Semiautomatic Knowledge Extraction from Unstructured Sources

J. M. Olivares-Ceja

Abstract

The World Wide Web built on top of Internetis populated with thousands of text-based knowledge sources that must be analyzed by humans to answer knowledge queries or to transform them into machine readable formats, for example, semantic networks or first order logic predicates. A similar process is followed by Knowledge Engineers during knowledge base construction. In this paper we propose a semiautomatic method to obtain knowledge from unstructured texts taken from the Internet. Our method uses techniques from Natural Language processing to map into Fuzzy Semantic Networks. A system that uses the method is under development, it consists of two main parts: knowledge extraction and question answering.

Keywords: Knowledge Extraction, Semantic Web,Semantic Networks.

Introduction

Nowadays Internet is populated with a huge amount of information and knowledge that is being used to satisfy requirements. One problem is that most of that information is in text form. An attempt to provide meaning is the Semantic Web, but to do so, humans still must analyze texts, in order to get the knowledge stored in that sources. We are developing a method to transform text into machine-readable knowledge representation, this is used to answer knowledge queries. Our aim is also to help Knowledge Engineers in knowledge-base construction. An expert user supervises the knowledge that is recorded in the knowledge base and provides fuzzy certainty measures to each fact. We have selected semantic networks [1] [2] as knowledge representation format, because it resembles the sentences structure (noun, verb, object; here verb is the link among noun and object that are represented as nodes).

Our work is an alternate direction on current works of the Semantic Web. Several works are aim to enrich the visualization of the Semantic Web, one of them is VisWeb [3]. Web mining [4] is an active area where it is attempted to solve some problems like: a) finding relevant information (here a problem is the precision), b) extraction of potential useful knowledge (mining tasks), c) personalization of information, d) learning about individual users. Lin [5] focus on answers to web

queries using text annotations and mining in web sources. This paper is organized as follows. In the section 2 we explain how to transform plain text into machinereadable format, here we use a fuzzy semantic network.

Section 3 how the semantic network is updated using the facts obtained in section 2 given as result a fuzzy semantic network. In section 4 comments the system under development. Conclusions and references are given at the end of this paper.

Text Analysis

In our work, text is transformed from plain text into facts of the form:

$$o_1 \; r \; o_2$$

where o1 and o2 are objects obtained from nouns and their adjectives. r is the relationship among o1 and o2, typically formed by verbs and prepositions or verbs and adverbs. The facts are recorded into the knowledge representation schema, in this case, a semantic network1, each fact is stored with a certainty value between -1 and 1.

As we can observe, plain text has different structure than the o1 r o2 structure required to naturally integrating facts into the semantic network. Therefore, we apply different transformations (figure 1) to the plain text for obtaining sets of objects linked by relationships. Some of these transformations are done manually and others have been automated obtaining a semiautomatic method for knowledge extraction.

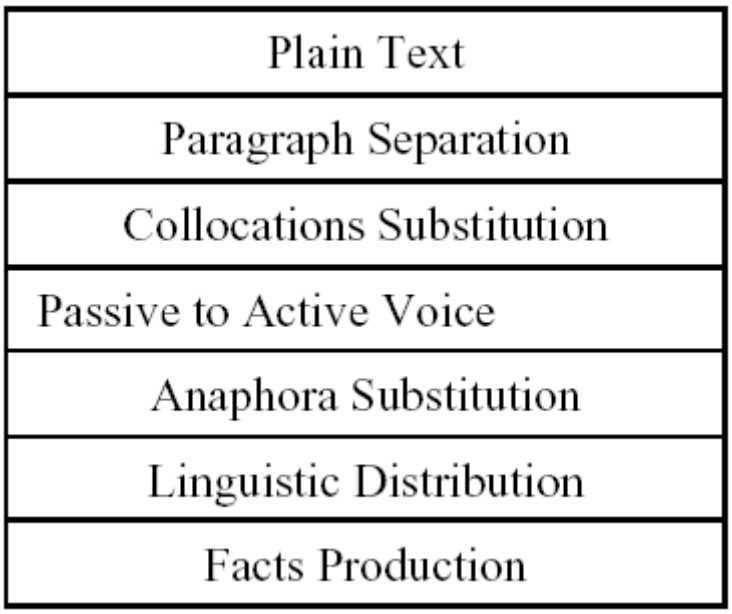

Figure 1. Text Analysis Transformations

Text is separated into paragraphs because it is a way to manage pragmatic knowledge, here pragmatic refers to objects that are related, for example, if a text talks about oranges in one paragraph and other paragraphs talk about lemons, then similar paragraphs are analyzed together due to natural relationships. In our work, we are not verifying language coherence and we presuppose that sentences are correct, it means that non-sense statements are not filtered, like for example “a car is eats meat”.

Let us consider the following text (original text) obtained from the web in Spanish to illustrate the transformations:

Biografía de Benito Juárez, quien nació en San Pablo Guelatao, Oaxaca, en 1806. De extracción indígena, habló solamente zapoteco durante gran parte de su niñez. En la ciudad

de Oaxaca vivió con su hermana Josefa, quien servía en la casa de don Antonio Maza. Estudió en el Seminario de Santa Cruz, único plantel de secundaria que existía en Oaxaca.

