Domain Ontology Construction from a Tunisian Spoken Dialogue Corpus

Jihen Karoui\(^1\), Marwa Graja\(^1\), Mohamed Mahdi Boudabous \(^1\), and Lamia Hadrich Belguith\(^1\)

\(^1\)Miracl Laboratory, Automatic Natural Language Processing Research Group, ANLP-RG, University of Sfax, Tunisia

{jihen.karoui, mehdiboudabous}@gmail.com
{marwa.graja, l.belguith}@fsegs.rnu.tn

Abstract. In this paper, we present a hybrid method for building a 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 corpus construction and treatment, the ontology construction and the ontology evaluation. The proposed method is implemented through the ABDI system to generate the RIO ontology which is used to semantically label Tunisian dialect utterances.

Keywords: Ontology, concept, term, semantic relation, spoken dialogue, Tunisian dialect.

1 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 the extraction of explicit and implicit semantic knowledge from a field.

Today, the main objective in the ontologies design field is to define shareable ontologies. However, the task of building ontologies is very expensive in terms of time, maintenance and updating which are performed manually. Consequently, automatic building starts to emerge in current research works which aims to create ontologies [1]. While the ontology building 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 the process for ontology building. Methodology gathers a method to build ontologies and a set of tools to automatically generate the ontology [2].

The two most popular methodologies to built ontologies are from scratch and training methodologies. The first one consists in building ontologies manually without knowledge [3]. In the literature, the most popular meth-
odologies from scratch are METHONTOLOGY [4] and SENSUS [5]. The second methodology consists in building ontologies from knowledge based on concept training. According to [6], there are many learning approaches for each input types. We distinguish the training approaches from texts, dictionaries [7], knowledge bases [8], semi-structured schemas [9] and from relational diagrams [10].

In this paper, we propose a hybrid method for semi-automatic building of domain ontologies. 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. Our method generates the “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 helps to generate RIO ontology. Conclusion is drawn in section four.

2 Proposed method for building domain ontology

In this section, we propose a semi-automatic method for building domain ontology from dialogue corpus in Tunisian dialect. This method is based on a statistical method for the extraction of terms and concepts, 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.

1.1 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.

2.1.1 Corpus construction

Building ontologies from texts requires a step of corpus construction which represents the domain knowledge [11]. Since we are interested in building domain ontology for railway information services in Tunisian dialect, we used the TuDiCol (Tunisian Dialect Corpus Interlocutor) corpus [14] 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.

2.1.2 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.
The nouns base represents a lexical dictionary which gathers synonyms and all different form of orthographic transcriptions, since the transcription has been done manually and given the lack of standard for the Tunisian dialect. For example, the word “reservation/رزرفسيون” is transcribed in four different ways: “رزرفسيون”, ”رازرفسيون”, ”رارارفسيون”, ”ريزروفسيون”. So, we tried to gather them in a single one. However, the verbs base consists of possible morphological transformations for verbs. Thereafter, we build the stop words base which consists of terms considered by the domain expert as useless. We also build compound words base which includes all possible compound words. For example, if we gather the two words ”بير”, ”بورية”, we obtain the word ”بير_بورقه” which represents a city name. Finally, we build the cities base which consists of all train station names.

2.1.3 Corpus normalization

The normalization step is based on all lexicons created in the corpus treatment step. This step consists in 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. Figure 2 shows an extract of the corpus before and after treatment.

Figure 2. A sample of raw version from the corpus and its corresponding normalized version

2.1 Ontology construction

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

2.2.1 Concept extraction

This step consists in extracting domain concepts specific to the railway information services from the TuDiCol 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 oc-
currence value higher than this concept threshold is regarded as a domain
concept. To justify this choice, we conducted an empirical study which con-
sists of taking each time a small part of the corpus and calculating the fre-
quency of the fixed concept threshold. The occurrence value of the concept
threshold “&lt;rtor” increases with the size of the corpus but it still has the low-
est 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>
<thead>
<tr>
<th>Concepts</th>
<th>Concepts translation</th>
<th>Terms</th>
<th>Terms translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>مدينة</td>
<td>City</td>
<td>مونس/صفاقس/تونس</td>
<td>Tunis/Sfax/Sousse</td>
</tr>
<tr>
<td>قطار</td>
<td>Train</td>
<td>تران</td>
<td>Train</td>
</tr>
<tr>
<td>تذكرة</td>
<td>Ticket</td>
<td>تشكره</td>
<td>Ticket</td>
</tr>
</tbody>
</table>

2.2.2 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 [7]. Before starting pattern
definition, it is necessary to form a study corpus which represents approxi-
mately 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.

2.2.3 Relations extraction

This step consists of projecting the obtained patterns on all the corpus utte-
rances 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:

\[ \text{hello how much round-trip train to Tunis} \]
\[ \text{بقداه رتور ألي تونس إلى رتر نجاء} \]

The corresponding pattern is:

\{ Politeness, train\text{\(\rightarrow\)}city, arrival, ticket\text{\(\rightarrow\)}type, request \}

The implicit relations are detected by pattern projection and the intervention of the human expert. The following example explains this step:

\[ \text{وأربعه ساعتين عادي تران} \]
\[ \text{normal train twenty past two} \]

The corresponding pattern is:

\{ train, train\text{\(\rightarrow\)}time \}

The human expert indicates the existence of the semantic relation “train\text{\(\rightarrow\)}have\text{\(\rightarrow\)}type/تران من نوع” between the first concept “train/تران” and the second concept “train\text{\(\rightarrow\)}type/نوع تران” which connects the two concepts “train/تران” and “train\text{\(\rightarrow\)}type/نوع تران”. 

Our method generates the RIO domain ontology which contains 14 concepts, 25 semantic relations and 387 instances.

2.2 Evaluation

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

2.3.1 Operational evaluation

To evaluate the RIO ontology, we exploit it for semantic annotation which aims to attribute semantic labels to transcribed utterances [13]. This method has already been introduced in [14], which used a manually built ontology. Now, we are going to use the RIO ontology to check its feasibility 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).

\[ F\text{-}measure = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad \text{and} \quad \text{CER} = \frac{\#\text{incorrect concept prediction}}{\#\text{concept_ref}} \]

<table>
<thead>
<tr>
<th>#Training words</th>
<th>#Test words</th>
<th>Correct prediction</th>
<th>Incorrect prediction</th>
<th>Reference concepts</th>
<th>CER</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>5710</td>
<td>2104</td>
<td>1213</td>
<td>162</td>
<td>1718</td>
<td>9.4%</td>
<td>78%</td>
</tr>
</tbody>
</table>

Obtained results for semantic annotation are very encouraging since we have improved the F-measure in comparison with result obtained in [14] 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.

2.3.2 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.

3 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 3). This system provides interactive interfaces for all steps of the 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.

4 Conclusion

In this paper, we proposed a hybrid method for building domain ontology from a spoken dialogue corpus in Tunisian dialect. This approach combines a statistical method for term and concept extraction with a linguistic method for semantic relation identification. The generated ontology helps us to annotate semantically 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


