The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. 96-109

Towards automatic enrichment of standardized electronic dictionaries by semantic classes

Elleuch Imen*, Gargouri Bilel* and Ben Hamadou Abdelmajid[1]

Abstract

In this paper we propose an approach for the automatic enrichment of standardized electronic dictionaries by the semantic classes. This approach consists of three phases. The first phase treat the semantic classification process founded on the studies of Gaston Gross. The second phase profites from the existed subject fields in the dictionary's lexical entries in order to attribute the suitable semantic classes. The final phase realizes syntactic analyses of the textual content of meanings's lexical entries. This phase, aims to refine the subject field based enrichment and also treats the non enriched meanings in the second phase. In addition, it attributes the same semantic classes for the synonym meanings. We used an available standardized Arabic dictionary to tested the performance of the proposed approach.

Keywords: Automatic enrichment, standardized electronic dictionnaries, semantic classes, Arabic language.

1. Introduction

Semantic knowledge, especially semantic classes which aim to characterize meanings of lexical units in dictionaries, have attracted considerable interest in both linguistic (Stede, 1998), (Dorr, 1997) and computational linguistics (Kipper et al 2000). Such semantic class can be definite as a semantic linguistic propriety classifying meanings and can therefore be used as a valuable means of comprehending the specific meaning of polysimous lexical units. Thus the need of dictionaries with semantic classes has become a necessity for Natural Language Processing (NLP) applications.

For various languages, various semantic classifications are now available. We can list the verbs classification (Pinker, 1989; Jackendoff, 1990; Levin, 1993, Dubois and Dubois-Charlier, 1997) that regroups together verbs that share both a common semantics and a set of syntactic alternations. Also, we notice WordNet (Fellbaum, 1998) that provides semantic ontological classification and FrameNet (Fillmore, 1985) that hierarchically classify lexical units using various relationship as synonymy, antonym and is-a relations. However, the referential classification is based on semantic features like [+/- human], [+/- concrete], etc. characterizing semantically each lexical unit outside of the meaning's contexts. Object classes (Gross, 1994) defines a semantic classification based on surface realization of predicate argument structure. A semantic class groups together predicates as arguments having the same syntactic constrictions. Rely on a semantic classification; two methods of enrichment lexical resources by semantic classes exist. The first one is manual. It is characterized by the large number of lexical units to be classified FrameNet

* FSEGS, B.P. 1088, 3018 Sfax, Tunisia
E-mail: { imen.elleuch, bilel.gargouri}@fsegs.rnu.tn

** ISIMS, B.P. 242, 3021 Sakiet-Ezzit Sfax, Tunisia
E-mail: abdelmajid.benhamadou@isimsf.rnu.tn

(Fillmore, 1985) this is why it is a costly and time-consuming method. The second method is automatic. It can use corpora (Fuchs & Habert, 2004), (Condamines, 2005) or in some cases, texts of the treated lexical resources (Rastier, 2001) and (Valette et al, 2006). The automatic method does not necessitate the intervention of the human expert during the enrichment process (Wilson et al., 2004). Both manual and automatic method of enrichment lexical resources with semantic classes requires the institution of the semantic classification. In addition, the ability of the structure's dictionary to receive semantic classes is important. In fact, some models of lexical resources do not supply the affectation of the semantic classes to lexical units.

In order to provide a unified framework for modeling lexical resources, in general, and to facilitate the exchange and integration into NLP applications, the LMF (Lexical Markup Framework) standard (Francopoulo & George, 2008) ISO 24613 is published. This standard allows the modelization of all linguistics levels such as the morphological, the syntactic, the semantic and the syntactico-semantic ones.

Considering the importance of the semantic classes to characterize the meaning of lexical units, and profiting from the fine model of LMF lexical resources to receive semantic classes, we propose in this paper an automatic approach for the enrichment of standardized LMF electronic dictionaries by semantic classes. In fact, the LMF standard offers particular fields (i.e., SubjectField) that can assist the identification of the relevant semantic class and provides synonymy relationships that can be used to improve the enrichment process. Also, in an LMF dictionary, the meaning of lexical entries is accompanied with a rich textual content. The proposed approach is founded on a semantic classification initiated by the Gaston Gross studies. An experimentation of this approach is carried out on an available standardized LMF Arabic dictionary.

The next part of this paper is organized as follows: We will start with a presentation of some related works related to semantic classification and enrichment methods. Then, we will present the LMF standard. Thereafter, we will detail the proposed approach for the enrichment of LMF standardized dictionaries with the semantic classes. After that, we will describe the experiment carried out on a standardized LMF Arabic dictionary and discuss some of the obtained results. Finally, in the conclusion, we will announce some future works.

2. Related works

This section is devoted to the representation of some related works of available semantic classifications and the semantic enrichment methods of lexical resources.

2.1. Semantic classification

Several semantic classifications exist in literature. We can mention the verbs classification (Pinker, 1989; Jackendoff, 1990; Levin, 1993, Dubois & Dubois-Charlier, 1997). It based on both a common semantics and a set of syntactic alternations to grouped lexical units into semantic classes. This type of classification is restricted to certain class types and treats only verbs. So no comprehensive classification is available limits the usefulness of the class for practical NLP tasks.

Moreover, we can note the ontological classification like WordNet (Miller, 1990) that intended to classify philosophical things as they exist in the world. It is particularly appropriate for object modeling, including their relationships and properties. Therefore, content of ontology does not interact directly but rather with relationships (i,e synonymy, antonym, part of, is-A,…). This semantic classification does not consider the use's context of lexical units, further it groups word into classes as presented in the real world without referring to the linguistics features.

Also, we can cite the referential classification (Gross, 1975) (Dichy, 2000) that used semantic features like [+/- concrete], [+/- human]. Those features are attached to lexical units to describe their appurtenance to the

semantic classes. This semantic classification assigned semantic features to lexical entries without taking into account the uses of the lexical units.

