LMF-based approach for detecting semantic anomalies in electronic dictionaries

Wafa WALI, MIR@CL Laboratory, FSEGS, Sfax, Tunisia; Bilel GARGOURI, MIR@CL Laboratory, FSEGS, Sfax, Tunisia; and Abdelmajid BEN HAMADOU, MIR@CL Laboratory, ISIMS, Sfax, Tunisia.

Abstract
Dictionaries are used for learning and disseminating natural languages. This important role implies that it is necessary to perform the operations of creating, enriching and updating carefully. Even in electronic versions, dictionaries may contain anomalies notably when the used acquisition system is not efficient. Several researches have been made in recent years in order to perform the detection process automatically. They addressed this issue in a superficial way given the major problems it poses. In the present paper we explain how to take advantage from the fine structure of dictionaries proposed by the Standard ISO-LMF (Lexical Markup Framework) in order to refine the detection process of anomalies. We focus in this paper on the anomalies of the semantic level starting from a study of the potential anomalies that can occur in an editorial dictionary. The approach that we propose is based on a set of coherence and consistency rules. An illustration of this approach was carried out on an LMF-standardized dictionary for Arabic language available in our team of the MIRACL research Laboratory. This dictionary contains more than 37000 entries.

Keywords: Automatic detection; Coherence and consistency rules; Electronic dictionaries LMF-ISO 24613; Semantic Anomalies.

1. Introduction
Dictionaries are linguistic and cultural resources which play an important role in learning and diffusing of natural languages. Initially, dictionaries were intended for editorial use in paper versions for a heterogeneous population of human. With the advent of computer science, the support of dictionaries becomes electronic and their use has been expanded to meet the requirements of Natural Language Processing (NLP) applications.

Considering the important role of dictionaries, these later should include consistent information. Incoherent content can not only affect the correct response estimated by the user but it may influence the results of other types of works such as merging dictionaries. However, dictionaries are often published with errors and inconsistencies that are hard to find manually (Fersoe et al., 2008). Until today, the task of analyzing the content of dictionaries is difficult, slow and expensive in requiring a stronger human expert in linguistic. Several researches have been made in recent years in order to perform the detection process automatically (références de quelques travaux: sajac et autres). They addressed this issue in a superficial way given the major problems it poses.

Few years ago, the LMF standard (Lexical Markup Framework) - ISO 24613 was published by the technical committee TC37/SC4 that it is responsible the language resources within ISO (Francopoulo et al., 2008). This standard provides a framework for unified representation of multilingual lexical data in conjunction with all levels of language processing. It provides a fine structure that highlights all kinds of relationships between entries knowledge and distinguishes the role of each available text such as definitions and examples.

In this paper, we take advantage of the LMF standard to propose an approach for the automatic detection of semantic anomalies in LMF standardized dictionaries. An illustration of the proposed approach is given on an Arabic normalized dictionary that is available in our team of the MIRACL laboratory. This dictionary contains more than 37000 entries (Khemakhem et al., 2013).

The present paper is outlined as follows. The next section is devoted to a review of the state-of-the-art of the evaluation of dictionaries contents. Section 3 exposes an overview on the LMF standard. Section 4 reports a study on the relevant semantic anomalies in LMF standardized dictionaries. Section 5 presents the basics of our approach for detection semantic anomalies in
LMF standardized dictionaries and gives a set of coherence and consistency rules on the corresponding stages. Section 6 presents an illustration of our approach, realized on an LMF standardized Arabic dictionary.

2. Related works

The majority of works on dictionaries has focused on the modeling and exploitation tasks, but it has abandoned the control and evaluation of the content. Indeed, there have been a few works that concern this type of problems even including references to dictionaries of paper versions. For monolingual dictionaries, most of the works focused on problems such as false derivation, incoherence of definition and incoherence between example and definition. These works deal with paper versions of old dictionaries like (Alkhatib, 1967), (Alchidyâq, 1899), (Ben Mrad, 1987), and (Hamzaoui, 1986) that are dedicated for the Arabic dictionaries. Other works (Asfour, 2003), (Khoury, 1996), (Kasimi, 1998) dealt with the evaluation of bilingual dictionaries. They specially deal with translation problems.

