Multi-Language Semantic Search Engine

Rawad Hammad, Mohammed Hassouna *
Department of Computing
Goldsmiths, University of London
rawadhmd@gmail.com

* *School of Information Systems, Computing and Mathematics, Brunel University*
Mohammed.Hassouna@brunel.ac.uk

***Abstract* – Next generation semantic search engine should offer more than just relevance. It should provide greater insight beyond fixing semantics of simple categories. Fully functional semantic search engine should provide semantic related results for multilingual inquiries. In that light, this paper proposes a combination of Semantic Web, Natural Language Processing and Ontologies to handle multilingual inquiries. In addition, this paper discusses and elaborates on the conceptual design and the implantation of a web based application to manipulate texts in four languages: English, Spanish, Basque, and Arabic.**

***Keyword* — Natural Language Processing, Ontology, Semantic Web Search Engine, WordNet, Watson.**

I. INTRODUCTION

The main goal of this research is to manipulate multi language texts through web-based interface to give some language-related feedback. This research explains the influence of the intimate link established between different technologies like Semantic Web and Natural Language Processing. This paper contains two parts: theoretical and practical. We will give a research background about the topic, and then extend the stage by talking in details about the core technologies and the human languages that suppose to be handled in Multi-Language Semantic Search Engine (MLSSE). Furthermore, we will analyse and evaluate the application and give conclusion about learned lessons and future work.

The main function of our proposed application is to distinguish the language of a specific text and then try to get some classifications and meanings for that text using some applications like WordNet and Watson semantic search engine. The value of this application comes from its ability to combine two areas: Natural Language Processing (NLP), and semantic web. In general NLP can be seen as a way to give a machine the ability of dealing with languages in advanced ways. While semantic web is another technology aiming at assigning meaning to data over web. NLP research has been going on since the late 1940s and Machine translation (MT) was the first application related to natural language processing with the aim of translating text from one language to another. It was built on the idea that the differences among languages reside in vocabularies and the order of the words that constitutes the language. Therefore, MT reduced the whole language concept to what is known as dictionary-lookup to find the translation of a specific word and then try to re-order the words to fit the rules of target language. This reductionist approach shows a shallow understanding of natural languages, and provides evidence on how NLP still need linguistic theorists to get better understanding for lexical ambiguity, like the work done by Chomsky [1] and his syntactic structure theory.

Another point should be taken into account is the context of text since getting the meaning of a specific text is related in a way or another to the area of that text weather it is social, scientific, or political. In other words, languages are dynamic and there is a need to respond to the text beyond the standard semantic of the language.

II. BACKGROUND AND RELATED WORKS

A. NLP:

One of the main concerns of NLP as Jackson and Moulinier [2] mention is for a machine is to be able to analyse or synthesize natural language whether it is spoken or written. According to this background, it is obvious that NLP is a field that combines both computer science and linguistics with the main aim of improving the capabilities of interaction between human and machine. Talking about interaction through a specific language reflects the need of being careful about some factors. First, what is literally said? Second, what is intended? Finally, the relationship between both of them. Also NLP intersects with other fields especially Artificial Intelligence (AI), Computational Linguistics (CL), and Speech Processing or Recognition, this intersection gives an indication to the need of using different techniques in NLP applications. As it is clear from this introduction, there is a huge gulf between computer and human. NLP believes that gulf can be, at least partially, crossed and it intends to take us on this journey. This gulf is based on the vast difference between "natural language" which refers to human language (i.e. English, French, etc) and computer language like Python, C++, etc. We find a room here to draw a difference between NLP and Natural Language Understanding (NLU). The main goal of NLP is to achieve human- like language processing. In the early days of AI, NLP was referred as Natural Language Understanding (NLU) but today and after a series of successful projects in NLP we can say that the goal of NLP is true NLU or true understanding for human languages. That goal has not yet been completely accomplished since NLU has more advanced features than NLP. NLP applications made some accomplishments in paraphrasing an input text and answering questions about the content of specific text, while NLU should feature more complex faculties like drawing inferences from the text itself, and that's why NLU is still the goal of NLP.