Another kind of semantic classification is proposed by Gaston Gross (Gross, 1994). It classifies lexical units into semantic classes based on predicate-argument structure. Thus, a semantic class groups together predicates as arguments sharing syntactic and semantic behaviors. Therefore, this classification insures the taking into account the multiple meanings of senses lexical entries depending on a specific use context.

Finally, we can conclude that the ontological and the referential classification do not guarantee the polysemy of lexical entries because they do not take into account meanings in the classification process. Or the verbs classifications classify only verbs and neglect the other part of speech whereas, the Gaston Gross semantic classification defines a syntactico-semantic classification based on predicate-argument structure. Thus, the variety meaning of senses lexical entries related to an applicable context was ensuring.

2.2. Semantic enrichment

Firstly, the semantic enrichment was done manually. Doing so, this enrichment necessitates high linguist capacities in order to affect the pertinent semantic class to meaning. The LADL tables (Gross, 1975) is one of studies that is based on a manually affectation of semantic features to lexical units meanings.

It is clearly that this manual enrichment is the most relevant one, but it requires a costly time because the vast number of lexical units to be classified and it necessitate the availability of the linguist who attribute the adequate semantic classes to meanings.

With the progress characterizing the computational linguistic domain, the enrichment methods become automatic. This automatic enrichment uses both linguistics features and mathematics techniques to classifying lexical units. This enrichment is marked by three ways. The first uses the linguistic tools for preparing the corpus before classifying lexical units by means of clustering tools (Wilson et al., 2004). In fact, the construction of the corpus requires the annotation steps that represent a heavy and time consuming task. Several clustering algorithms can be used as Ripper (Cohen, 1996). The second way uses techniques of automatic clustering (Hatzivassiloglou & McKeown, 1997). In this case, it is necessary to add syntactic and semantic features in order to achieve the automatic enrichment. The third way consists of using linguistic and statistical approaches. The purpose of this way is to build several types of classifiers and combine their results, either by voting systems or by clustering algorithms (Dziczkowski & Wegrzyn-Wolska, 2008). This kind of enrichment needs heavier treatments than the other manners listed above.

3. LMF standardized model

LMF is a standard ISO 24613 for modeling lexical knowledge of the majority of natural languages (Francopoulo & George, 2008). It provides a common model for the representation of electronic lexical resources with guarantees the exchange of data between and among these resources. The LMF model is composed of a core package and a range of extensions referring to the various levels of linguistic analysis (i.e., morphological, syntactic, semantic and syntactico-semantic). The LMF core package describes the basic hierarchy of lexical entry information, including information on the form. The LMF extensions are added to the LMF core components in conjunction with the additional components required for the specific resource modeling. Indeed, to obtain lexical resources according to the LMF standard, it is sufficient to have the core package, then, optionally select packages of extensions necessary to the representation of the modeled dictionary. It is also, essential to select from each extension the corresponding LMF classes required to the treated language. For example, the core package provides the *Sense* and the *Definition* classes to describe the meaning of a lexical entry. The MRD (Machine Readable Dictionary) extension reserves the *Subject Field* class to represent the domain of use of a *Sense* and the *Context* class to describe the authentic context for the use of the word form managed by the lexical entry. The LMF semantic extension designates the *Sense Relation* class to describe the possible relationship between *Senses* instances such as synonymy and autonomy. Then,

the resulting model will be decorated with the Data Categories Registry (DCR)[2] required for the modelization of the dealt language.

4. Proposed approach

In this section, we detail the proposed approach for the automatic enrichment of LMF standardized electronic dictionaries by the semantic classes. The following figure 1 illustrates steps of this approach.

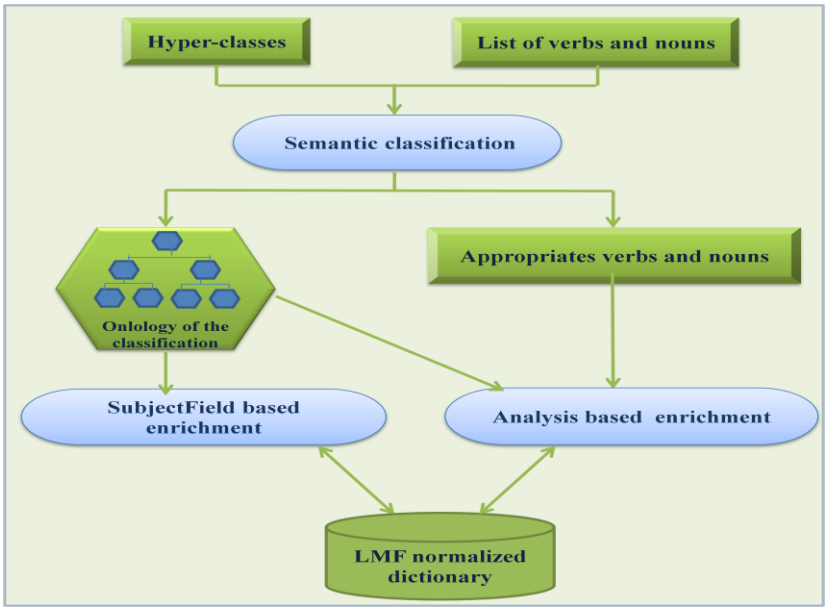

Figure 1: Proposed approach

The proposed approach is composed of three steps: a semantic classification and two phases of automatic enrichment. To accomplish the aims of the semantic classification, this step requires the hyper-classes of the Gaston Gross classification and a list of verbs and nouns of the studied language in input. The results of the semantic classification step are the ontology of the classification and a list of appropriate verbs and nouns characterizing this classification. Whereas, the SubjectField based enrichment uses the ontology of the classification to enrich the LMF normalized dictionaries by identifying semantic classes. The analysis based enrichment requires achieving the enrichment of the LMF normalized dictionary, both the ontology of the classification and the list of appropriates verbs and nouns identified previously.

