modules.

In order to use any document available on the web and understand it, it is desirable to have a system that, given a text written with technical terms, translates it in a plain language and provides additional information with the same kind of language. In this way, the user is not required to increase his/her knowledge to a level where he/she is able to understand the text being examined but knowledge is somehow brought back to the user level so greatly facilitating his/her understanding and learning. This system can be very useful when a user needs precise information in a timely manner and cannot waste time in exploring the huge amount of information provided by the web. Moreover, besides facilitating understanding and learning, it can help the

communication between experts and non-experts (e.g., physicians and patients or scientists and laymen) by providing a sort of two-way translation of terms.

In this work we describe a system that, given any kind of text, finds the technical terms, translates them in a plain language and provides additional information for them (in plain language) greatly facilitating understanding and learning of the generic e-learner. We have developed a prototype of such system for the health/medical field.

The paper is organized as follows. The second chapter describes the basic principles of our methodology that allows to find the technical terms and provide their plain equivalents and additional information. The third chapter describes the implementation details and some practical use of the system we have developed for the health/medical field. The final chapter presents some conclusions and future work.

TRANSLATING WEB CONTENTS FOR UNDERSTANDING AND LEARNING

When an average user wants to learn about a specific topic on the web, generally he/she uses a search engine, finds a potentially relevant web page and examines the text present in the document [3]. As mentioned above, any web page that deals with a specific subject (mainly a technical one) often uses a technical language related to that subject. For a better comprehension of the text, the user may need some external help to understand the technical terms, translate them in a familiar language and find additional information. This external help comes in form of different resources (online or not) such as vocabularies, dictionaries and thesauri (Fig. 1):

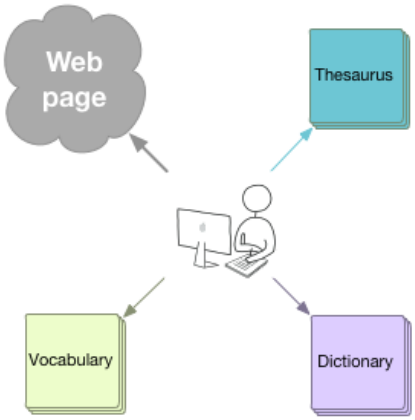

Figure 1. Accessing different resources for text understanding

- **a vocabulary** is a selective list of words and phrases used in a specific field and can be used to find the technical terms in a text;
- **a thesaurus** contains synonyms and antonyms and can be used to find a plain synonym of a technical term;
- **a dictionary** gives information about the meaning of the words and can be used to find additional info on a technical term.

In some cases, a single resource can have multiple functionalities, e.g., it can contain both definitions and synonyms. Table 1 contains some examples of the three types of resources for different subjects.

Table 1. Vocabularies, thesauri and dictionaries for different subjects

Subject	Vocabulary	Thesaurus	Dictionary
Medicine	UMLS http://www.nlm.nih.gov/research/umls/	CHV http://www.consumerhealthvocab.org/	WebMD Medical Dictionary http://dictionary.webmd.com/
Mathematics	Math Vocabulary Word List http://www.myvocabulary.com/word-list/math-vocabulary/	Basic math glossary http://www.basic-mathematics.com/basic-math-glossary.html	A Maths Dictionary for Kids http://www.amathsdictionaryforkids.com/dictionary.html
ICT	FOLDOC http://foldoc.org/		Tech Terms Computer Dictionary http://www.techterms.com/
Economy	Essential Economics http://www.economist.com/economics-a-to-z	STW Thesaurus for Economics http://zbw.eu/stw/versions/latest/about	Economic Glossary http://glossary.econguru.com/

The user will usually look at those resources in an unestablished order to find the technical terms, their synonyms and additional information so, ultimately, to understand the whole text and learn on it. However, this approach assumes that the additional resources are readily available and this is not always the case. Moreover, it is a disorganized process, time consuming and can lead to dispersion and ultimately to an information overload that makes more difficult the learning process.

It is then important to develop a process that uses the resources, such as the ones reported in Table 1, in a coherent and efficient way; even better if this process is automatized. The final objective of this process is the translation of the technical terms of this document to plain language and the provision of additional information for increasing understanding and learning. This removes the need to use many resources to interpret a technical text, in particular when the user accesses the contents through a mobile device.

The basic steps of this process can be summarized as follows :

1. Take a document from the web on a specific subject (e.g., engineering, science, or medicine) and find and highlight the technical terms (words or combinations of words) by using the vocabulary of technical terms;
2. Translate the highlighted technical terms to non-technical, or plain, terms with the thesaurus;
3. Finally, provide additional plain information with the dictionary of plain terms.

By doing so, the user will have everything at hand dealing with a single document that contains all the useful information for its understanding and learning.

A CASE STUDY: A SYSTEM FOR HELPING HEALTH CONSUMERS

In the frame of a collaboration with some Italian hospitals for providing advanced tools to health consumers, we have developed a system that automatically finds the medical (technical) terms in a medical document, translates them in plain or 'consumer' terms and provides additional information.