On the other side, one of the major issues in NLP is the text ambiguity. In general, human communications usually contain different aspect of ambiguity; therefore there is a need to measure this ambiguity first, then to propose a mechanism to

deal with this phenomenon. Berry et al [3] proposed two- step approach to identify ambiguities in natural language but before going through this point, what kind of ambiguity? Generally, ambiguity refers to the possibility of holding more than one meaning for a specific text or to be uncertain about specific meaning for text. To be more specific here, there are different kinds of ambiguity like lexical, syntactic, semantic, etc; it is not the main part of our project to discuss this point in more details but we will talk later on some details of semantic analysis. Back to Berry et al approach to identifying ambiguity in natural language. The purpose of the first step is identifying the ambiguous sentences based on applying a set of ambiguity measurement steps.

While as the task of the second step is to show what specifically is potentially ambiguous about each potentially ambiguous sentence. Since the ambiguity is a crucial issue in NLP, a lot of researches done in this area but unfortunately results were largely modest since the source of ambiguity is wide. The simplest example on this is the word polysemy (i.e. how it is possible for automated application to decide the right meaning of 'bank').

The importance of ambiguity comes from the fact that disambiguation is a primary requirement for NLP and NLU as well, and it is clear that disambiguation needs first and foremost to identify ambiguity. Claudio et al [4] presented an effective approach for word sense disambiguation. They used a combination of basic kernel functions to independently estimate syntagmatic and domain similarity, building a set of word-expert classifiers that share a common domain model acquired from a large corpus of unlabeled data.

Additionally, there are other important points that cannot be discussed here in details like context sensitivity and the ability of automated applications to classify and categorize the text in order to get more semantic outputs.

Natural language processing applications has a wide range of applications, Goyal and Lehal [7] presents web based machine translation system from Hindi to Punjabi. The shared point with our project is about translation. However, we have a tiny strategy for translation but our project includes four different languages which totally offer different features. Goyal and Lehal works on both Hindi and Punjabi languages which are similar since their originated from one language. This reflects their similarity in syntax and other language components. They use direct algorithm for doing the task which includes lexicon based translation, but in the same way they use machine learning techniques to improve system behaviour after awhile in addition to find a way to solve word sense disambiguation.

On the other side, Calsavara and Schmidt [8] show the efficiency of using semantic search engine in the area of e-business applications. They propose a semantic search engine that is able to stores semantic information about Web resources and to solve complex queries. They reveal that how semantic search engine can be used in order to let customers get the best customizable information they need, weather this information relates to product, services, seller, etc.

NLP has different analytical levels with different outputs for each one, levels are divided into: Phonology, Morphology, Lexical, Syntactic, Semantic, and Discourse. Due to time, we are going to shed lights on some specific levels in order to get a grip on what we are doing in this piece of work.

1) Morphological level: In order to understand this level of NLP, we have to mention to the concept of morpheme, which is the smallest units of meaning. For example, the word "unlikely" can be morphologically analyzed as shown in figure 1.

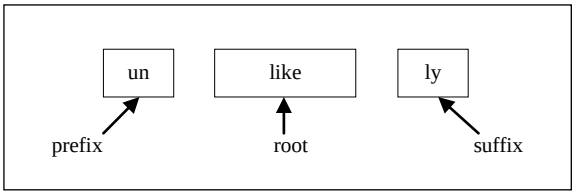

Figure 1: Morphological Analysis of "unlikely"

This level depends on decomposing the words into its components. The significant importance of this step starts from the idea that humans decompose unknown word into its constituent morphemes in order to understand its meaning. In the same way NLP application can recognize the meaning of each morpheme in order to gain and represent the meaning for specific word. Moreover, in some cases these morphemes hold further indications like the suffix -ed when it is added to a verb NLP application knows that the action happened in the past and so on. This level of NLP is used in our application, and it is the first step to do with the word after taking it from the text.

2) Syntactic level: The goal of this level is to reveal the grammatical structure of the sentence via analysing the words in that sentence; NLP should have both grammar and parser to do this level. The output of this level is a representation that shows the structural dependency relationships between the words of the sentence. Indeed, this is a very useful approach since syntax conveys meaning in many languages because order and dependency contribute to the meaning. For example "Mark hits Denis" has a different meaning from "Denis hits Mark", both sentences differ only in the order of their constituent words.

3) Semantic level: Although all levels are important to determine the meaning, but this level is the most important since it uses valuable techniques like the semantic disambiguation of words. In other words, if you have a word holding more than one meaning, then you need to check with the rest of the sentence in order to get its accurate meaning. Also there are other alternatives related to the domain of text or frequency of specific words/parts. This level has been taken into considerations in the intended application, and therefore we used some techniques to find useful clues via semantic web methodologies.