4.1. Semantic classification

4.1.1. Basic concept

Our semantic classification is based on the studies of the Gaston Gross (Gross, 1994) semantic classification (see section 2). This classification uses the predicate-argument structure to classify lexical units. Thus, the simple sentence represents the minimum unit of analysis. Indeed, two major semantic classes characterize this classification namely: the semantic classes of predicates and the semantic classes of arguments. However, prior to the object classes, and based on syntactic features, the classification maintains classes that regroup all predicates that share the common syntactic behaviors named Hyper-classes. Thus, hyper-classes of predicates, specified by this classification are: "ACTION, EVENT, STATE and PREDICATIVE HUMAN." While hyper-classes of arguments are: "HUMAN, CONCRETE, PLANTS, ANIMALS, TIME, RENTAL and ABSTRACT." These hyper-classes are subject to sub classifications by means of arguments permutations (distributional criteria) appearing in one or more positions of arguments related to a given predicate. Thus, if a permutation of a noun by another contributes to a rupture of the meaning of a predicate sense, then a new object class is required to be created. These object classes allow highlighting the different uses of a polysemous predicate.

[2] www.isocat.org

4.1.2. Steps of the semantic classification

We propose in figure 2 the general semantic classification process.

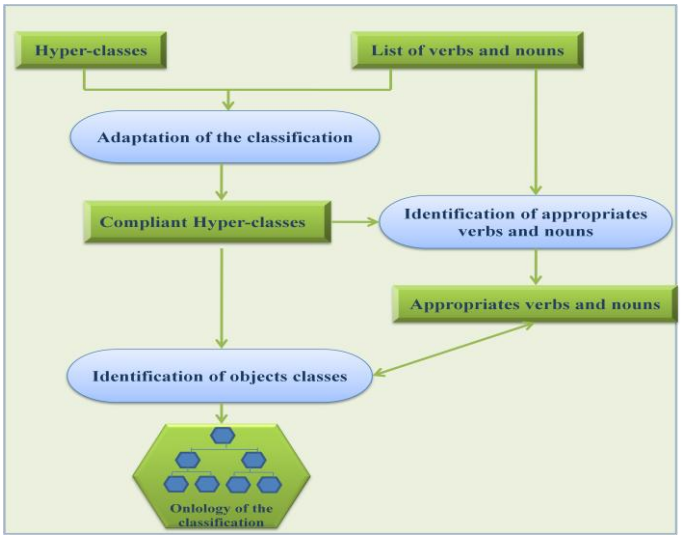

Figure 2: Semantic classification

The process of the proposed semantic classification is realized manually by a linguist. It composed by three steps: (i) Adaptation of the classification, (ii) Identification of appropriates verbs and nouns and (iii) Identification of object classes.

i. Adaptation of the classification:

Hyper-classes of the Gaston Gross studies (see section 4.1.1) and a *list of verbs and nouns* of the studied language perform together in order to accomplish the adaptation of the adaptation of the classification step. Considering that the semantic classification is performed by a linguist, this step requires the abilities of this expert and the syntactic features of the studied language in order to study the possibility of the adaptation of the semantic classification on the studied language. On the basis of syntactic features of the studied language, the expert can identify new hyper-classes appropriate to the treated language, delete or rename the existing semantic hyper-classes. Therefore the *compliant hyper-classes* represent the result of this step.

ii. Identification of appropriates verbs and nouns:

On the basis of the novel list of *hyper-classes* identified in the previous step, related to the specific studied language, the identification of appropriates verbs and nouns take place. This step aims to detect the *appropriate list of verbs and nouns* characterizing each hyper-classes of the proposed semantic classification.

iii. Identification of object classes:

The object class concept represents the characteristic of the proposed semantic classification. Thus, the aim of this step is the *identification of object classes* for each semantic class. To accomplish this objective, this step requires the *compliant hyper-classes* of the studied language and the list of *appropriates verbs and nouns* recognized in the last step. The results of this step affect predicates-semantic classes to as well as arguments. Indeed, the expert benefits from the syntactic features of the studied language in order to identify objects classes relating respectively to hyper-semantic classes of predicates and arguments. As hyper-classes, the identification of the object classes outcomes a list of verbs and nouns characterizing each object class. This list performs to update the list of appropriates verbs and nouns of the classification. An *ontology of the classification* that's regroups all complaint hyper-classes and object classes related to the studied language represent the result of this step.

4.2. Enrichment of LMF standardized dictionaries

After developing a semantic classification, the enrichment process of the standardized LMF dictionaries with semantic classes will take place. It composed of two main phases: (i) the Subject Field based enrichment that benefited from the LMF dictionaries structure, particularly from the uses domains related to meanings of lexical entries (ii) the analysis based enrichment that uses features of the obtained semantic classification.

4.2.1. Subject Field based enrichment

This enrichment is based on the field "SubjectField" according to the LMF model. As shown in figure 3, it consists of two steps described as follow:

i. Searching senses with "SubjectField": the domains of uses for each "*Senses"* of lexical entries in *LMF normalized dictionary* are represented through a class named "SubjectField". The aim of this step is the extraction from the dictionary, *Senses* related to treated lexical entry containing the "SubjectField" field.

ii. Identification of semantic classes: a pretreatment realized on the obtained semantic classification and the existed "Subjectfield" in an LMF stadardized dictionary can made a directly correspondence between the hyper-semantic classes or the object classes with the "SubjectField". If this is the case, this step identifies the semantic class from the *ontology of the classification* related to the founded *"SubjectField"* and updates the LMF standardized dictionary by the addition of the retained semantic class to the corresponding *Sense*.

