SOLVING THE HETEROGENEITY PROBLEM IN E-GOVERNMENT USING N-GRAMS

Cornoiu Sorina

Consiliul Legislativ
sorina@clr.ro

Abstract: e-Government represents an important provocation for the improvements of applications based on architecture of Web services and on semantic dimension of Web based on specific ontology . In one scenario, the Public Administrations are structured on organizational levels (local, regional, national), having an hierarchical organization, a central coordonation, and a industry of services. This requires the usage of Web Services based on ontologies. The services used in e-Government use a diversity of ontologies which interoperate between them. Ontology mapping play an important role when semantic interoperability is needed (Kevorchian *and al*., 2007). To solve the heterogeneity problem, it is necessary to use the similarity concept on the mapping of ontologies.

Keywords: Ontology, Interoperativity problem , e-Government, Ontology mapping, n-grams

1. PRELIMINAIRES

for the improvements of applications based on architecture of Web services and on semantic dimension of Web based on specific ontology (Fernando Ortiz *and al.*, 2006).

In one scenario, the Public Administrations are structured on organizational levels (local, regional, national), having an hierarchical organization, a central coordonation, and a industry of services. This requires the usage of Web Services based on ontologies (Xia Wang *and al.*, 2007).

E-government is summing up a system of technologies for transmitting and processing information of public services. The main idea of e-government is based on the integration of heterogeneous and distributed applications. The architecture orientated to services in general, and to technologies, in particular, are the base for the integration of flexible applications and for the implementation of multiple processes.

Semantic Web supplies the technological infrastructure for building the e-Government intelligent applications, based on Web services and ontologies.

The Semantic Web was proposed by Tim Berners-Lee (Berners-Lee et al., 2001) and it appears to be a new research field, and according to the World Wide Web Consortium (W3C) the Semantic Web is defined as" an extension of current Web where the information is understood by computers and it allows a cooperation between humans and computers". It is based on the idea of having data on the Web defined and linked such that it can be used for more effective discovery, automation, integration, and reuse across various applications.

In a dynamic environment, as e-Government, we can talk about transactions that suppose interactions between different clients and providers. Using different representations and terminologies, it is necessary to have a communication between clients and providers of services.

The e-Government services use ontologies that could be different from the syntactic and semantic point of view. The heterogeneity problem of ontologies is caused by different representations and terminologies used in their construction. For this reason it is necessary an evaluation of similarities between those ontologies.

The evaluation of similarities between these ontologies could be accomplished if the representations of the concepts take in consideration the same components (hierarchies, relations, types) and compare them. If two ontologies have at least one common part, they may be compared.

The Semantic Web is considered to be the infrastructure upon which all intelligent e-Government applications will be built in the near future. Within the objectives of the Semantic Web the ontologies play an important role.

In the field of the Artificial Intelligence, Neches was the first to define an ontology, and he did : ”Ontology defines the basic terms and the relations that include the vocabulary of a specific area, in addition to the rules to combine terms and relations to define extensions to the vocabulary” (Fernando Ortiz *and al.*, 2006). Gruber defines the ontology as : “Ontologies are defined as a formal specification of a shared conceptualization”. Conceptualization refers to an abstract model of some phenomenon . Formal refers to the fact that the ontology should be machine-readable. Shared reflects the notion that the ontology captures consensual knowledge, that is accepted by a group.

We consider the characteristics of relations and types a common components of ontologies. The proposal is to combine the methods that allow to calculate the sintactic and semantic similarities between the concepts representated in different ontologies.

2. ONTOLOGY MAPPING

Ontology mapping is the process of finding correspondence between the concepts of two ontologies (similarities).

Ontology mapping is the complex process, that aims at finding sintactic similarity, semantic similarity based WordNet, and the concept’s description similarity for the two ontologies.

For this process we are using the lexical similarity measure to start the mapping process.

We use the similarity found without an ontology merging an ontology alignment. The ontology alignment consists in establishing different types of mappings between ontologies but keeping the original ones. The ontology merging consists in generating new and unique ontology from original ones.

Let’s consider the e-Government scenario, where the problem is the cause displacement from the Tribunalul judetean 1 to Tribunalul judetean 2. Each Tribunal Judetean has its own format, concepts ,and ontologies. One of the problems of using different ontologies is that there are different representations and terminologies . There is not a formal mapping between ontologies.