The architecture of the system is shown in Fig. 2. The HIGHLIGHT module takes as input an arbitrary text and, using the vocabulary, underlines all the technical terms in the area of the chosen subject. The MAP module connects each technical term previously found to its equivalent consumer term by using the thesaurus. Notice that

MAP highlights the technical words that have a consumer translation showing their translation with a tooltip instead of replacing the word or inserting an explanation in the text [14]. The DEFINE module provides a list of synonyms of the technical term (including the consumer one, if applicable) and a description retrieved by a consumer dictionary (in a separate frame). This definition will also be processed by the whole system and transformed in an annotated hypertext that highlights the technical terms so to allow the user a deeper analysis and navigation.

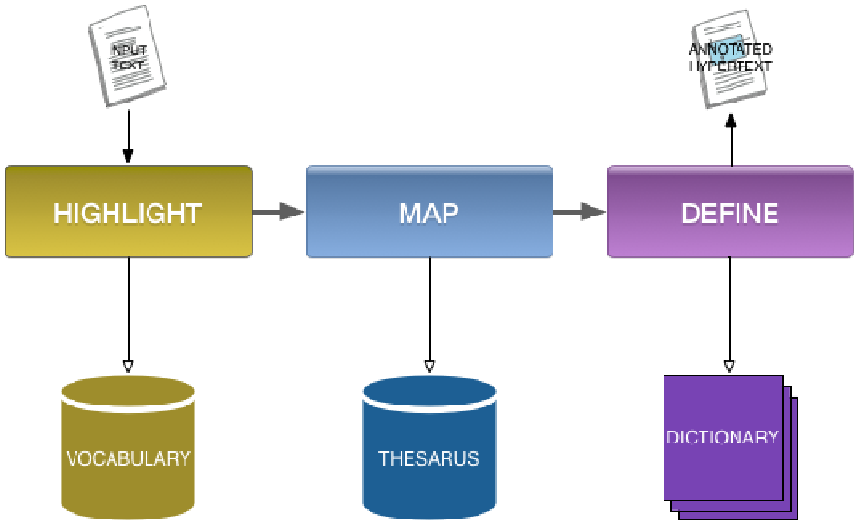

Figure 2. System architecture

Our system uses, as a vocabulary, the "Unified Medical Language System (UMLS)", a large collection of multi-lingual vocabularies that contains information about biomedical and health related concepts [13]. As a thesaurus, we use the "Open Access Collaboratory Consumer Health Vocabulary (OAC-CHV)" that is a relationship file that links commonly used terms to associated medical terminology represented by the UMLS [8]. The mapping from UMLS to OAC-CHV is accomplished by means of the Concept Unique Identifier (CUI) when available [4]. As health dictionaries, we use different Italian web sources such as "Ok Salute" [7] and "Dizionario della Salute" [2].

To test our system and evaluate its effectiveness, we used different types of documents such as medical reports and medical web pages. As an example, Fig. 3 shows an Italian medical report of a magnetic resonance (MRI) where all the technical terms are underlined after being processed through our system. In particular, the word *encefalo* (encephalon in English) is selected and its consumer translation, *cervello* (brain in English), is shown together with its explanation in the below frame.

Further examples of processed texts extracted from the web can be found at the address http://www.cs.unipa.it/H_search/HTM/.

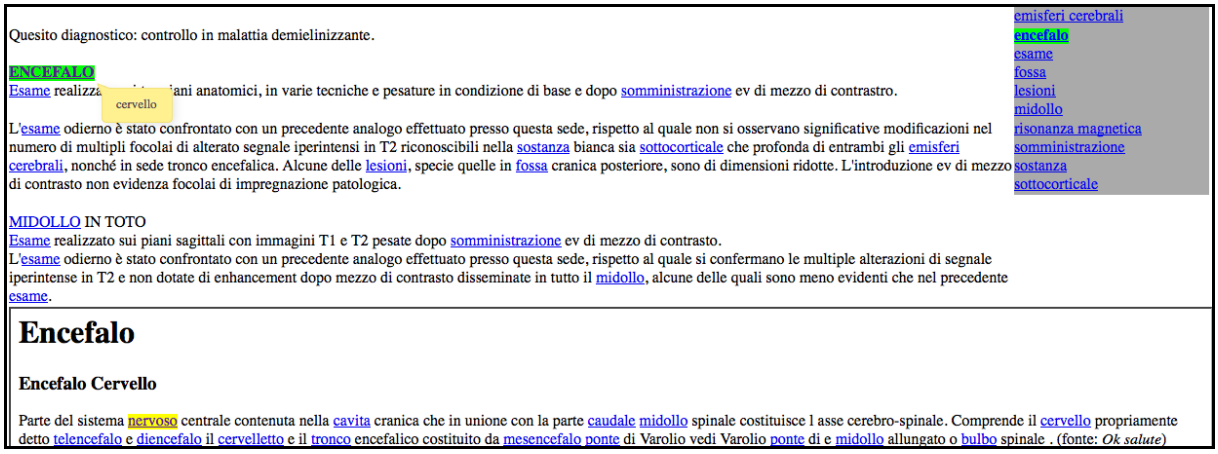

Figure 3. Processed web text and definition frame

Notice that, as said above, our system does not create any change in the original text (this could disorient the user) but only provides a translation (as a tooltip) and additional info (on a separated frame) on request, leaving the user fully in charge of his/her navigation path through the text as it was originally created.