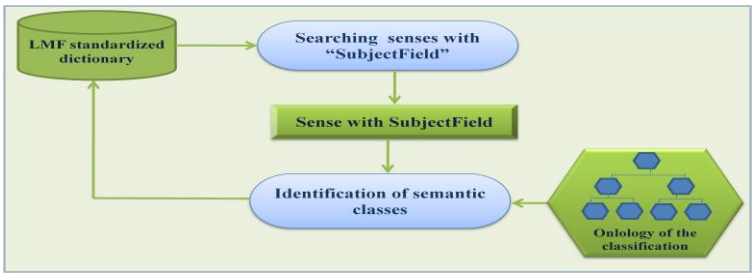

Figure 3: Subject Field based enrichment

4.2.2. The analysis based enrichment

The analysis based enrichment uses the features of the retained semantic classification. The following figure 4 illustrates the steps of this kind of enrichment.

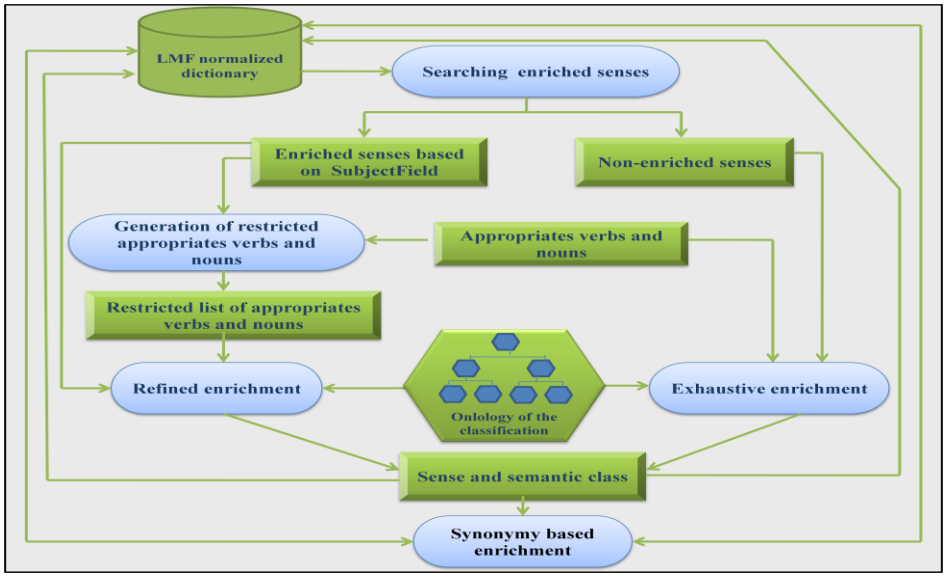

Figure 4: Analysis based enrichment

The list of appropriates verbs and nouns and the ontology of the classification represent the input of analysis based enrichment. It is composed by the following five steps:

i. Searching enriched senses: this step aims to search from *LMF normalized dictionary* the *enriched senses* with semantic classes based on the SubjectField based enrichment and in the same time the *non-enriched senses*. A specific treatment will be affected to those senses in the next step.

ii. Generation of restricted appropriates verbs and nouns: the assignment of the semantic class identified by the SubjectField based enrichment is not a definitive assignment. Indeed, in order to achieve the definitive enrichment, this step requires for the realization of its process both the *Appropriates verbs and nouns* and the *Enriched senses based on SubjectField. The restricted appropriates verbs and nouns* represent the result of the generation of restricted appropriates verbs and nouns phase.

iii. Refined enrichment: the *restricted list of verbs and nouns* identified in the last step, *the sense already enriched based on SubjectField* and *the ontology of the classification* represents the input of this step. Indeed, this step uses the *restricted list of verbs and nouns* to analyze the textual content of the enriched "Sense" in order to refine the semantic class assignment. Thus a relevant *semantic class* is identified based on the ontology of the classification and will definitive be attribute to the treated *Sense*.

iv. Exhaustive enrichment: the exhaustive enrichment concerns the *non-enriched senses*. In fact, a specific treatment is performed to those non-enriched senses by the means of the *Appropriates verbs and nouns identified* by the retained semantic classification. This treatment consist of an analyze of the *"Contexts"* and the *"Definitions"* field related to a Sense of a lexical entry in the *LMF dictionary* using the appropriate verbs and nouns. This analyze identify the relevant semantic class from *the ontology of the semantic classification* which will be affected to the Sense in order to enrich semantically the LMF dictionary.

v. Synonymy based enrichment: in this step we have identified and affected a semantic class to the treated Sense. After that, the synonymy based enrichment takes place, it aims to search the synonymy senses related to the treated sense. Then, the same semantic class identified by the exhaustive or the refined enrichment will be affected to the synonymy senses. At the end of this step, we obtain an enriched sense with the relevant semantic class and also the related synonymy senses enriched by the same semantic class.

5. Experimentation on the Arabic language

This section focuses an experimentation of the proposed approach of the automatic enrichment of standardized dictionaries by semantic classes. An Arabic LMF dictionary is used to test the performance of this approach.

5.1. Choice of the Arabic language

With respect to the Arabic language and to our knowledge there has been no works treated effectively an Arabic semantic classification. In fact, in literature available works are limited to some attempts of specialized dictionaries without related to any theoretical semantic classification. We can note for example, the *" فقه اللغة وسر العربية/ fiq.hu all~uγaħi wa sir~u alçarabiy~ati"* dictionary created by *" أبو منصور الثعالبي/ Aabuw mansuwr alθ~açaAlibiy"* which groups lexical units into thirty chapters. Each chapter is subdivided into sub-chapters grouping together lexical units sharing the same semantic meaning. The chapter " في اللباس وما يتصل به والسلاح وما ينضاف إليه وسائر الأدوات / والآلات وما يأخذ مأخذها fiy al~*ibaAs wa maA yat~asilu bihi wa als~ilaAH wa maA yan.DaAfu Ălay.hi wa saAŷiri alAadawaAti wa alĀlaAti wa maA yuA.xaDu maA.xaDahaA*" includes forty-nine sub-chapters as *"في تقسيم النسيج/fiy taq.siym aln~asiyj"* "the division of tissues", "في تقسيم الخياطة/*fiy taq.siym alHiyaATaħi* " "the division of sewing ", " في تقسيم الخيوط وتفصيلها/ *fiy taq.siym alxuyuwT wa tafSiyluhaA* " "the division of thread and its peculiarities "....

