Semi-automatic Domain Ontology Construction from Spoken Corpus in Tunisian Dialect: Railway Request Information

http://dx.doi.org/10.3991/ijes.v1i1.2925

Jihen Karoui, Marwa Graja, Mohamed Mahdi Boudabous, and Lamia Hadrich Belguith
University of Sfax, Sfax, Tunisia

***Abstract*—In this paper, we present a hybrid method for semi-automatic building of domain ontology from spoken dialogue corpus in Tunisian Dialect for the railway request information domain. The proposed method is based on a statistical method for term and concept extraction and a linguistic method for semantic relation extraction. This method consists of three fundamental phases, namely the corpus construction and treatment, the ontology construction and the ontology evaluation. The proposed method is implemented through the ABDO system to generate the RIO ontology that contains 14 concepts, 25 semantic relations and 387 concepts instances. The generated domain ontology is used to semantically label Tunisian dialect utterances in spoken dialogue.**

***Keywords*—concept, ontology, semantic relation, spoken dialogue, term, Tunisian dialect.**

I. INTRODUCTION

During the last decade, ontologies are largely used in many fields such as: artificial intelligence, information retrieval, semantic Web, Natural Language Processing (NLP), etc. The purpose of ontology development for NLP system is to add a semantic level and then improve its quality. In fact, ontology allows to represent knowledge explicitly. However, the task of ontologies construction is very expensive in terms of time, maintenance and updating which are performed manually. Consequently, the automatic construction starts to emerge in current research works which aims to create ontologies [1]. While the ontology construction is a complex process, several actors are involved in the different stages of this process. For that, it is necessary to define methods or methodologies to assist this ontology construction process. The two most popular methodologies to built ontologies are from scratch [2] and learning methodologies [3].

In this paper, we propose a hybrid method for semi-automatic construction of a domain ontology. Our method belongs to the learning methodologies. In fact, this method is based on a spoken dialogue corpus in Tunisian Dialect for the railway request information domain. The obtained ontology is called "Railway Information Ontology" (RIO) which is used to semantically label Tunisian dialect utterances.

This paper is organized as follows: the second section presents the proposed method for building a domain ontology. The third section presents an overview of the "Assistant for Building Domain Ontology" (ABDO) system which allows to generate the RIO ontology. Conclusion is drawn in section four.

II. PROPOSED METHOD FOR BUILDING DOMAIN ONTOLOGY

In this section, we describe a hybrid method for domain ontology construction from transcribed speech. This method is based on a statistical method for terms and concepts extraction, and a linguistic method for the identification of semantic relations between concepts. The proposed method consists of three fundamental phases, namely corpus construction and processing, ontology construction and ontology evaluation. These phases are illustrated in figure 1.

A. Corpus construction and processing

This phase is crucial because it represents the starting point on which are based the other phases of our method. It is composed of three steps which are corpus construction, corpus treatment and corpus normalization.

1) Corpus construction

Building ontologies from texts requires a step of corpus construction which represents the domain knowledge [4]. Since we are interested in building domain ontology for railway information services in Tunisian dialect, we used the TuDiCoI (Tunisian Dialect Corpus Interlocutor) corpus [5] which is a corpus of spoken dialogue in Tunisian dialect. It gathers a set of dialogues between the staff of the National Company of the Tunisian Railways and the customers who seek information about train schedules, fares, reservation, etc. Given the small size of this corpus, we have increased its size by manual transcription of recorded dialogues based on phonetics to highlight the Tunisian accent in the dialogue. The transcribed corpus consists of 1825 dialogues. These dialogues consist of 6533 customer utterances which represent 21682 words.