4) Discourse Level: Unlike previous levels of analysis (specially semantic and syntactic) this level works on different units. Previous levels are concerned with the sentence as a unit of handling texts, but sentences here are not the main target since the output of this level should be related to the properties of the text.

Fernando [5] made a good contribution in this area when he presented a modal logic for translating a sequence of English sentences to a sequence of logical forms, characterized by Kripke models with points formed from input/output sequences, and valuations determined by entailment relations.

Output of this level can influence the inference in different ways, for example, if you have a document about solving mathematical equations, then this level is expected to divide this document to introduction to the problem, methods of equation solving, examples, etc. On the other side, MacCartney and Manning [6] proposed a mechanism for inference grounds on what they named as natural logic. They used this term to refer to inferences based on lexical and syntactic properties without paying too

much attention to semantic interpretation.

B. Semantic Web:

Semantic web is the second scaffold of our research since it is used in our application. 'Semantic' as a word means 'meaning', and that's indicate that adding semantic to the web means adding meaning to the web. The historical importance of this approach stems from the large amount of information published on the web. Currently, there are several billion documents on the World Wide Web (www), which are used by hundreds of millions of users. The continuous growth in information published on the web increased the difficulty of finding, organizing, and accessing specific information. Having this difficulty, Tim Berners-Lee, the inventor of semantic web, clarify it as it is an extension of the current Web in which information is given well-defined meaning to enhance computer people cooperation [9]. Furthermore, semantic web can be considered as having the data defined and linked to be used by machines not just for display purposes, but for automation, integration, and reuse across various applications1. That's means the concept of semantic web is going to bring a paradigm shift to the way that www is working. Furthermore, there is abundant evidence on this from many areas especially from the area of knowledge management [10].

In general, meta data used for a long time to describe html pages and give more meaning to web contents. Increased amount of contents makes the use of meta data and keywords less effective and shows the need to get further in this way. For that purpose, semantic web depends on two of W3 Consortium recommendations that are known as Resource Description Framework and Web Ontology Language [11].

Having the idea that how important metadata is to the Semantic Web Resource Description Framework (RDF) is the language used to construct these metadata files. So, it is the basic building block for supporting the semantic web. Point ahead about RDF schema (RDFS) which is a language to create a vocabulary for describing classes, subclasses, and properties of RDF resources. Moreover, RDFS adds semantics to what is known as 'RDF predicates and resources': it defines the meaning of a given term by specifying its properties and what kinds of objects can be the values of these properties. Web Ontology Language (OWL) is one of the most popular languages for creating ontologies today, and it is built on the top of RDFS, but has more advanced feature. Therefore, all RDFS classes and properties can be used when creating an OWL document. OWL and RDFS have the same purpose: to define classes, properties, and their relationships. The difference between both of them is that OWL gives the capability to express much more complex and richer relationships. So it offers constructing tools with enhanced reasoning faculties.

C. TOOLS AND CONCEPTS:

1) WordNet: WordNet began in 1990 by its founder George A. Miller. Different groups of researchers contributed to its development, so it evolved in terms of size and mechanism [12]. WordNet is a lexical database for English language; it groups the words into sets called synsets. In some cases, WordNet appears as a dictionary since it gives the definitions and sample sentences for most of its synsets. It also contains information about morphologically related words. For its relation, there is a distinction between conceptual-semantic relations, which link concepts; and lexical relations, which link individual words. It provides short definitions and records the various semantic relations between these synsets. WordNet's structure makes it a useful tool for computational linguistics and natural language processing. Snow et al [13] demonstrates a model to extend WordNet capabilities via solving the disambiguation of word senses with two-relation assumptions.

In this research, WordNet has been chosen to be used as a lexical database; for other languages rather than English some steps has been taken to translate text from non-English to English and then to manipulate text as what will be described later. Similarly to WordNet, EDBL is a large lexical database used in Basque text-processing tasks for natural language processing purposes. It provides specific levels of language analysis that can be considered as specific features for Basque language [14].

2) Watson: Watson is a web-based gateway for the Semantic Web, which has been guided by the requirements of Semantic Web applications and by lessons learnt from previous experience in the field of semantic web2. Its main aim is to provide access for online ontologies and semantic data. The way to find the required data is to search via keywords and the results will be a list of URIs of semantic documents. However it looks simple and fast, Allocca et al [15] developed a mechanism for finding equivalent ontologies in a large-scale ontology repository, such as Watson. They described a method based on a Knowledge Compilation technique that transform the ontologies in a specific form they named it as Canonical Prenex Conjunctive Normal Form (CPCNF). There are three different ways to use Watson on a specific application, yet all of them return the same data with different mechanisms. In principles Watson is not the only semantic web search engine, there are other similar systems, but Watson is the only one that provides appropriate level of services to be used in semantic web [16].