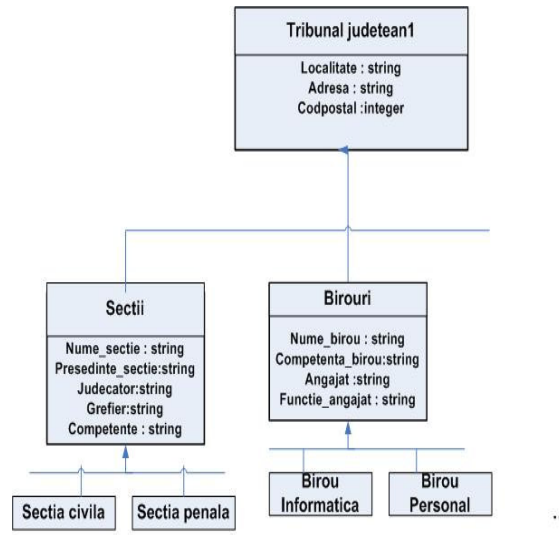

Fig. 1 : (a) Part of the Tribunal judetean1 ontology

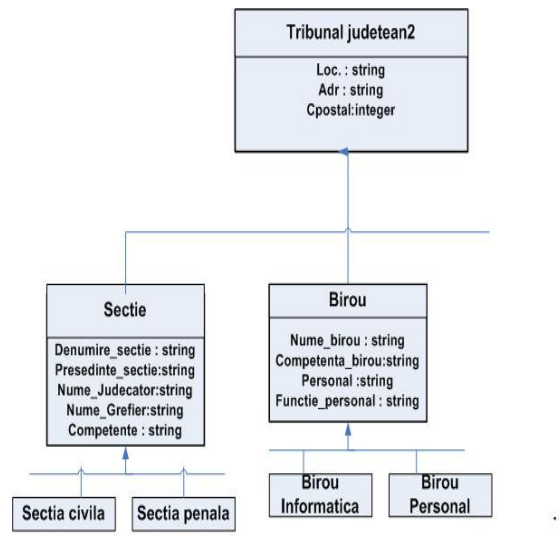

Fig. 1 : (b) Part of the Tribunal judetean2 ontology

2.1. Lexical similarity identification

If two different ontologies are used in the same domain (juridical domain), there is a high probability of the description concepts having similar written or even the same attributes.

To discover the similarity between two strings, we use n-grams algorithm. N-grams takes as input two strings and computes the number of common n-grams between them. N-grams are subsequences of n items from a given sequence. This subsets are called grams and the quantity of characters in each gram is defined by n. The algorithm looks for subsets from one string into another one.

N-grams are used in various areas of statistical natural language processing and genetic sequence analysis (*N-grams-Wikipedia*). N-grams have been use as alternatives to word-based retrieval in a number of systems and to distinguish between documents in different languages in multi-lingual collections and the gauge topical similarity between documents in same language. For this paper we use 3-grams. For example, the word "TEXT" would be composed of following n-grams :

1 letter combination
uni(1)-gram : {T,E,X,T}

2-letter combinations
bi(2)-gram :{ _T, TE, EX, XT, T_}

3-letter combinations
tri(3)-gram : {_TE,TEX,EXT,XT_,T__}

For the identification of lexical similarity between two ontologies, we have classified concept's atributtes according to their values'data types : string, integer and considered the relation **"has-part".**

For example, in Table1, we use n-grams to calculate the similarity between the attribute "Codpostal" and "Cpod" : (integer) , "Localitate" and "Loc" , "Adresa" and "Adr" (string) and the **"has-part"** relation following concepts : Tribunal judetean1 si Tribunal judetean2.

To calculate the similarity between two words we use the following formula:

$$sim = \frac{2n}{n_1 + n_2} \quad (1)$$

where :
n = number of common n-grams
n_1 = number of n-grams in the first word
n_2 = number of n-grams in the second word

n-grams may be 1-grams, 2-grams,3-grams…

Table 1. Compare concept's atributtes usig n-grams

	Atribute Tribunal judetean1	Atribute Tribunal judetean 2	N-grams
Int	Codpostal	Cpostal	0.6
			Max=0.6= sim_{attr1}
String	Localitate	Loc	0,3529
	Adresa	Adr	0,4615
			Max=0,4=sim_{attr2}
has-part	Sectii	Sectie	0.625
	Birouri	Birou	*Max=0.625*

The final results for each attribute types, we must calculate using the following formula (Malucelli *et.al*, 2006) :

$$r_{n\_grams} = \frac{\sum_{1}^{n} \max_i}{n} \quad (2)$$

where :
max_i = maximum for all comparison results that exist for one attribute;
n = number of max_i .

For example, for "integer" we have one attribute , for "string" we have two attributes, and for "has-part" we have two attributes.

The final result is the real number that range between 0 and 1, where 0 signifies no similarity at all and 1 occurs if the words are indentical.
The final similarity $sim_{attr1/attr2}$ take into account comparisons of all attribute types (Malucelli *et.al*, 2006).

$$sim_{atr1/atr2} = \frac{\sum r_{n\_grams}}{n} \quad (3)$$

where :
n = number of different attribute types

For our case we have :
$sim_{attr1/attr2}$= (0.6+0,4072+0.625)/3= **0.5440**

This value of the syntactic similarity for concept's attribute indicates a quite significant similarity.