" *المعجم العربي لاسماء الملابس/almuç.jam alçarabiy lias.maA'i almalaAbis"* is another Arabic dictionary specialized in the classification of Arabic nouns of clothes. *" رجب عبد ابراهيم/ rajab çabd Aib.raAhiym"* the writer of this lexical resource grouped more than 1250 clothes Arabic nouns.

5.2. Illustration of the Arabic semantic classification

5.2.1. Classification of Arabic arguments

In this section, we experiment the process of the semantic classification (see section 4.1) on the Arabic language. We were interested in this experimentation on the "CONCRETE" hyper-class of arguments. This

hyper-class is also retained for the Arabic language from the classification of Gaston Gross. Among the object classes belonging to the "CONCRETE" hyper class we note the "Clothes" class. Indeed, the Arabic verb "لبس" "to wear", represent the appropriate verb characterizing this object class. Thus, one meaning of this verb describe an "ACTION" realized by a first "HUMAN" argument and highlighting another "CONCRETE" argument. The example bellow illustrates three sentences detailed the mean of the "لبس" "to wear" verb:

(1)	لَبِسَ التِّلْمِيذُ القُبَّعَةَ	Labisa alt~*il.miydu alqub~açaħa*	The pupil wears the hat
(2)	لَبِسَ التِّلْمِيذُ التُّفَاحَةَ	Labisa alt~*il.miydu* alt~*ufaHaħa*	The pupil wears the apple
(3)	لَبِسَ المَاءُ القُبَّعَةَ	Labisa almaA*'u alqub~açaħa*	The water wears the hat

All sentences (1), (2) and (3) are syntactically correct. But, only the sentence (1) is semantically acceptable. Indeed, in sentence (1), a "*pupil"* can "*wear a hat"*, while in sentence (2) a *"Pupil"* cannot wear an *"apple"* because an *"apple"* is an "Aliment" so it can be eaten but not wean. Whereas, in sentence (3) the *"water"* is an "Aliment/ water" and cannot be wean. Those examples explicate the requirement of the creation of the "Clothes" and the "Aliment" objects classes under the "CONCRETE" hyper-class. Thus, he "Clothes" object class includes all nouns that can be worn by a "HUMAN". Arabic verbs such as: *"خلع /xalaça/to undress" "ارتدى/Air.tady'/to dress", "لبس/labisa/to wear"*, and nouns like: *"كساء/kisaA''/cloth"*, "لباس/libaAs/wear", *"ثوب/θaw.jbũ/dress"* characterize the "Clothes" object class.

Arguments instances of the "Clothes" object class can be: *"حذاء/HidaA'/shoes", "نعل/naç.l/sock", "خف/xuf~ũ/slipper", "قبعة/qub~açaħũ/hat", "سروال/sir.waAlũ/pant", "قميص/qamiuSũ/shirt"*. Thus, the appropriate verbs of the "Clothes" object class like *"خلع /xalaça/to undress" "ارتدى/Air.tady'/to dress"" "لبس/labisa/to wear"* can be correctly introduces the arguments instances list before. But in Arabic language, some verbs select from the "Clothes" arguments instances a specific ones but cannot use all of them. For example, the (احتذى/ Ain.taçala / to wear shoes, انتعل /AiH.tady /to wear shoes) verbs cannot precede allof the arguments instances "Clothes" but only <shoes> <الحذاء>. Thus, the sentence (احتذى القميص) (he wear shoes shirt) is semantically incorrect because (احتذى/Ain.taçala / to wear shoes) is an appropriate verb to <shoes> <الحذاء> class and (القميص/ *alqamiuSũ/shirt*) does not represent <shoes> <الحذاء> but rather <fringues> <ثياب >. Therefore, it is necessary to create two objects classes under the "Clothes" namely <shoes> <الحذاء> and < fringues > <ثياب>.

The table 1 in following summarizes the previous idea:

Table1: Appropriate verbs for the <shoes> <الحذاء> and <fringues> <ثياب> object class

		Examples of nouns / verbs	لَبِسَ *labisa to wear*	احتذى AiH.tady to wear shoes	خَلَعَ *xalaça to undress*	إرتدى *Air.tady' to dress*	انتعل Ain.taçala to wear shoes
<Clothes> object class	<shoes> <الحذاء>	خف/xuf~ũ/ slipper	👍	👍	👍	👍	👍
		حذاء/HidaA'/ shoes	👍	👍	👍	👍	👍
		نعل/naç.l/ sock	👍	👍	👍	👍	👍
	<fringues> <ثياب>	قميص/qamiuSũ/ shirt	👍	👎	👍	👍	👎
		سروال/sir.waAlũ/ pant	👍	👎	👍	👍	👎
		قبعة/qub~açaħũ/ hat	👍	👎	👍	👍	👎

5.2.2. Results of the Arabic semantic classification

In this section we present an example of the semantic classification ontology for Arabic language. The below figure 5 illustrates some recognized hyper-classes and object classes using the proposed process of the semantic classification (see section 4.1). This figure is created with the OWL ontology.

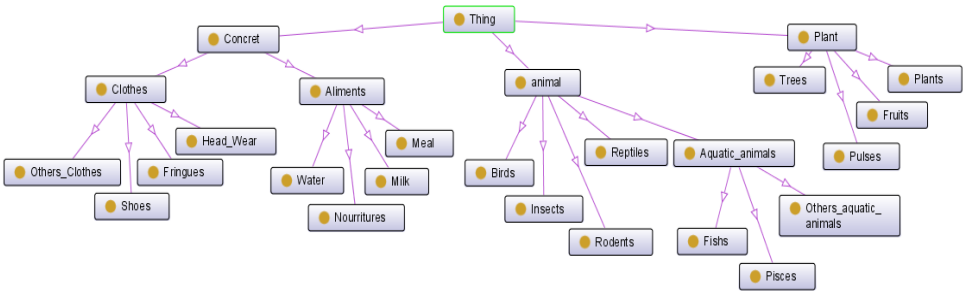

Figure5: Examples of Arabic hyper-classes and object classes of arguments

As presented in the figure 5, a "CONCRETE" is a hyper class. Among the object classes founded under the "CONCRETE" hyper class we note the "Clothes" and the "Aliments" subclasses. The "Clothes" object class is subdivided into the following object-sub-classes: "Head_wear", "Fringues", "Shoes" and "Others_Clothes". And so on for the other hyper and object classes.