3) Ontology: In principle, ontology intersects with many areas in computing and it is applied in different technologies from different perspectives. This background brings this concept to different platforms like what has been attributed to W3C as ontology is the terms used to describe and represent an area of knowledge. According to this view, it is easily understood that ontology is domain specific since it represents an area of knowledge. A domain is simply a specific subject area of knowledge, such as photography, education, government, etc. Jovanović et al [17] applied this ontological concept to develop an integrated ontology-based solution for elearning environment with the aim of building customizable environment. Additionally, ontology contains terms and the relationships among these terms. Terms are often called classes or concepts. The relationships between these classes can be expressed by using a hierarchical structure: super classes represent higher-level concepts and sub classes represent finer concepts, and the finer concepts have all the attributes and features that the higher concepts have. Ontology features some advantages such as: first, provides a common understanding about certain key concepts in the domain. Second, provides a way to reuse domain knowledge. Third, makes the domain assumptions explicit. Sari and Ayuningtyas [18] work is a good example on how to exploit these advantages and some others to build an electronic journal citation system. The main feature of this system is to classify information reported in a specific article in electronic journal and relate it with articles cited in it. Also the formal specification of knowledge gives the ground for users to find the relation among articles cited by particular article. It is worth mentioning

that there is a possibility to use a ready-made ontology as an alternative for building your own. Reflecting this background on our project, we build our own simple model of ontologies as a relational database. Ontologies are simple, just four ontologies related to different domains: food, music, accommodation, and places.

In general, we use the text entered by user to search in database to determine the domain of the text as a first step. Then, to provide the user with some links to other ontological resources and more details about specific parts chosen by user. Figure 2 shows the structure of one of those ontologies.

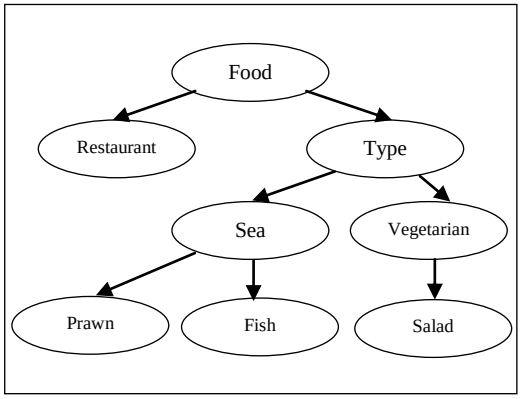

Figure 2: sample of ontology

D. Implications of Multi-languages:

Designing a multi-lingual application requires careful considerations of major issues that mainly correlate with language properties. Our application designed to manipulate four languages: English, Arabic, Spanish, and Basque. Generally, English language developed over a long period of time and many words were constructed based on some forms from Greek and Latin languages. But in comparison with the previous-mentioned languages English is the most obvious one in terms of grammatical and semantic issues.

Second, Spanish is a Latin language. Its writing system is based on Latin alphabets with some complicated or mixed letters like 'k' and 'w' that indicates that those words have been evolved from foreign languages. In addition to this feature, there is another point related to the semantic manipulation of this language since speaker accent can affect and change the meaning of specific words within specific context. Spanish like English is a two-gender language system.

Third, Basque language writing system is based on Latin alphabet, and it is developed over a long period of time with different ways and dialects. That's was the main reason to create a unified language which is almost 30 years old [19]. It follows Spanish in its grammatical syntax, but it is more complicated than English, for instance the verb refers to all grammatical information like: subject, object(s), tense, etc. Furthermore, the order of elements in sentence is free; it can be changed up to the topic. And it is one-gender language, but there is a room to differentiate between male and female. Those are only very simple examples on how such kind of difficult features can affect NLP applications. Finally, Arabic language seems to be the most difficult one, because it differs than all the previous-mentioned languages. It is two-gender language with three style of speaking: singular (one), double (two), plural (three or more). The plural words comes in three forms, for example the plural of student can be written in the following ways: 'Mothakar' طلاب, 'Moanath';طالبات , 'Takseer' طلبة . Arabic language flourished a lot and was affect and affected by other languages especially European languages. Also Arabic has two kinds of sentence structures (noun sentences and verb sentences), and there is a huge range of flexibility of word order. Another distinctive feature for Arabic language is the diacritical mark التشكيل. The effects of diacritical mark are huge, since it can change the sentence meaning, sentence structure (subject - object or object - subject), sentence voice (direct or passive), etc.