Finding Collocations

The first type of transformations applied to a text is collocations because many words that appear together might affect the meaning in the text. Therefore, groups of words like “Association for Computing Machinery” and “Eiffel Tower” are consider as one token instead of four and two respectively. We link the words with an underscore, it is possible to automate this task by using a collocations dictionary. The tokens look like this: Association_for_Computing_Machinery and Eiffel_Tower respectively. In the sample text we obtain:

Biografía de Benito_Juárez, quien nació_en San_Pablo_Guelatao, Oaxaca, en 1806. De extracción_indígena, habló_solamente zapoteco durante gran_parte_de_su_niñez. En la_ciudad_de_Oaxaca vivió_con su_hermana_Josefa, quien servía_en la_casa_de_don_Antonio_Maza. Estudió_en el_Seminario_de_Santa_Cruz, único_plantel de_ secundaria que existía_en Oaxaca.

Passive to Active Voice transformation

It is common in English the use of passive voice where the object appears at the beginning of a sentence. In Spanish subject appear at the beginning because is more often used the active voice. Our method requires that sentences are written in active voice to map the subject relation object into the semantic network.

Therefore, when passive voice structures appear they are changed into active. Our sample text as is written in Spanish does not requires this transformation.

Anaphora Substitution

The second step is finding direct and indirect anaphora and substitutions are made in order to obtain sentences without referential ambiguity on subjects and objects.

Words like that, these, here are substituted with the correct subject or object. In the sample text we substitute *who* and *that*.

Biografía de Benito_Juárez, Benito_Juárez nació_en San_Pablo_Guelatao, Oaxaca, en 1806. De
extracción_indígena, habló_solamente zapoteco durante gran_parte_de_su_niñez. En la_ciudad_de_Oaxaca vivió_con su_hermana_Josefa, su_hermana_Josefa servía_en la_casa_de_don_Antonio_Maza. Estudió_en el_Seminario_de_Santa_Cruz, único_plantel_de_secundaria el_Seminario_de_Santa_Cruz existía_en Oaxaca.

Linguistic Distribution

Once we have changed words that must be considered as one token (collocations) and anaphora substitution, the next step is assigning type to each token using a dictionary, ambiguity is solved asking the user (in future implementations it could be done using contextual information). Nouns and adjectives become nodes in the semantic network, verbs and prepositions forms relations.

We apply linguistic distribution [2] to the sentences to relate subjects with objects, it occurs when it is said something about a subject in a text. In our example we observe that Benito_Juárez is related with San_Pablo_Guelatao, but apparently Oaxaca is isolated. When we apply linguistic distribution we obtain three facts talking about Benito_Juárez

From the sentence and applying linguistic distribution: Benito_Juárez nació_en San_Pablo_Guelatao, Oaxaca, en 1806 We obtain the following facts: Benito_Juárez nació_en San_Pablo_Guelatao (Benito_Juárez was born in San_Pablo_Guelatao) Benito_Juárez nació_en Oaxaca (Benito_Juárez was born in Oaxaca) Benito_Juárez nació_en 1806 (Benito_Juárez was born in 1806)

Semantic Network Structuring

The semantic network structuring involves two operation: storing and retrieval. Storing is done taking the facts obtained during Text Analysis. An expert user assigns the truth value for each fact using values between –1 and 1, 0 represents facts that are false. –1 is used with facts with complete uncertainty, 1 is used in facts with complete certainty. Each fact is of the form o1 and o2 as was obtained previously. If a fact is already in the network the user is asked to use the best truth value. Other options on the certainty value are possible. The certainty value is placed in the arc that links two nodes.

Figure 2 shows one fragment for the semantic network of the sample text. Truth values are showed in ellipsis.