Table 2 present the list of appropriate verbs and nouns related to the "CONCRETE" <clothes> class:

Table2: Appropriate verbs and nouns of the <clothes> object class

			Appropriate										
	Object classes		Verbs			Nouns							
CONCRETE hyper-class	Clothes		لَبِسَ labisa	خَلَعَ xalaça	إرْتَدَى Air.taday	لِبَاس libaAs	ألْبِسَة Aal.bisaħ	ثَوب θaw.jb	أثْوَاب Aθ.waAb	كِسَاء kisaA'	أكْسِيَة Ak.siyaħ	رِدَاء ridaA'	أرْدِيَة Ardiyaħ
			to wear	to undress	to dress	wear(s)		dress(es)		clothe(s)		apparel(s)	
	Sub-object-classes	<fringues> <ثوب>	تَبَرْدَسَ tabar.nasa	تَجَلْبَبَ tajal.baba	تَبَرْقَعَ tabarqaça	قصقاص faD.faAD				وَاسِع waAsiç			
			to wear the bumous	to wear	to wear the veil	widish				ample			
		<shoes> <الحذاء>	احْتَذَى Ain.taçala	انْتَعَلَ AiH.tady		حذاء HidaA'	أحْذِيَة AH.diyaħ			نَعْل naç.l	نِعَال niçaAl		
			to wear shoes	to wear shoes		shoes				*sock(s)*			
		<head wear> <أغطية الرأس>	تَعَمَّمَ taçam~ama			رَأس raA.s	رؤوس ruŵius	حَاسِر Hasir		حَافٍ HaAf			
			to wear a hat			head		bare		unshod			
		<others wear> <غيرها>	دَرَزَ daraza	بَكَلَ bakala		رجل rij.l	رجلون rijlay.n	أرجل Arjul	يَد yad	يَدَيْن yadayn	أيدي Ay.dy	كف kaf	كفين kaf~ay.n
			to sew	to buckle		foot			hand(s)			forehand(s)	

5.3. Illustration of the enrichment of the Arabic LMF dictionary

5.3.1. Arabic LMF standardized dictionary

The Arabic LMF standardized dictionary is a lexical resource conforms to the LMF standard ISO-24613. The model of this dictionary (khemakhem et al 2013) covers all lexical levels: morphological, syntactic, semantic and syntactico-semantic. This dictionary contains about 37000 lexical entries among them 10800 verbs and 3800 roots.

5.3.2. Experimentation of the "SubjectField" based enrichment

In Arabic LMF standardized dictionary, three classes namely *Definition*, *Context* and *SubjectField* characterize the *sense* of lexical entry. The *Definition* determines the meaning of sense. While the *Context* gives an example of using sense. Regarding the *SubjectField* it describes the use's domain related to a given sense of a lexical entry. The table below contains some examples of domains available in the Arabic LMF standardized dictionary.

Table3: Examples of available Subject Field in the Arabic standardized dictionary

Subject Field		
In Arabic	**Transliterated**	**In English**
حَيَوَان	HayawaAn	Animal
حَشَرَة	Hašaraħ	Insect
نَبَات	nabaAt	Plant
هَنْدَسَة	han.dasaħ	Geometry
طَبْخ	Tab.x	Culinary
جُغْرَافِيَا	juγ.raAfãyaA	Geography
مُوسِيقَى	musiyqaA	Music
رِيَاضَة	riyaADaħ	Sport
طِب	Tib	Medicine
عَسْكَر	ʕas.kar	Military

In the Arabic LMF normalized dictionary, the "حيوان" "Animal" "HayawaAn" and the "حشرة" "Hašaraħ" "Insect" *SubjectFields* can be grouped into the "animal" hyper-class. It is important to indicate that the "حشرة" "Hašaraħ" "Insect" *SubjectField* corresponds directly to the object class named “Insect” and the "حيوان" "HayawaAn" "animal" *SubjectField* may correspond to the object classes: “Bird”, “Rodents”, “reptiles” and “Aquatic-animals” as shows in figure5.

The following figure 6 illustrates an example of the SubjectField based enrichment.

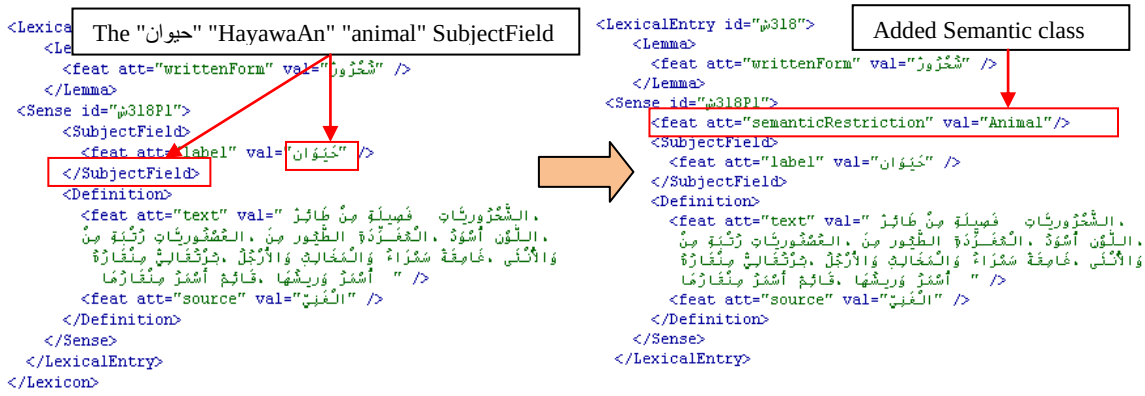

Figure 6: The SubjectField based enrichment applied to a sense of lexical entry

5.3.3. Experimentation of the analysis based enrichment

The analysis based enrichment is subdivided into two kind of enrichment. The first enrichment appointed refined enrichment requires for the progress of its process the restricted list of verbs and nouns in order to refine the primary enrichment carried out in the SubjectField based enrichment. Or the second enrichment is exhaustive, concerning only non-enriched senses, uses the appropriate verbs and nouns of the semantic classification in order to realize the semantic enrichment of the dictionary.

