doi:10.2498/cit.1002234

Semantic and Contextual Knowledge Representation for Lexical Disambiguation: Case of Arabic-French Query Translation

Souheyl Mallat[1], Mohamed Achraf Ben Mohamed[1], Emna Hkiri[1], Anis Zouaghi[2] and Mounir Zrigui[1]

[1] Department of Computer Sciences, University of Monastir, Tunisia, LATICE Laboratory Research

[2] Department of Computer Sciences, Higher Institute of Applied Science and Technologies Sousse, Tunisia, LATICE Laboratory Research

We present in this paper, an automatic query translation system in cross-language information retrieval (Arabic-French). For the lexical disambiguation, our system combines between two resources: a bilingual dictionary and a parallel corpus. To select the best translation, our method is based on a correspondence measure between two semantic networks. The first one represents the senses of ambiguous terms of the query. The second one is a semantic network contextually enriched, representing the collection of sentences responding to the query. This collection forms the knowledge base of our disambiguation method and it is obtained by alignment with the relevant sentences in Arabic. The evaluation of the proposed system shows the advantage of the contextual enrichment on the quality of the translation. We obtained a high precision, relatively proportional to the precision provided by the used alignment. Finally, our translation demonstrates its potential by comparing its Bleu score with that of Google translate.

Keywords: cross-language information retrieval systems, machine translation, lexical disambiguation, semantic and conceptual indexing, contextual relations, matching, automatic evaluation metrics

1. Introduction

Today a staggering number of multilingual documents of different kinds are available on the web, which has necessitated the implementation of multilingual information retrieval systems (IRS). Such systems are generally used to retrieve documents written in one or more languages different from that used for the formulation of the query; this is called the cross-language information retrieval (CLIR). In this work, the user submitted his Arabic query and seeks to retrieve responding documents in French.

With Cross Lingual Information Retrieval (CLIR), unlike monolingual IR, we cannot evaluate the adequacy of given result by merely applying a similarity function to queries and documents based, for example, on the vector model [29]. Because the central problem in CLIR is how to retrieve relevant documents in the target language, most CLIR systems include an automatic translation module that is applied to documents and/or queries in order to bring both into a single repository. Here we present an overview of the different approaches:

- Approach based on the translation of documents translates all documents to the language of the query. Its main advantage is to provide a high retrieval precision [39].
- Approach based on the translation of the query and documents: It is based on a heavy linguistic analysis because it must translate documents and queries to pivot language [7] [10] [36]. The disadvantage of this approach is that it requires some resources such as thesaurus, which are not always available (e.g. EuroWordNet).
- Approach based on query's translation: the query in the language source is translated

to many target languages, and its different translations are sent to the search engine [10]; [24]. Currently, most of the work in the area of multilingual information research focuses on the translation of the query. This translation is cheaper than the translation of all the documents in the collection [40] [11].

In our work, we are interested in the translation of the query; these systems are based on the translation of key and isolated terms [17]. This causes serious problems at the performance level, mainly due to their inability to solve lexical ambiguities that characterize natural languages.

In this context, our contribution is the proposition of a method for lexical disambiguation associated with the query translation system (Ar-Fr), based on the use of semantic networks [19] [30]. Semantic networks have been used in various applications of NLP [3] such as parts of speech labeling [18], information extraction [14], automatic summarization [20], etc. The semantic relations (relations of taxonomy, synonymy, etc.) between the concepts are extracted from French EuroWordNet [49]. As for the contextual relations, they are obtained from semantic association rules generated by the Apriori algorithm [1] [10]. This type of relations ensures full lexical and semantic coverage of the collection of sentences in the target language responding to the need of information expressed by the query. We build two semantic networks. The first represents the ambiguous terms of the translated query and the second represents the knowledge base (KB) represented by a list of relevant sentences (KB ($\mathrm{list}_{\mathrm{RSF}}$)).