Moreover, a few efforts are made to detect anomalies for electronic dictionaries as (Zagic et al., 2011) and (Rodrigues et al., 2011). The authors elaborated methods for detecting and correcting OCR problems in Urdu-English digital dictionaries by using Dictionary Language Modeling (DML). However, these dictionaries are poorly structured resulting in the digitalization of paper versions. Furthermore, this situation generates a handicap for the evaluation of electronic dictionaries that require fine structure of the dictionary entries.

Finally, we believe that the lack of works on automatic detection of anomalies in the contents of dictionaries can be explained by the complexity of this task.

3. LMF standard

LMF - ISO 24613 (Francopoulo et al., 2008) provides a common meta-model for the construction of Machine Readable Dictionaries (MRD) and lexicons for the Natural Language Processing use. This standard aims at managing the exchange of data between these resources, and enabling the merging of a large number of individual electronic resources from extensive global electronic resources. The meta-model is organized of eight optional extensions (see Figure1) and a mandatory core package. It should be enriched by data categories (DC) that are selected from the normalized Data Category Register (DCR) ISO-12620.

DC represents basic linguistic concepts (e.g. /part of speech/, /grammatical number/ and /grammatical gender/) and describes the specific linguistic properties to a given language with respect to the ISO 12620.

4. Study on semantic anomalies

The present study was carried out on LMF standardized dictionaries. We will aim to determine pertinent semantic anomalies in such dictionaries. These anomalies may be the result
of the use of a non efficient acquisition system or resulting from an automatic conversion of a numerical and lightly structured dictionary to a well structured version.

In the following subsections, we will present the semantic model for LMF standardized dictionaries and the relevant anomalies at this level.

### 4.1 Semantic model for LMF-standardized dictionaries

In this paper we are limited to present the semantic level for LMF normalized dictionaries as presented in Figure 2 given below. This model is related with the syntactic model, the MRD (Machine Readable Dictionary) model and the core of LMF. The semantic model presents the meanings of word that may be general or specific to one field (Subject Field class). In addition, The Sense Relation allows connecting the meanings belonging to different lexical entries with several types of relationships such as the synonym and antonym. The Sense Example represents an instance of a given sense.

Further, the standard has represented the overlap between syntax and the semantics in the semantic extension. This overlap is mainly the syntactic arguments and their semantic equivalents. The originality of the model is in the representation of the semantic predicate (using Semantic Predicate and Semantic Argument classes) and its link with the syntax through Predicative Representation class that represents the overlap between syntactic and semantic levels. It is specific to one sense and consists of a syntactic behavior, a predicate semantic and the correspondence between them.

![Figure 2: Semantic extension of LMF](image)

### 4.2 Semantic anomalies in LMF dictionaries

For the semantic model, we might find the anomalies which related to the meanings, examples and semantic relationships.

To launch with incoherence of semantic relationships, we can find two types of anomalies: the first is related to existence of contradictory relationships between the same senses of two or more lexical entries as shown the figure below. Figure 3 illustrates synonymous and
antonymous relationships between three items. The sense 1 of word 1 “ذهب” Go “is a synonym with the sense 3 of word 2” "walk" and the sense 2 of word 3 “ممر” "pass" is a synonym with the sense 3 of word 2 "walk". So, by transitivity, the sense 2 of word 3 "walk" and the sense 1 of word 1 “ذهب” Go “are synonyms. However, we can identify an anomaly and we define the previous senses of antonyms.

Figure 3: Example of contradictory relationships between three items

The second is, the senses which belong to different domains but involved in relationships that apply to a same domain. The figure 4 schematized below presents a synonymous relationship between two senses of word cancer-

Figure 4: False semantic relation between senses belonging to different domains

Then, the second anomaly which relates to semantic model is the incoherence of senses. This anomaly deals with the existence of contradictory relationships between different senses of the same lexical entries like illustration in figure 5 given below.