The table 4 given in the following, contains the restricted list of appropriates verbs and nouns related to the "حيوان" "HayawaAn" "animal" arguments hyper-class.

Table4: Restricted list of appropriate verbs and nouns of "Animal" hyper-class

Hyper class	Restricted list of appropriate nouns and verbs		Object classes	Restricted list of appropriate nouns and verbs			Sub-Object-classes
Animal	حمام	عصفور	Bird				
	طير	طَائِر / طيور					
	زاحف	زواحف	Rodents				
	قواضم	قوارض	Reptiles				
	ماء	البِحَار	Aquatic-animals	سمك	أسماك		Fish
				حِيتَان	حوت	الثَّدْيِيَّاتِ	Pisces
	بحري	مياه		القِشْرِيَّاتِ	رَخْوِيَّةٍ	مَحَارِيَّاتِ	Others-aquatic-animals

The application of the process of the refined enrichment by using the restricted list of verbs and nouns (table5 in yellow) on the last extracted fragment used in the SubjectField based enrichment (figure 6) can give the enrichment presented in the following figure 7.

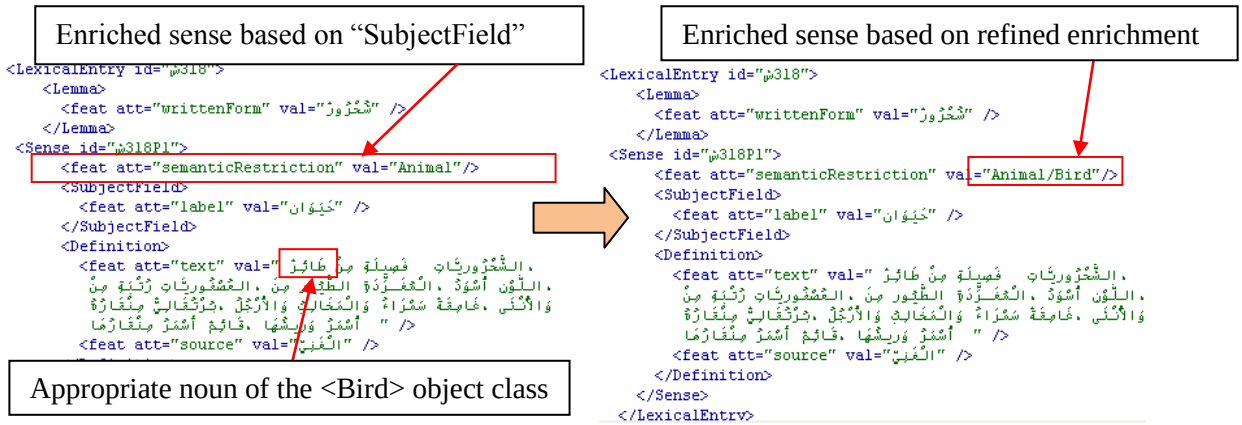

Figure 7: The refined enrichment applied to a sense of lexical entry

In the following, we present an experimentation of the exhaustive enrichment using the appropriate verbs and nouns applied to non-enriched senses. The analysis of *Contexts* and *Definitions* of senses related to a lexical entry in the Arabic LMF standardized dictionary by using the appropriate verbs and nouns (table4) can identify the relevant semantic class.

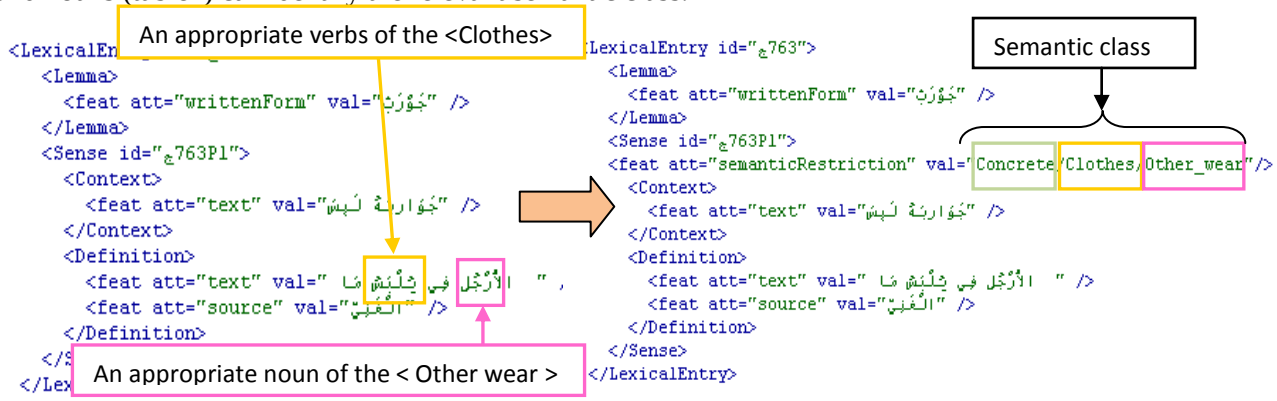

Figure 8: The exhaustive enrichment applied to a sense of lexical entry

5.4. Results

To test the performance of the carried out experimentation, we have realized a statistical evaluation.