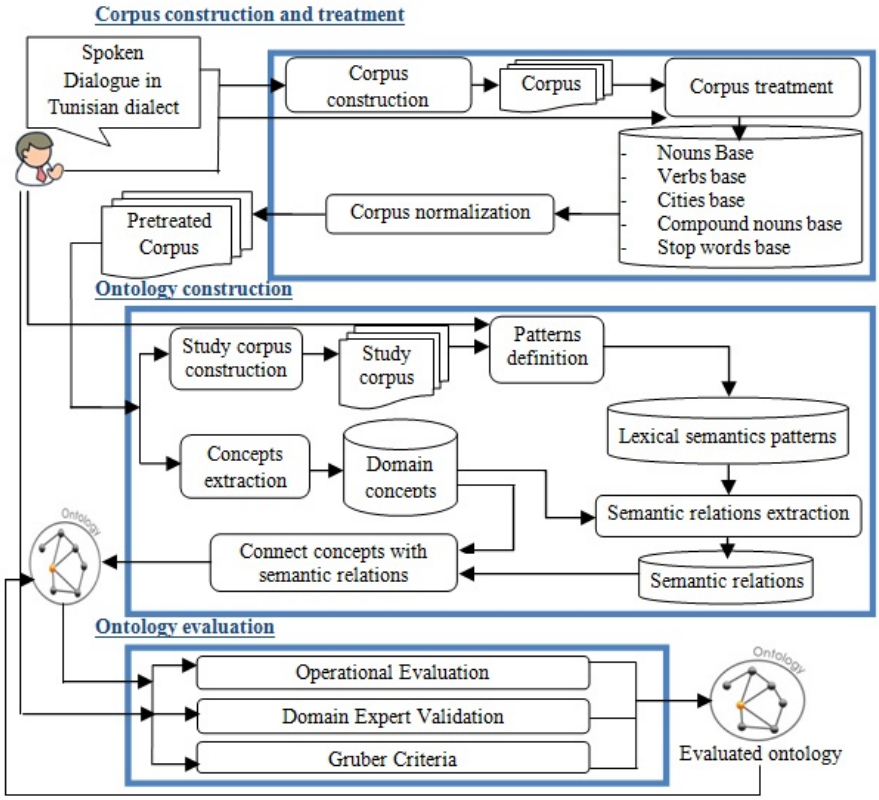


Figure 1. The ontology construction process

2) Corpus treatment

The main idea of this step is to treat the corpus in order to standardize the domain terms. This requires the construction of five lexicons which are nouns base, verbs base, cities base, compound words base and stop words base. This processing step presents a major challenge because we are performing the Tunisian dialect which suffers from lack of tools and linguistic resources such as the morphological, syntactic and semantic analyzers as well as the lack of dictionaries. This step is carried out manually by the domain expert.

3) Corpus normalization

The normalization step is based on all lexicons created in the corpus treatment step. This step consists in orthographic correction, replacing all words by their synonyms, remove stop words, detect compound words and perform morphological analysis of verbs and nouns in order to obtain a new version of the corpus called pretreated corpus.

B. Ontology construction

The ontology construction phase consists of three steps: concept extraction, pattern definition and relations extraction.

1) Concept extraction

This step consists in extracting domain concepts specific to the railway information services from the TuDiCoI corpus based on statistical method. Indeed, domain terms have an important frequency given the important corpus size. Statistical study consists in calculating the frequency of each term of the pretreated corpus. Terms which have an important frequency represent terms of the domain concepts. The only problem is with the domain terms which have a low occurrence frequency. In this case, we have relied on the domain expert to specify the frequency threshold to be fixed. Indeed, the domain expert fixes the concept threshold and each term which has an occurrence value higher than this concept threshold is regarded as a domain concept. To justify this choice, we conducted an empirical study which consists of taking each time a small part of the corpus and calculating the frequency of the fixed concept threshold. The occurrence value of the concept threshold "رتور" increases with the size of the corpus but it still has the lowest occurrence frequency.

The result of this step is a list of domain terms. But it is so important to know that several set of terms represent the same concept. So, it is necessary to gather them and represent them by a single concept. For this, we have relied on the domain expert to define for each set of terms a well defined concept. Table 1 represents some candidate terms and their domain concepts.