2.2. WordNet-based semantic similarity

The WordNet (WordNet Search - 3.0) is a lexical database that provide a combination between the traditional lexicographical information and modern computing. WordNet contains more than 118000 different word forms and more than 90000 different word senses and include synonymy (same-name) , antonymy (opposite-name), hyponymy (sub-name),

hypernymy (super-name), meronymy (part-name) and holonymy (whole-name) relations. The lexical database WordNet is particularly suited for similarity measures, since it organizes nouns and verbs into hierarchies of "is-a" relations. Using the CPAN(Comprehensive Perl Archive Network) module, we are able to measure the semantic similarities between words by use of algorithms. There will be used the Leacock and Chodrow algorithms (LCH).

The measure of semantic similarity by using LCH algorithm finds the shortest way between the two concepts, counting up the number of edges between the senses in the "is-a" hierarchy of WordNet (Pedersen, S *et al.*, 2004).

The LCH measure technique requires two word senses as input parameters. The input format is word#pos#sense , where word is a term, pos identifies the type of the word (n for noun, v for verb, a for adjective and r for relation) and sense is a positive integer and represents the meaning of the word in WordNet.

The LCH algorithm (Budanitsky et al., 2001) measures the similarity between two concepts based on the formula:

$$sim(c1, c2) = \log \frac{len(c_1, c_2)}{(2D)} \qquad (4)$$

where :
c1, c2 = concepts
len(c1,c2) = the shortest path between c1 and c2
D = max deep in taxonomy

For example, if we want to calculate the semantic similarity between "Angajat" si "Personal", we use the translation to English of those 2 concepts.

Angajat $\xrightarrow{tr}$ employee (engleza)
Personal $\xrightarrow{tr}$ staff (engleza)

For the "employee" concept, there is one meaning in WordNet and for the "staff" concept there are six different meanings.

Using the LCH algorithm (*WordNet::Similarity web interface)* , we compare each meaning of one concept with each meaning of the other concept. The maximum value of these conceptions is the value that indicates the similarity between those two concepts.

sim_semantica (employee#n#1,staff#n#1) = 1.0726
sim_semantica (employee#n#1,staff#n#2)= 1.0726
sim_semantica (employee#n#1,staff#n#3)= 1.1527
sim_semantica (employee#n#1,staff#n#4)= **1.335**
sim_semantica (employee#n#1,staff#n#5)= 1.1527
sim_semantica (employee#n#1,staff#n#6)= 0.9985

In our example, semantic similarity between "employee,, and "personal" is **1.335.**

2.3. N-grams for description

To measure the similarities between the two concepts, we have to eliminate the words that belong to the class"stop words" (articles , adverbs , prepositions and conjuctions) . Afterwards , we have to build a matrice of n-grams similarity.

According to the DEX dictionary (*DEX online)* , there is the next description of "Personal" and "Angajat "concepts:

Personal= "totalitatea persoanelor apartinand unei institutii, unei intreprinderi"
Angajat = "(persoana) incadrata intr-un loc de munca".

By eliminating the words that bellong to the class"stopwords" , we obtain:

Personal="totalitatea persoanelor apartinand institutii, intreprinderi"
Angajat="persoana incadrata loc munca"

Afterwards, it is building the matrice that contains the n-grams results for the description of "Angajat" and "Personal" concepts (Table 2.).

For n-grams we use the following formula (1) :

$$r_{n\_grams} = \frac{\sum_{1}^{n} \max_i}{n} \qquad (5)$$

where :
max_i = maximum of all comparation results that exist for one attribute type.
n = number of max_i
For our example,
r_{n_grams}= (0.6086+0.1739+0.1) /3 =**0.2941**

2.4. Calculation of the final result

To find the connection between the two ontologies, it has been used multiple calculations methods : the measurement of attributes similarity-using formula (3), the measurement of description by using formula

Tabel 2. The matrice that contains the n-grams results for the description of "Angajat" and "Personal" concepts

	totali-tatea	persoa-nelor	aparti-nand	intre-prindere	institutii
persoana	0,0869	**0,6086**	0	0	0
inca-drata	0,0833	0	0	0,1538	**0,1739**
loc	0	0	0	0	0
munca	**0.1**	0	0	0	0

(5), and the measurement of semantic similarity of concepts by using LCH algorithm- formula (4).

Using only one result it is not sufficient. For this reason there is used a formula that takes in consideration the result of each method.