Arabic language is complex and strong as well. The root word can generate many words and the ability of developing the language still alive. Unlike English, Arabic has different word ordering system, this feature is reflected heavily on NLP applications specially Arabic-English translation application. This problem is known as phrase re-ordering and was widely discussed by Green et al [20]. In simple words subject and its modifiers should be moved to produce a correct grammatical translation. This is applicable for a wide range of rules in Arabic language.

III. TECHNICAL ANALYSIS

The main title (on the first page) should begin 1-3/ The variety of technologies used in our application has been reflected on the structure of MLSSE; therefore the main components of the System are:

1. Language analysis component
2. Ontology component
3. Semantic links component
4. Word meanings component

A. Algorithm:

Component	Steps
Language Analysis	1. Receive text from user via the main interface. 2. Manipulate text to get the most repeated words. 3. Decide the language of the text with a limited margin for errors. 4. Eliminate unnecessary words. For example "I eat apple", having in mind the purpose of this application, there is no need for the pronoun "I" because this system is designed to decide the area of text not to translate or decide who did the action. Only nouns and adjectives are excluded from elimination. 5. If text language is not English, then call translation procedure
Ontology	6. Our ontological model is a relational database; therefore deciding ontologies require a query using words gotten from the previous step. The result includes ontologies, categories, and keywords in database.
Semantic Links	7. In this stage, there is a need to find more knowledge about text. For this purpose I use OntologyDataSearch class to find semantic links with Watson semantic engine. For that purpose the keywords sent to the function getResults to return a list of semantically related URLs
Word Meaning	8. The choice of getting more details about a specific word is available for the user. 9. Once the user clicks on any word the application will pass it as a parameter to getDefinition function in the class WordNetJAWS.

	10. getDefinition function will use the WordNet path specified within the system 11. A list of possible meanings will be printed on the screen.

Table 1: Algorithm Steps

B. Result Evaluation:

In this part we will describe sample of the system result via detailing the results of each component through testing the application with one English text and another Arabic text.

Text 1 (English Text):

I went with my classmates to a restaurant to get a meal after a working day. My friends wanted to discover new kinds of food. We all ordered different kinds of salad. Some tried salads mixed with apple and many other different fruits. I tried a delicious salad with a lot of cheese, while took a regular meal

Text 2 (Arabic Language):

رغبت في تناول وجبة بعد يوم شاق من العمل فذهبت مع أصدقائي لأحد مطاعم لندن. ولقد طلبنا جميعا أنواعا مختلفة من السلطة ، منها سلطة مخلطة بالفواكه والتفاح. أنا قمت بتجربة سلطة لذيذة ممزوجة بالجبن، بينما حاولت سارة الحصول على طبق تقليلدي من السلطة. من المؤكد أن قرارنا كان صائبا في تجربة أنواع جديدة من الطعام.

Component	Steps
Language Analysis	Text 1: most repeated words are: salad and meal language is English Text 2: most repeated words are:وجبة ، سلطة language is Arabic There is an additional step for text 2, which is translating to English.
Ontology	Text 1: Ontology: food Category: restaurant, vegetables Text keywords: meal, salad, restaurant, food, apple Text 2: Ontology II: food Category: vegetables Text keywords: salad, fruit Ontology II: places Category: city Text keywords: London
Semantic Link	Sample of Semantic related URLs for Text 1: http://www- agentcities.doc.ic.ac.uk/ontology/restaurant.daml http://www.csd.abdn.ac.uk/research/AgentCities/ontologies/restaurant- review.daml http://www.csd.abdn.ac.uk/research/AgentCities/ontologies/restaurant-v4 Sample of Semantic related URLs for Text 2: http://www.archiplanet.org/w/index.php/Special:ExportRDF/Moose_Creek_Administrative_Site?xmlmime=rdf http://www.archiplanet.org/w/index.php/Special:ExportRDF/Burnham%2C_G.A.%2C_House?xmlmime=rdf http://www.archiplanet.org/w/index.php/Special:ExportRDF/Grangeville%2C_Idaho?xmlmime=rdf
Word Meaning	Sample of Text 1 results: The following synsets contain 'vegetarian' or a possible base form of that text: vegetarian: eater of fruits and grains and nuts; someone who eats no meat or fish or (often) any animal products Sample of Text 2 results: The following synsets contain 'City'or a possible base form of that text: city, metropolis, urban center: a large and densely populated urban area; may include several independent administrative districts city: an incorporated administrative district established by state charter city, metropolis: people living in a large densely populated municipality.