CONCLUSIONS AND FUTURE WORK

In this paper we have presented a system that, given any text, finds the technical terms, translates them in plain language and provides additional information for them with the same kind of language. We have built a prototype of such a system for the health/medical field. It can greatly help e-learners in understanding any web page found on the web even when presenting a very technical content.

Although the prototype is complete and working, there is room for some improvements. As a first step, more experiments are needed in order to verify that the system is able to find all the technical terms by using other vocabularies and thesauri in other fields as the ones shown in Table 1.

A potential extension of the system comes from providing the user with direct access to an external web search engine, either a generic one (i.e. Google, Bing) or a specific one (such Quertle or Pubmed for the health field), for finding further information beside the one already provided. We have developed a system that classifies web pages on the basis of their level of health information and used language [1] and we plan to integrate it in the prototype in order to optimize its retrieval capabilities.

As a further possible extension, the system can be prepended by a CLASSIFY module that, given any arbitrary text, finds the technical terms for each field using different vocabularies and labels the page as belonging to one of different possible categories. This category can then be used to select the proper thesauri and dictionaries in the following steps. We plan to complete the prototype so that it takes as an input the URL of web pages of different disciplines and automatically provides, as output, the same pages with all the underlined technical terms and their definitions.

This work was partially funded by the PON Smart Cities PON04a2_C "SMART HEALTH – CLUSTER OSDH – SMART FSE-STAYWELL" project.

REFERENCES

[1]. Alfano, M., Lenzitti, B., and Lo Bosco, G. "A web search methodology for health consumers," Proc. 15th Int. Conf. on Comp. Sys. and Tech. (CompSysTech'14).

[2]. Dizionario della salute, http://www.corriere.it/salute/dizionario/

[3]. Hölscher, C., and Gerhard S., "Web search behavior of Internet experts and newbies", Computer Networks, vol. 33, no. 1, pp. 337-346, 2000.

[4]. Keselman, A., Smith, C. A., et al., "Consumer health concepts that do not map to the UMLS: where do they fit?", Journ. Am. Med. Inform. Ass., vol. 15, no. 4, pp. 496-505, 2008.

[5]. Koper R., Tattersall C., "Learning Design. A Handbook on Modelling and Delivering Networked education and Training", Berlin, Springer, 2005.

[6]. Multimedia Educational Resource for Learning and Online Teaching (MERLOT), http://www.merlot.org/merlot/index.htm

[7]. Ok Salute, http://www.ok-salute.it/dizionario-medico

[8]. Open Access Collaboratory Consumer Health Vocabulary (OAC-CHV), http://www.consumerhealthvocab.org/

[9]. Open Educational Resources (OER) Commons, https://www.oercommons.org/

[10]. Plain language.gov, http://www.plainlanguage.gov/

[11]. Simple English Wikipedia. https://en.wikipedia.org/wiki/Simple_English_Wikipedia

[12]. Thrush, E. A., "Plain English? A study of plain English vocabulary and international audiences." Technical Communication, vol. 48, no.3, pp. 289-296, 2001.

[13]. Unified Medical Language System (UMLS), http://www.nlm.nih.gov/research/umls/

[14]. Zeng-Treitler, Q., Goryachev, et al. "Making texts in electronic health records comprehensible to consumers: a prototype translator", Proc. AMIA Annual Symposium, vol. 2007, pp. 846-850, 2007.

ABOUT THE AUTHORS

Marco Alfano, PhD, Anghelos Centre on Communication Studies and Dipartimento di Matematica e Informatica, University of Palermo, Palermo, Italy, Phone: +39 091 341791, E-mail: marco.alfano@anghelos.org.

Assist. Prof. Biagio Lenzitti, Dipartimento di Matematica ed Informatica, University of Palermo, Palermo, Italy, Phone: +39 091 23891101, E-mail: biagio.lenzitti@unipa.it.

Assist. Prof. Giosuè Lo Bosco, Dipartimento di Matematica ed Informatica, University of Palermo and I.E.ME.ST. Istituto Euro-mediterraneo di Scienza e Tecnologia, Palermo, Italy, Phone: +39 091 23891075, E-mail: giosue.lobosco@unipa.it.

Valerio Perticone, Dipartimento di Matematica ed Informatica, University of Palermo, Palermo, Italy, Phone: +39 091 23891111, E-mail: valerio@pertico.net.

The paper has been reviewed.