TABLE I.
SOME DOMAIN CONCEPTS

Concepts	Concepts translation	Terms	Terms translation
مدينة	City	سوسه / صفاقس/ تونس	Tunis/Sfax/Sousse
قطار	Train	تران	Train
تذكرة	Ticket	تسكره	Ticket

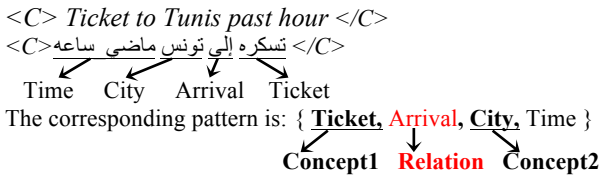
2) Pattern definition

After domain concept extraction, we proposed a linguistic method for the extraction of semantic relations between concepts. In order to define the patterns, we make a linguistic study of the utterance structure. This enabled us to notice that utterance in a limited field respect lexico-semantic patterns in order to extract the semantic domain relations. In fact, this idea consists in using this kind of patterns is introduced by Hearst [6]. Before starting

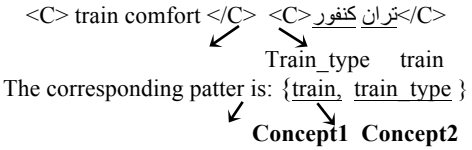
pattern definition, it is necessary to form a study corpus which represents approximately 30% of the pretreated corpus and which is used to manually define patterns. Note that a pattern can be detected from several utterances. In this case we must keep only one definition of this pattern. Defined patterns are validated by the human expert.

3) Relations extraction

This step consists of projecting the obtained patterns on all the corpus utterances to detect semantic relations. On this level, it is necessary to use the results obtained in the concept extraction and pattern definition steps in order to detect semantic relations between domain concepts. We noticed after this projection that the semantic relations are of two types namely the explicit semantic relations and the implicit semantic relations. An explicit relation is detected by a pattern which connects two domain concepts. This is explained by the following example:



The implicit relations are detected by pattern projection and the intervention of the human expert as mentioned in the following example:



The human expert indicates the existence of the semantic relation "train_have_type/ترانمننوع" between the first and the second concept which connects the two concepts "train/ترانا" and "train_type/نوعترانا".

C. Evaluation

The ontology evaluation is a critical phase. This phase can be achieved in deferent ways and with various manners. To evaluate the obtained ontology, we check its conformity with: our needs through an operational evaluation, its conformity with Gruber criteria [7] and human expert evaluation.

1) Operational evaluation

To evaluate the RIO ontology, we exploit it for semantic annotation which aims to attribute semantic labels to transcribed utterances [8]. This method has already been introduced in [5], which used a manually built ontology. Now, we are going to use the obtained ontology to check its fusibility for semantic annotation of the Tunisian dialect in the field of railway information. The semantic annotation based on the RIO ontology consists of attributing semantic label to each word based on ontology concepts. Then, we improve the semantic annotation by using ontology semantic relations in order to take into account the local context of each utterance. Indeed, semantic relations increase the level of understanding by highlighting relations between words in the same utterance. Before the annotation step, each utterance must be analyzed by the same treatments used in the first step to build the RIO ontology. So, we use the same bases created after corpus treatment step. The evaluation of the semantic annotation is given in term of F-measure and concept error rate (CER).

$$F_{mesure}=\frac{2\times precision\times recall}{precision+recall}$$

and

$$CER=\frac{\#incorrect\ concept\ predcition}{\#concept\_ref}$$

TABLE II.
SEMANTIC LABELING RESULTS USING THE RIO ONTOLOGY

#Training words	5710
#Test words	2104
Correct prediction	1213
Incorrect prediction	162
Reference concepts	1718
CER	9,4%
F-measure	78%

Obtained results for semantic annotation are very encouraging since we have improved the F-measure in comparison with result obtained in [5] which obtained an F-measure equals to 66%. In addition to that, semantic annotation based on RIO ontology is very interesting because we have dealt with speech phenomena which are incorporated in the RIO ontology.