Concerning the construction of the collection of French sentences KB ($\mathrm{list}_{\mathrm{RSF}}$) corresponding to the list of relevant sentences in Arabic KB ($\mathrm{list}_{\mathrm{RSA}}$) presented in the work [31] [32] [33], it was obtained with the MkAlign tool [22]. After this step, a conceptual and semantic indexing method is applied to this KB ($\mathrm{list}_{\mathrm{RSF}}$) for the construction of conceptual index (representative concepts of the KB ($\mathrm{list}_{\mathrm{RSF}}$) with reduced size, based on the lexical French resource EuroWordNet). Finally, a matching mechanism selects one of the networks which is the most similar to the KB ($\mathrm{list}_{\mathrm{RSF}}$). This corresponds to the best translation associated to one or more ambiguous words of the query.

This article is organized into six sections. Section 2 presents a state of art of translation approaches in the domain of cross-language information retrieval, also the problem of lexical ambiguity in queries translation and gives the state of art on existing disambiguation methods and limitations. Section 3 includes the steps of our automatic query translation system (Arabic-French), and details the proposed method of lexical disambiguation. Section 4 presents experiments and evaluations conducted on a collection of queries using the "Monde Diplomatique" Corpus (MD), which is a parallel aligned corpus (Arabic-French) of the ARCADEII campaign (Concerted Research Action on Alignment and Evaluation Documents). In Section 5 we make a comparison between our translation system and Google translation system. Finally, the last section concludes our paper and presents our future work.

2. State of Art of Query Translation and Presentation of Our Disambiguation Method

For the query translation various approaches exist, such as automatic translation, translation based on predefined vocabulary, on aligned corpus, on dictionaries and translation based on the disambiguation of the translations. These approaches are presented below:

2.1. Approach Based on Automatic Translations

This type of query translation requires the use of automatic translator to translate the query or the collection of documents so that both of them are in the same language, with or without the assistance of an expert. These automatic translators allow saving time by avoiding recourse to dictionaries and massive encyclopedias. The most used softwares are Systran[1] Power Translator Pro[2].

Yamabana [52] worked on automatic translation of queries, but it showed lower performance

[1] http://systran-office-translator.software.informer.com/

[2] http://power-translator-pro.software.informer.com/

than other approaches. This is due to the fact that the query includes ambiguous words. In this case, automatic translators do not produce good translations [41].

2.2. Approaches Based on Predefined Vocabulary

This approach is based on the use of controlled vocabularies, represented by multilingual thesaurus. This approach aims to represent documents and queries by a list of classes based on an indexing method. Thus, information retrieval is therefore to retrieve documents expressed in different languages and indexed by the list of classes representing the query.

In this context, the first work based on a predefined multilingual thesaurus was the work of Salton (1970). In his experiment, the author used a list of concepts expressed in English with their translation into German and a bilingual corpus of abstracts (English-German). He showed that the average precision was approximately 95% in terms of performance, compared to the results of the monolingual thesaurus.

The main problems related to this approach are the ambiguity and incomplete coverage: the vocabulary is fixed; it is likely that it is not exhaustive, compared to the contents of documents. Even with a predefined language, some technical terms are probably absent, which affects the search results.

2.3. Approaches Based on Aligned Corpus

Alignment can be parallel or comparable. The approach based on aligned corpus provides translations of terms related to the topic of the query, and does not correspond to translation by word of the terms [45].

In order to succeed, the corpus should be parallel and aligned by sentence. Then, the system creates a global representation to translate a term into a set of terms that have a high probability of translation in the target language according to the position of words in sentences. This is done by IBM statistical translation models. These models attempt to calculate the conditional probability $p(fj|ei)$ between words ei and fj.

2.4. Approaches Based on Dictionaries

These approaches offer translation by words without worrying about syntax; this is done through a machine-readable dictionary (MRD). These approaches are not entirely satisfactory because of ambiguities of terms in the source language. Indeed, the dictionary does not contain all the query words because the user is able to derive words in many forms. Terms that are semantically ambiguous have many possible translations into target languages (synonyms ambiguities and ambiguities of polysemous terms) [2] [53].

2.5. Approach Based on the Disambiguation of Translation

A significant number of terms constituting the Arabic query may have several interpretations (polysemous words and homonyms) [6] [46] [28]. The sense of these ambiguous words is determined by the context of their occurrence. For example, "ذهب" may have translations as "gold" or "go", and the word "قانون" can have translations as "law", "rule" or "canon" (musical instrument). These ambiguities of translation cause the recuperation of documents that do not match the query.