Figure 5: Example of contradictory semantic relationships

Afterwards, we can have another semantic anomaly that is the inconsistency of senses. That is to say, in a dictionary we can find a false definition or false semantic relation that reflects the senses of lexical entry.

Finally, in a dictionary we can find another type of semantic anomalies that relate to consistency and redundancy of examples. Usually, the examples of the same definition of sense are semantically similar and the examples of different definitions are dissimilar as indicated figure 6.

Figure 6: Similarity or dissimilarity between examples and definitions
5. The proposed detection approach of semantic anomalies

In this section, we will present the proposed detection approach of semantic anomalies in LMF standardized dictionaries (ISO-24613). Figure 7 describes this method that consists of three phases. We shortly explain: i) the verification of senses, ii) verification of examples and iii) verification of semantic relations.

![Figure 7: Semantic evaluation method](image)

5.1 Verification of senses

The aim of this phase is to verify sense of each lexical entry. This step consists of checking the redundancy, the consistency and the coherence. Usually, a lexical entry has two different meanings; the definition of these two senses and these examples must be semantically dissimilar. Indeed, the algorithms of similarity are different such as giving Jaccard index (Jaccard, 1901) and Dice (Dice, 1945).

To check the verification of redundancy according the following R1 rule:

**Rule1:**
- \( S_m L_1 \) be the sense \( m \) of the lexical item \( L_1 \)
- \( S_n L_1 \) be the sense \( n \) of the lexical item \( L_1 \)
- \( C_1 \) be the example linked to the sense \( m \) of the lexical item \( L_1 \)
- \( C_2 \) be the example linked to the sense \( m \) of the lexical item \( L_1 \)
- \( TH \) be the threshold defined by an algorithm of similarity index
- \( \text{Sim()} \) be the semantic similarity function
- \( \text{SemCoherence}() \) be a predicate that checks the semantic coherence

We define the following semantic coherence rule:

**Algorithm1:**

Input
\( S_i L_i = \{ \text{senses } S_j \text{ of the lexical item } L_i \} \)
Output
Compt: Number of incoherent semantic relations
After that, we will check the coherence of sense. This verification is used to control whether there is any contradictory semantic relations such as synonymy/autonomy, hyperonymy/hyponymy and holonymy/meronymy between different senses of same lexical entries.

We will detect this incoherence according to the following rule R2.

**Rule2:**

- $S_mL_1$ be the sense $m$ of the lexical item $L_1$
- $S_nL_1$ be the sense $n$ of the lexical item $L_1$
- $S_mL_2$ be the sense $m$ of the lexical item $L_2$
- $S_nL_2$ be the sense $n$ of the lexical item $L_2$
- $SR_c$ set of contradictory semantic relation
- $SR_1, SR_2 \in SR_c$
- SemCoherence () be a predicate that checks the semantic coherence

We define the following semantic coherence rule:

If $SR_1 (S_mL_1, S_nL_2)$ and $SR_2 (S_nL_1, S_mL_2)$ then  
SemCoherence ($S_mL_1$, $S_nL_2$) =false 
Or  
SemCoherence ($S_nL_1$, $S_mL_2$) =false

5.2 Verification of examples

In this phase we will check the consistency of examples. The examples of same sense definition are similar (index of similarity is high). However, a lexicographer can give one or more examples of senses that are semantically dissimilar. To detect this anomaly we will apply the rule R3 described as below.
5.3 Verification of semantic relations
The verification of semantic relations is done in two steps. Firstly, we check whether there are any contradictory semantic relationships between the same senses of two or more lexical entry.