2) Gruber criteria and human expert evaluation

After achieving the RIO ontology construction, we evaluate it according to the Gruber criteria which are clarity, extensibility, minimal ontological commitment, coherence and encoding minimum deformation. RIO ontology has provided clarity because all its concepts are represented with terms from the domain in Tunisian dialect. Also, the extensibility of RIO ontology is guaranteed in our method by the increase of the corpus size and by an infinite back cycle to the second phase to update the ontology. Equally, the minimal ontological commitment is ensured by the presence of all terms covering the studied domain in Tunisian dialect, this guaranteed the sharing of knowledge related to this domain. About the coherence criteria, the domain expert helped us to check the RIO ontology coherence. Finally, the encoding minimum deformation criterion is provided by the use of one of used terms as a label for concept. In fact, we try to keep the used terms of the studied domain in Tunisian dialect to nominate concepts. The satisfaction of all these criteria is noted by the human expert who has followed us during the construction process and evaluation of our ontology. Also, the human expert has manually evaluated and validated each step.

III. IMPLEMENTATION

To generate a domain ontology based on the proposed method, we have implemented the ABDO system which is an Assistant for Building Domain Ontology (see figure 2). This system provides interactive interfaces for all steps of the proposed method to facilitate the intervention of the different actors. In addition, this system provides many

interfaces for creating and updating standardization databases, for pattern definition and finally for databases creation into XML files. Also, the system helps the expert to add implicit relationships through automatic detection of not connected concepts. After connecting all concepts, the system generates the domain ontology in OWL language.

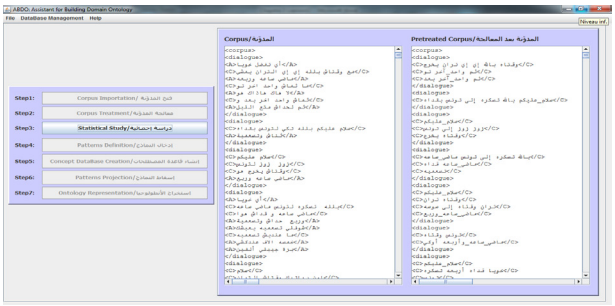


Figure 2. ABDO interface

D. *Presentation of the built ontology*

The domain ontology contains 14 concepts, 25 semantic relations and 387 instances of concepts. It can be visualized with the plugin OntoGraf of protégé 4.1 which allows the graphic visualization of the ontology represented with OWL language. Figure 3 shows the obtained ontology.

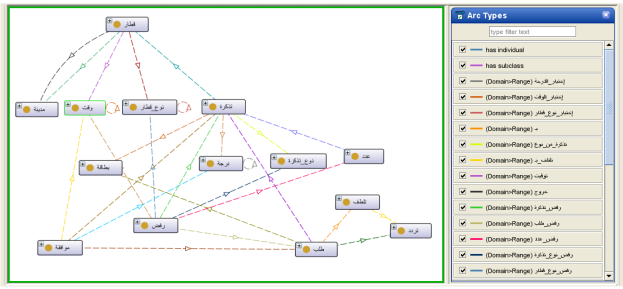


Figure. 3. Visualization of ontology with Protégé

IV. CONCLUSION

In this paper, we proposed a hybrid method for building domain ontology from a spoken dialogue corpus in Tunisian dialect. This method combines a statistical approach for term and concept extraction with a linguistic approach for semantic relation identification. The generated ontology helps us to semantically annotate the Tunisian dialect utterances. As perspective we intend to integrate the RIO ontology into the understanding component of a dialogue system for railway request information to ameliorate semantic interpretation.

REFERENCES

[1] Ding. Y, Foo. S, "Ontology Research and Development: Part 1 – A Review of Ontology Generation", Journal of Information Science 28 (2), pp. 123–136, 2002.