2.5.1. Strategy of queries disambiguation in cross-language information retrieval

Among the methods of queries disambiguation for MT systems, we find methods based on: corpus analysis, construction rules and lexical resources (dictionaries and generative lexical resources). Methods based on the analysis of corpus adapt well to the development of statistical models based on the study of frequencies in texts.

Methods of disambiguation based on lexical resources (electronic dictionaries, generative lexical resources): the methods based on electronic dictionaries have two major defects: rigorous information is not easily available and dictionaries have large inconsistencies. Mihalcea [34] tried to improve this type of method by disambiguating the senses of words using a statistical classification tool. Various studies attempt to use generative resources WordNet [21] to perform lexical disambiguation. However, the absence of generative dictionary for the French language and the lack of information for the enrichment of a semantic network led us to replace the missing resources.

Our method combines several techniques mentioned above to provide an original solution for the query translation and aims to improve the results provided by the existing methods of lexical disambiguation. We want to resolve the ambiguities of words related to the context by a disambiguation process based on French lexical resource EuroWordNet and Apriori algorithm. This algorithm is applied on an indexed knowledge base to provide a set of semantic association rules. These rules rise lexical coverage in terms of wealth of information and express existing semantic and contextual relations between the terms. These relations are mainly used to build a semantic network contextually enriched to develop a matching mechanism with semantic networks associated with ambiguous query terms. This method strengthens our translation system, by eliminating the translated terms by other senses that do not belong to the semantic context of KB (list$_{RSF}$). In addition, we have enriched the user query in the source language by meaningful terms in order to increase the precision in multilingual search [31] [32]. As part of our work, the issue of selecting the best translation for query ambiguous terms can be conceived in two ways:

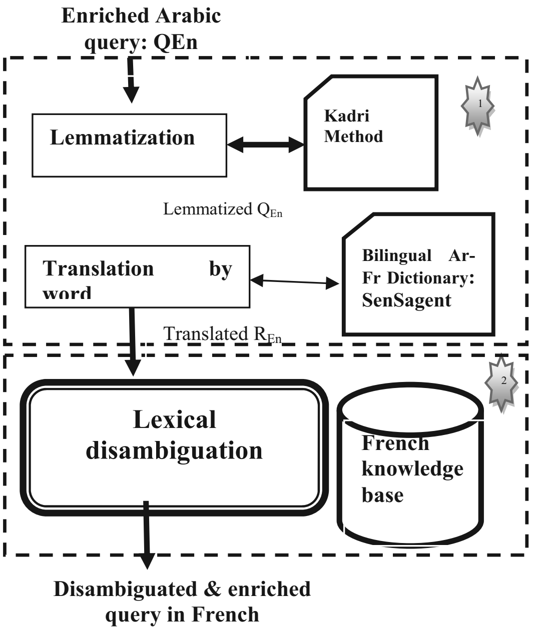


Figure 1. Functional architecture of the query translation system (Arabic-French).

1. First, the best translation is selected according to the semantic context in which the terms appear;
2. The second step is selection of the best synonym, because it is rare to find perfect synonyms that substitute each other in any context.

3. Architecture of Our Translation System

Figure 1 shows the steps of the proposed translation system.

Our system takes as input an enriched query generated by the enrichment method described in [31] [32]. This method essentially includes a double enrichment (linguistic and contextual). The first one is based on different types of linguistic analysis (lemmatization, morphological, syntactic and semantic), whose goal is to generate a descriptive list (list-desc) containing a set of language lexicons assigned to each significant term in the query. The second enrichment consists in integrating contextual information derived from documents of the corpus. This second one uses statistical analysis by the weighting functions of Salton (TF-IDF and TF-IEF).

The TF-IDF function is applied between the list-desc and documents of the corpus, it identifies relevant documents. TF-IEF function is between the list-desc and the sentences belonging to relevant documents. The role of this function is to identify relevant sentences, and then the words in these sentences are weighted. The terms of the highest weights are considered rich in terms of informative and contextual importance and they are added to the original query. In our work, the enrichment improved the performance of the research with a precision of 81% and a recall of 77% [31] [32].