Table 2: Results Evaluation

IV. CONCLUSION AND FUTURE WORK

This research is still in its early stages. It combines both semantic web and natural language processing techniques into one application, to apply some functionality in the area of natural language processing. First implication will be drawn from this project is the importance of dealing with a successful NLP applications like Google translator services.

Multilingual feature seems an ambitious idea and should be taken in further researches in the future. Also it is obvious that there is too much effort need to be done in NLP area for Arabic language, the problem of phrase re-ordering. It needs more practical developments related to semantic web. The other valuable parts of this application are the parts that deal with Watson and WordNet. It is a real demonstration on how simple works can do tangible achievements whenever there is a possibility to cooperate with other technologies like open source software. Finally, there is a need to develop more RDF and OWL related functions and show higher performance in statistical points, and to develop our ontologies more and more.

REFERENCES

[1] Chomsky, N. 1957. Syntactic Structures. The Hague: Mouton & Co. Reprinted 1978, Peter Lang Publishing.

[2] Jackson, P., and Moulinier, I. 2002. Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization. John Benjamins Publishing

[3] Berry, D., Kiyavitskaya, N., Zeni, N., Mich, L. 2008. Requirements for tools for ambiguity identification and measurement in natural language requirements specifications.

[4] Giuliano, C., Gliozzo, A., Strapparava, C. 2009. Kernel Methods for Minimally Supervised WSD, Computational Linguistics. Vol. 35, No. 4: 513-528.

[5] Fernando, T.1999, A Modal Logic for Non- Deterministic Discourse

Processing, Journal of Logic, Language, and Information 8: 445-468. Kluwer Academic Publishers: Netherlands.

[6] Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In Proceedings of the 22nd International Conference on Computational Linguistics

[7] Goyal, V. and Lehal, G. 2010 Web Based Hindi to Punjabi Machine Translation System. Journal of Emerging Technologies in Web Intelligence Vol. 2, No. 2

[8] Calsavara, A., Schmidt. 2004. G. Semantic Search Engines. ISSADS 2004, LNCS 3061:145- 157:Springer: Brazil

[9] Yu, L. 2007. Introduction to the Semantic Web and Web Services. Taylor and Francis Group

[10] Davies, J., Fensel, D., Harmelen, F. 2003. Towards The Semantic Web: Ontology-Driven Knowledge Management, John Wiley and Son

[11] Xie, K., Dong, L. 2010. Ontology of general concept for semantic searching. Second International Conference on Computer Modeling and Simulation

[12] Fellbaum, C. 1998. WORDNET: An Electrical lexical database. Massachusetts Institute of Technology

[13] Snow, R., Jurafsky, D., Y. Ng, A.2006, Semantic Taxonomy Induction from Heterogeneous Evidence

[14] Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernández, G., Lersundi, M., 2001, EDBL: a General Lexical Basis for the Automatic Processing of Basque. IRCS Workshop on Linguistic Databases, Pennsylvania

[15] Allocca C., d'Aquin M., Motta E. 2008. Finding Equivalent Ontologies in Watson. Proceedings of the International Semantic Web Conference.

[16] d'Aquin, M., Sabou, M., Motta, E., Angeletou, S., Gridinoc, L., Lopez, V., Zablith, F.What can be done with the Semantic Web? An Overview of Watson-based Applications. Knowledge Media Institute (KMi). The Open University: United Kingdom

[17] Jovanović, J., Gašević, D., Devedžić, V. 2009. TANGRAM for Personalized Learning Using the Semantic Web Technologies. Journal of emerging technologies in web intelligence, Vol. 1, No.1:6-21

[18] Sari, R., Ayuningtyas, N. 2010. Implementation of Web Ontology and Semantic Application for Electronic Journal Citation System. Journal of emerging technologies in web intelligence, Vol. 2, No. 1

[19] Alegria I., Artola X., Sarasola K., Urkia M. (1996). Automatic Morphological Analysis of Basque. Literary and Linguistic Computing 11, N. 4.

[20] Spence Green, Conal Sathi, and Christopher D. Manning. 2009. NP subject detection in verb-initial Arabic clauses. In Proceedings of the Third Workshop on Computational Approaches to Arabic Script-based Languages