Our Arabic standardized dictionary contains in total 34000 lexical entries including 62157 senses. Concerning the SubjectField based enrichment experimentation; we have used 4 "SubjectField" (Animal, Insect, Plant and Culinary) among the 19 available in the Arabic dictionary. And for the analysis based enrichment we have choice only the "CONCRETE" hyper class and specially the "Clothes" object class to experiment the process of this kind of enrichment.

The table 5 below gives the statistical evaluation of the "SubjectField" and the analysis based enrichment.

Table 5: Evaluation of the enrichment

		SubjectField based enrichment	**Analysis based enrichment (exhaustive step)**
Number of Subject Field	**Animal**	197	
	Insect	19	
	Plant	242	
	Culinary	39	
Total		**497**	
Correct assignment		454	90
Incorrect assignment		43	52
Recall		91,34 %	26 %
Precision		98 %	63 %

6. Conclusion and perspectives

In this paper, we have proposed an approach for the automatic enrichment of LMF standardized dictionaries with semantic classes. This approach is composed of a semantic classification based on the Gaston Gross studies and two types of enrichment. The first enrichment named SubjectField based enrichment, takes advantages from the structure of an LMF dictionary where meanings contain the domain of use of a lexical entry. The second enrichment called analysis based enrichment, uses the features of the proposed semantic classification based on appropriates verbs and nouns specifying each semantic class and applied to the available text components in the dictionary.

We carried out experimentation, by using an available Arabic standardized dictionary. The obtained results are satisfying concerning the SubjectField based enrichment. The synonymy based enrichment can reduce the enrichment effort at thirds because on average, the synonymy relation connects three or more senses.

In the future, we opted to achieve the experimentation on the others semantic classes of the proposed semantic classification for Arabic language and to complete the rest of SubjectField existed in the Arabic LMF standardized dictionary. In addition, we consider improving the analysis based enrichment by adding more efficient syntactic-semantic analysis. Finally, we foresee that the enrichment can offer the flexibility to create new oriented versions of the semantic knowledge needed for different NPL applications.

Reference

Cohen, D. (1996). Law, social policy, and violence: The impact of regional cultures. *Journal of personality and Social Psychology*, 70.961-978.

Condamines A. (2005). Sémantique et corpus, quelles rencontres possibles ? Sémantique et corpus, A. Condamines, éd., Paris : Hermès.

Dichy, J. (2000). Morphosyntactic Specifieers to be associated to Arabic lexical Entries- Methodological and Theoretical Aspects. *Proceddings of ACIDA'2000. Monastir, Tunisia, 22-24 March 2000. Corpora ans Natural Language Processing. Volume, p.55-60.*

Dorr, B. (1997). Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–325.

Dubois, J. & Dubois-Charlier, F. (1997). Les verbes français (LVF) . *Jean Dubois et Dubois-Charlier Françoise Diffuseur exclusif. Larousse Bordas, Paris.*

Dziczkowski, G. & Wegrzyn-Wolska, K. An autonomous system designed for automatic detection and rating of film reviews. Extraction and linguistic analysis of sentiments. *In IEEE/WIC/ACM International Conferences on Web Intelligence*. WI'08, Australie.

Fellbaum, C. (1999). The organization of verbs and verb concepts in a semantic net. In P. Saint-Dizier, editor, *Predicative Forms in Natural Language and in Lexical Knowledge Bases*, pages 93–110. Kluwer Academic Publishers, Netherlands.

Fillmore, C.J. (1985). Frame and the semantics of understanding.Quaderni di semantica, 6(2):222–254.

Francopoulo, G. & George, M. (2008). ISO/TC 37/SC 4 Rev.16. *Language resource management- Lexical markup framework (LMF).*

Fuchs, C., Habert, B., éd. (2004). Traitement automatique et ressources numérisées pour le français, Le français moderne, Vol. 72, n°1.

Gross, G. (1994). Classes d'objets et description des verbes. *Langages*, 115, 15-30.

Gross, M. (1975). Méthodes en syntaxe : Régimes des constructions complétives. Hermann, Paris, France.

Habash, N., Soudi, A., Buckwalter, T., et al. (2007*). In Arabic Computational Morphology: Knowledge-based and Empirical Methods*. ISBN: 978-1-4020-6045-8

Hatzivassiloglou, V., McKeown, K. (1997). Predicting the Semantic Orientation of Adjectives. *ACL '98 Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.* Pages 174-181.

Jackendoff, R. (1990). *Semantic Structures*. MIT Press, Cambridge, Massachusetts.

Khemakhem, A., Gargouri, B., Haddar K. Ben Hamadou, A, et al. (2013). LMF for Arabic, chapter in the book *"LMF:Lexical Markup Framework".* Wiley Editions, ISBN: 9781848214309, pp.83-96, March 2013.

Kipper, K., Dang, H. T., Palmer, M., et al. (2000). Classbased construction of a verb lexicon. In *Proc. of the 17th National Conference on Artificial Intelligence*. Austin, TX.

Levin, B. (1993). *English Verb Classes and Alternations*. Chicago University Press, Chicago.

Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, Massachusetts.

Rastier, F. (2001). Arts et sciences du texte. Paris : PUF.

Schapire, R. E. & Singer, Y. (2000). BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, 39, 135–168, 2000

Stede, M. (1998). A generative perspective on verb alterntions. *Computational Linguistics*, 24(3):401–430.

Valette, M., Estacio-Moreno, A., PetitJean, E., Jacquey, E. (2006). Eléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens. Verbum ex machina, Actes de la 13ème conférence sur le traitement automatique des langues naturelles (TALN 06). Piet Mertens, Cédrick Fairon, Anne Dister, Patrick Watrin (éds).

Wilson, G.V., Gorda, B., Lu, P., et al. (1994). Twelve Ways to Make Sure Your Parallel Programming System Doesn't Make Others Look Bad. *IEEE Computer*, 27(10), 1994.