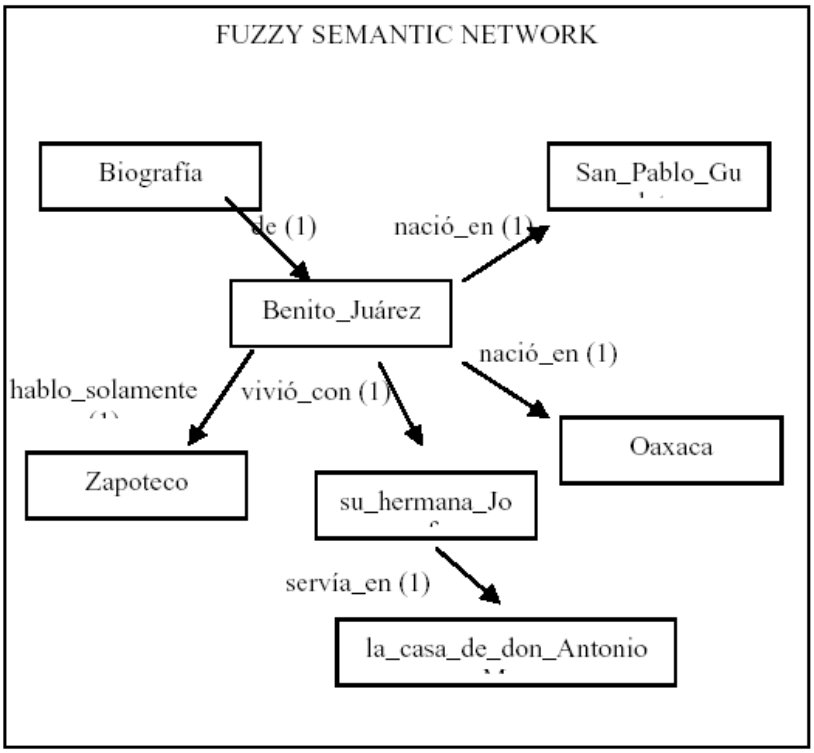

Figure. 2. An example of a fuzzy semantic network

System for Knowledge Extraction

We are building a system to implement the knowledge extraction. The systems consists of two main modules (figure 3). The knowledge acquisition module takes texts from the Internet and applies the operations described in the section 2 in this paper to build the semantic network assigning a truth value for each fact.

We validate the knowledge stored using a query module. The answers are given to the user in textual or graphical form. A navigational tool is also under development to facilitate to the user the navigation in the knowledge.

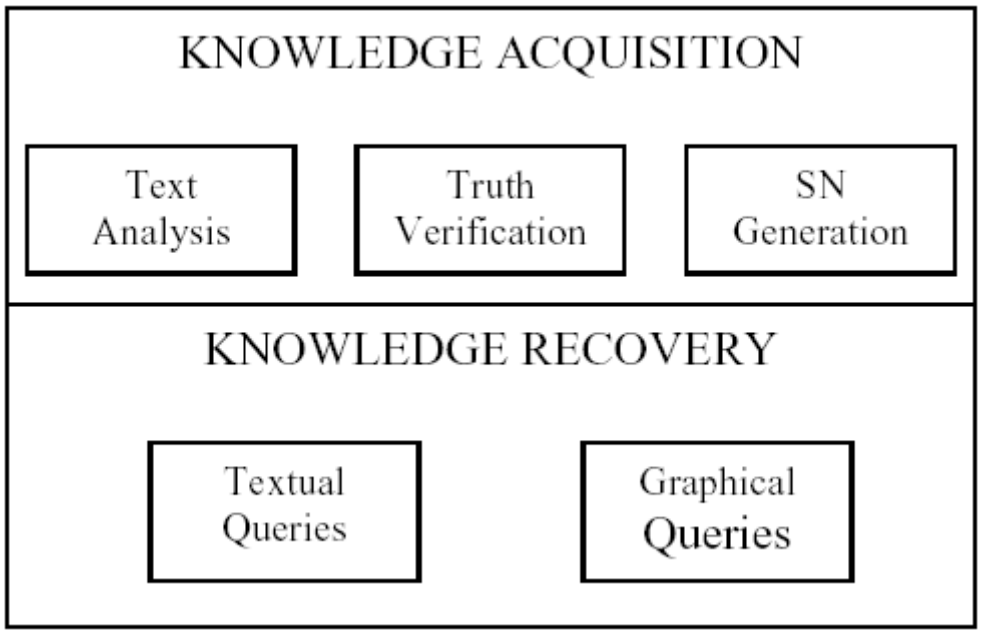

Fig. 3. Main modules of the system

Conclusions

We have presented a method to extract knowledge from text files obtained from the Internet. As we know some information in the Internet is uncertain or invalid, we reflect this using truth values from –1 to 1. A system is under development to automate the process and provide a tool to the users for navigating in the knowledge stored. Many experiments should be done to improve our method and to evaluate the qualification of the knowledge.

References

[1] M. Ross Quillian, Semantic Memory in *Semantic Information Processing*, (Editor Marvin Minsky), MIT 1968

[2] Jesús Manuel Olivares C., Sistema Evolutivo para Representación del Conocimiento (bachelor degree theses), IPN-UPIICSA, clasif. 7.152, Mexico City, abril 1991

[3] Dori Dov, ViSWeb—the Visual Semantic Web: unifying human and machine knowledge representations with Object-Process Methodology, *The VLDB Journal — The International Journal on Very Large Data Bases*, Volume 13 Issue 2 May 2004

[4] Raymond Kosala, Hendrik Blockeel, Web mining research: a survey, *ACM SIGKDD Explorations Newsletter*, Volume 2 Issue 1, June 2000

[5] Jimmy Lin, Boris Katz, Question answering from the web using knowledge annotation and knowledge mining techniques, *Proceedings of the twelfth international conference on Information and knowledge management*, November 2003

[6] Adolfo Guzmán A., Finding the Main Themes in a Spanish Document *en Journal Expert Systems with Applications*, Vol. 14, No. 1/2, 139-148, Jan./Feb. 1998