Based on the number of partial results(rez_{sing}) for each formula (3),(4) and(5), the final formula is (Malucelli *et.al*, 2006):

$$sim = \frac{\sum_{1}^{n} rez_{sing}}{n} \qquad (6)$$

unde :
rez_{sing} = partial result for each formula (3),(4) sau (5)
n = number of partial results.

The accuracy of the methods depends on the quantity of information (attributes and description) contained in the ontologies and if the words can be found in WordNet.

3. CONCLUSION

The current trends in e-Government application call for joined-up services that are simple to use, shaped around and responding to the needs of the citizen.

E-Government is an attractive domain for research . Recently, the emphasis is put on the modeling of Public Administration domain, and the application of the semantic web technologies to integrate of e-Government systems. Some discussion on applying semantic and Web Service technology in the e-Government domain are concluded as follows :
Public Administration is a huge, diverged and distributed environment layered in clearly defined organizational levels. It causes difficulties when applying semantic technologies in a large scale .
This requires the usage of Web Services based on ontologies. We assume and utilize Web Services as the executable application interfaces logically accessible using standard Internet protocols :
WSDL [1] (Web Services Description Language) and SOAP [2] (Simple Object Access Protocol) . Current languages for describing Web Service (WSDL) and their composition on the level of business processes BPEL4WS [3] (Business Process Execution Language for Web Services) lack semantic expressivity that is crucial for capturing service capabilities at abstract levels.

The e-Government services use ontologies that could be different as a syntactic and semantic aspects. The problem of heterogeneous ontologies occurs as a result of using different representations and terminologies. There does not exist a formal mapping between them. For this reason an evaluation of similarities between ontologies is necessary.

The mapping process is regarded as a promise to solve the heterogeneity problem between ontologies since it attempts to find correspondences between semantically related entities that belong to different ontologies.

Similarity evaluation among ontologies may be achieved if their concept's representations share same components. If two ontologies have at least one component, they may be compared.

It takes as input two ontologies, each consisting of a set of components (classes, instances, properties, rules, axioms, etc.) , and determines as output the similarity matching's.

It has been represented a methodology that evaluates lexical and semantic similarities between concepts of different ontologies.

The solution proposed to solve the interoperability problem applies methods from linguistic processing of data. This solution includes the detection of lexical similarities with n-grams algorithm and the usage of LCH in WordNet for semantic similarity.

The lexical measure algorithm compare attributes , relations and concept's description. The attributes are classified according to their data types and a "has-part" relation . The concepts are compared using the Leacock-Chodrow WordNet-based semantic similarity measure algorithm.

The lexical measurement compares the concepts description in natural language.

The measure of semantic similarity by using LCH algorithm finds the shortest way between the two concepts, counting up the number of edges between the senses in the "is-a" hierarchy of WordNet.

[1] http://www.w3.org/TR/wsdl
[2] http://www.w3.org/TR/soap
[3] http://www.ibm.com/developerworks/library/specification/ws-bpel/

REFERENCES

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). *The Semantic Web*. *Scientific American*, **284(5)** :34 – 43

Budanitsky A., Hirst G. (2001) , *Semantic distance inWordNet: An experimental, application oriented evaluation of five measures.* Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics.

Cristian Kevorchian, Sorina Cornoiu(2007).

A Computational Semiotic Approach of Textual Warehouse Linguistic Summary, 6Th Congress of Romanian Mathematicians, July 2007, Bucharest

DEX online., http://dexonline.ro/search.php?cuv=contract

Fernando Ortiz (2006). *EGO Ontology Model : law and regulation approach for E-Government*, ESWC06, Semantic Web for eGovernment Workshop, Budva, Montenegro, 11-14 June, 2006

Andreia Malucelli, Daniel Palzer, Eugénio C. Oliveira(2006). *Ontology-based Services to help solving the heterogeneity problem in e-commerce negotiations*, Electronic Commerce Research and Applications **5(1)**: 29-43 (2006).

N-grams-Wikipedia, the free encyclopedia . http://en.wikipedia.org/wiki/N-gram

T. Pedersen, S. Patwardhan, J. Michelizzi (2004). *Wordnet::Similarity - Measuring the relatedness of concepts*, In Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04).

WordNet Search - 3.0 http://wordnet.princeton.edu/perl/webwn

WordNet::Similarity web interface , http://marimba.d.umn.edu/cgi-bin/similarity.cgi

Xia Wang, Tomas Vitvar, Vassilios Peristeras, Adrian Mocan, Sotirios Goudos and Konstantinos Tarabanis(2007) . *WSMO-PA: Formal Specification of Public Administration Service Model on Semantic Web Service Ontology*, Hawaii International Conference on System Sciences (HICSS), Jan. 3-6, 2007, Waikoloa, Big Island, Hawaii