To detect this anomaly, it applies the rule R4 definitely as follows:

<table>
<thead>
<tr>
<th>Rule 3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• ( C_1 ) be the example 1 attached of the sense ( j ) of the lexical item ( L_i )</td>
</tr>
<tr>
<td>• ( C_2 ) be the example 2 attached of the sense ( j ) of the lexical item ( L_i )</td>
</tr>
<tr>
<td>• ( TH ) be the threshold defined by an algorithm of similarity index</td>
</tr>
<tr>
<td>• ( \text{Sim}() ) be the semantic similarity function</td>
</tr>
<tr>
<td>• ( \text{SemCoherence()} ) be a predicate that checks the semantic coherence</td>
</tr>
<tr>
<td>We define the following semantic coherence rule:</td>
</tr>
<tr>
<td>If ( \text{Sim} (C_1, C_2) &lt; TH )</td>
</tr>
<tr>
<td>then ( \text{SemCoherence} (C_1, C_2) = \text{false} )</td>
</tr>
<tr>
<td>/*Comparison strings of characters */</td>
</tr>
<tr>
<td>If ( (C_1 ) is identical to ( C_1 ) or ( C_2 ) is almost identical to ( C_2 )</td>
</tr>
<tr>
<td>Then return (“redundancy of examples”)</td>
</tr>
</tbody>
</table>

Secondly, we verify whether there are senses belonging to different domains but participated in same semantic relationships.
This anomaly will detect according to the subsequent rule R5:

<table>
<thead>
<tr>
<th>Rule 4:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• ( S_{L_1} ) be the sense ( i ) of the lexical item ( L_1 )</td>
</tr>
<tr>
<td>• ( S_{L_2} ) be the sense ( j ) of the lexical item ( L_2 )</td>
</tr>
<tr>
<td>• ( S_{L_3} ) be the sense ( k ) of the lexical item ( L_3 )</td>
</tr>
<tr>
<td>• ( \text{SR} ) set of contradictory semantic relation</td>
</tr>
<tr>
<td>• ( \text{SemCoherence()} ) be a predicate that checks the semantic coherence</td>
</tr>
<tr>
<td>We define the following semantic coherence rule between two lexical items ( L_1, L_2 ):</td>
</tr>
<tr>
<td>If ( \text{SR}<em>1 (S</em>{L_1}, S_{L_2}) ) and ( \text{SR}<em>2 (S</em>{L_1}, S_{L_2}) )</td>
</tr>
<tr>
<td>Then ( \text{SemCoherence} (S_{L_1}, S_{L_2}) = \text{false} )</td>
</tr>
<tr>
<td>We define the following semantic coherence rule between three lexical items ( L_1, L_2, L_3 ):</td>
</tr>
<tr>
<td>If ( \text{SR}<em>1 (S</em>{L_1}, S_{L_2}) ) and ( \text{SR}<em>1 (S</em>{L_1}, S_{L_3}) ) and ( \text{SR}<em>2 (S</em>{L_2}, S_{L_3}) )</td>
</tr>
<tr>
<td>Then ( \text{SemCoherence} (S_{L_2}, S_{L_3}) = \text{false} )</td>
</tr>
</tbody>
</table>

6. Illustrating the approach for Arabic language
The proposed approach is illustrated on an available LMF-standardized dictionary of Arabic language within our research team. Along with the present study, many experiments have
been carried out on the Arabic language to assess the applicability and the feasibility of the proposed method for detecting the semantic anomalies in the content of LMF normalized dictionaries. In the following subsections, we will, first, present the framework of illustration in giving an insight on LMF standardized Arabic dictionary and Jaccard index of similarity. Then, we will show the results of applying the three stages of our approach.

6.1 Context of the illustration

The LMF standardized Arabic dictionary was developed within the framework of the project of the electronic interactive Arabic dictionary managed by the ALECSO\(^1\). Our research team has developed the standard model of this dictionary (Khemakhem et al., 2013) and worked out an experimental version for research, and currently composed of 37000 lexical entries. This dictionary is constructed using an automatic conversion of a numerical version (reference de l’article de MEDAR 2009 Khemakhem et al.). Indeed, the anomalies that this dictionary may include can be seen from the initial construction by the lexicographer, the acquisition within the numerical version or the conversion process to the normalized representation.