[2] Ben Mustapha. N, Aufaure. MA, Baazaoui-Zghal. H, "Vers une approche de construction de composants ontologiques pour le Web sémantique – synthèse et discussion", In: Troisième atelier sur la « Fouille de données complexes dans un processus d'extraction des connaissances », 6èmes journées francophones « Extraction et Gestion des Connaissances » (EGC), Lille, France, 2006.

[3] Maedche. A, Staab. S, "Ontology Learning for the Semantic Web", IEEE Intelligent Systems & Their Applications 16(2), pp. 72-79, 2001. http://dx.doi.org/10.1109/5254.920602

[4] Condamines. A, "Sémantique et Corpus", Hermès Science Publications, ISBN 2-7462-1055-X, 2005.

[5] Graja. M, Jaoua. M, Hadrich Belguith. L, "Building Ontologies to Understand Spoken Tunisian Dialect", International Journal of Computer Science, Engineering and Applications Vol.1, No.4, 2011. http://dx.doi.org/10.5121/ijcsea.2011.1403

[6] Hearst. MA, "Automatic Acquisition of Hyponyms from Large Text Corpora", In Proceedings of the Fourteenth International Conference on Computational Linguistics, pp. 539-545, Nantes, France,1992.

[7] Raymond. C, Riccardi. G, "Generative and Discriminative Algorithms for spoken language understanding", Proceedings of Interspeech, pp. 1605-1608, Antwerp, Belgium, 2007.

[8] Gruber. TR, "A translation approach to portable ontology specifications", Knowledge Acquisition, 5(2), pp. 199–220, 2-22, 1993.

AUTHORS

Jihen KAROUI received her postgraduate diploma in Computer Science at Faculty of Economics and Management of Sfax (FSEGS) in 2010. She received her Master degree in Computer Science (SINT) at Faculty of Economics and Management of Sfax (FSEGS) in 2013. She is persuading her Master studies with ANLP Research Group within MIRACL Laboratory. Her research interests include Ontology and Tunisian Dialect (jihen.karoui@gmail.com).

Marwa GRAJA received her Engineer degree in Computer Science engineering from the National School of Engineers of Sfax-Tunisia (ENIS) in 2006, and Master degree in Computer Science (NTSID) from the National Engineering School of Sfax-Tunisia (ENIS) in 2008. Actually, she is persuading her PhD studies with ANLP Research Group within MIRACL Laboratory. Her research interests include Ontology, Tunisian Dialect, and Dialogue systems (marwa.graja@fsegs.rnu.tn).

Mohamed Mahdi BOUDABOUS received his postgraduated diploma in Computer Science at Faculty of Economics and Management of Sfax-Tunisia (FSEGS) in 2008. He received his master degree in Computer Science (SINT) Faculty of Economics and Management of Sfax (FSEGS) in 2010. Actually, he is persuading his PhD studies with ANLP Research Group within MIRACL Laboratory. His research interests include ontology, learning methodology (mehdiboudabous@gmail.com).

Lamia HADRICH BELGUITH received her postgraduate diploma in Computer Science at Faculty of Economics and Management of Sfax (FSEGS) in 1990. She received her master degree in Management Information Systems at High School of Management - Tunis (ISG) in 1992. Then, she received PhD degree from the Faculty of Sciences -Tunis, Tunisia, in 1999. From 1999 to 2009, she was an Associate Professor of Computer Science and Management at Faculty of Economic science and Management (FSEGS), University of Sfax. Since 2009, she is a professor of Computer Science and Management at FSEGS and head of Arabic Natural language Research Group (ANLP-RG) of Multimedia, Information systems and Advanced Computing Laboratory (MIRACL). Her research activities have been devoted to several topics: Arabic text analysis, Automatic Abstracting, Question-Answering systems, and Human-machine spoken dialogue systems (l.belguith@fsegs.rnu.tn).

Submitted 19 June 2013. Published as re-submitted by the authors 23 July 2013.