In order to calculate the similarity between texts, we are using the Jaccard index (Jaccard, 1901) that corresponds to the ratio between the number of words present in both texts (the intersection) and the number of words of the two texts (the union). Indeed, the Jaccard index is normalized between 0 and 1. Besides, if the Jaccard index exceeds to 0.5, both texts are similar.

6.2 Illustrative case

The applications of the three stages for detecting the semantic anomalies, as well as the resulting semantic evaluation report are presented in the paragraphs that follow.

**Check the senses:** figure 8 illustrates the principle of verification such as redundancy the senses of a lemma like definitions and examples. Thus, we are seeking if two definitions of senses “Sense1” and “Sense2” of the lemma “cancer-سَرَطَان” are similar. In addition, these examples are presented in the *Context* class according to rule R1. We find the following results:

- The similarity between the definition of sense1 and the definition of sense2 equals 0.
- The similarity between the example of sense1 and the example of sense2 equals 0.

![Diagram](image)

**Figure 8:** Portion of the lexical entry cancer “سَرَطَان”

In figure 9, we show the verification of existence of a contradictory relationship between two different senses from two lexical entries. Thereby, we are seeking if the relationship between the senses “S22” and “S11” is contradictory with the relation between the senses “S11” and “S21” given that the senses “S11” and “S21” are related to the lemma “deviation-إِنْحِراف” and the

\(^1\) The Arab League Educational, Cultural and Scientific Organization
senses “S12” and “S22” are related to the lemma “inclination- مَيْلٌََ”. At this verification, we use the rule R2.

Check the examples: figure 10 presents the ideas of the examples’ verification which relate to the same definition. Thus, we try to find out if the example “the student writes the lesson- كَتَبَ التِّلْمِيذُ الدَّرْسَ” is similar to the example “the boy writes the lesson- كَتَبَ الوَلَدُ الدَّرْسَ”. However, both examples use the same words except that the subject differs in the example1, “the student- التِّلْمِيذُ” and “the boy- الوَلَدُ” for the example 2. The semantic similarity between the previous examples that is calculated with Jaccard index equals to 0.66. To check this anomaly, we are applying the rule R3.

Check the semantic relationships: figure 11 illustrates how we verify the semantic relations between two or more lexical entries. The principle is to seek if the semantic relation between the senses "S11" and S12” is contradictory with the semantic relation between the senses “S12” and “S13” or the semantic relation between the senses “S11” and S13”. For example, the value of semantic relation is presented in Sense Relation class. It has been known that the senses “S11” relate to the lemma “go- ذهَب“; “S12” relates to the lemma “walk- سَارَ“; “S13” relates to the lemma “pass- مَر” and they are synonyms. We are checking this anomaly according to the rule R4.
In figure 12, we show the idea to verify the existence of a semantic relationship between two senses which belong to different domain. The principle is to seek if there is a semantic relation between the senses “S11” and “S21” giving that the sense “S11” which relates to the domain “medicine” and “S21” which relates to the domain “astronomy” that are presented in the SubjectFieled class. We are checking this anomaly according to the rule R5.

Figure 12: Portion of the lexical entry “cancer-سرطان”

7. Conclusion

In this work, we proposed an approach for the detection of semantic anomalies in LMF standardized dictionaries. Taking advantage from the fine structure of the dictionaries LMF model, we proposed and developed appropriate rules for the detection of some anomalies which relate to the semantic level. The detection approach is based on the following three phases: verification of the senses, verification of the examples and verification of the semantic relationships. Besides, the proposed approach is proven to be reliable through the illustration carried out on a normalized Arabic dictionary. We justify the use of such dictionary by two main reasons. The first is the great deficiency of works on the evaluation of Arabic dictionaries contents. The second concerns the availability within our research team of an LMF standardized Arabic dictionary which is automatically converted from a numerical version.

At present, we are setting to formalize all the detection rules of the semantic anomalies in the LMF standardized dictionary. In fact, we focus on the values of attributes such as definitions and examples and the relationships between lexical entries.

As future perspectives to our work, we plan to develop the necessary API for detecting the semantic anomalies of LMF normalized dictionary.

References