As presented in Figure 1, translation of the Arabic query to French involves two main phases: lemmatization and translation, and the disambiguation of the enriched query.

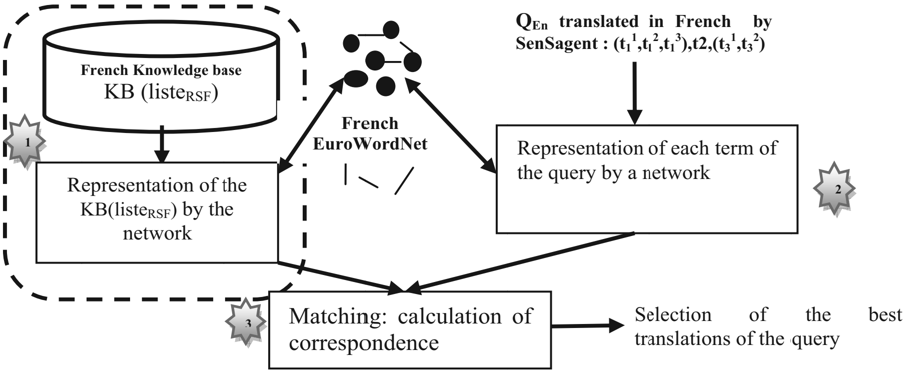

French Knowledge base
KB (liste$_{RSF}$)
1
Representation of the KB(liste$_{RSF}$) by the network
French EuroWordNet
Q$_{En}$ translated in French by SenSagent : $(t_1^1,t_1^2,t_1^3),t2,(t_3^1,t_3^2)$
Representation of each term of the query by a network
2
3
Matching: calculation of correspondence
Selection of the best translations of the query

Figure 2. Proposed disambiguation method.

3.1. Lemmatization and Translation of Arabic Enriched Query

A word is composed of a base (verb, noun) to which affixes (prefixes and suffixes) and clitics (proclitics and enclitics) are clumped together. For the lemmatization step we adopted the approach of Kadri [26]; it consists in truncating some prefixes (ال. اللك. وال. فب. وبال. فل. ولل. كال. فال. ول. وب.بال. ل. ب.. و) which are nothing more than prepositions attached to words, and some suffixes (ات. آنن. اتي. هما. وا. ك. نا. هم.ون. و. ين. ها. ت. ي..) which are generally pronouns granted to the end of words. For the translation step we used the online multilingual dictionary SenSagent[3]; it uses data available on the Web, such as Wikipedia, and incorporates the process of transliteration of named entities from the Arabic into French, among other sources. Sensagent is used in several studies (translation, multilingual research, etc).

3.2. Lexical Disambiguation of the Enriched Query

Figure 2 shows the architecture of our disambiguation method. The process of lexical disambiguation is based on three important steps, which are:

1. construction and representation of the knowledge base KB (list$_{RSF}$);
2. representation of the query by a network;
3. measuring of the match between them.

Figure 3 shows the detailed architecture of our lexical disambiguation method.

3.2.1. Construction and representation of the knowledge base from a parallel corpus

Meillet [35] defended the idea of contextual conceptualization of the meaning and assumed that a word has no sense by itself, but only in a context: "The sense of a word is defined by a means between the linguistic uses". This hypothesis is the base, in particular, for empirical approaches that acquire knowledge based on the senses of words from large corpus.

Therefore our disambiguation process requires construction of a knowledge base in the target language:

1. Construction of the knowledge base KB (list$_{RSF}$): The construction is based on Arabic relevant sentences and documents that were used during the enrichment step [31] [32] colored in blue. The construction of the knowledge base is presented in the Figure 4.
   The method of building the knowledge base KB (list$_{RSF}$) is detailed as follows:

- Extraction of relevant documents in French: The alignment of the relevant documents retrieves French documents corresponding to Arabic relevant documents extracted in our previous work (see Section 3). In this work, we used the MkAlign an alignment tool, based on bilingual parallel corpus (Ar-Fr), which takes as input a collection of relevant documents in Arabic in de-

[3] http://dictionnaire.sensagent.com

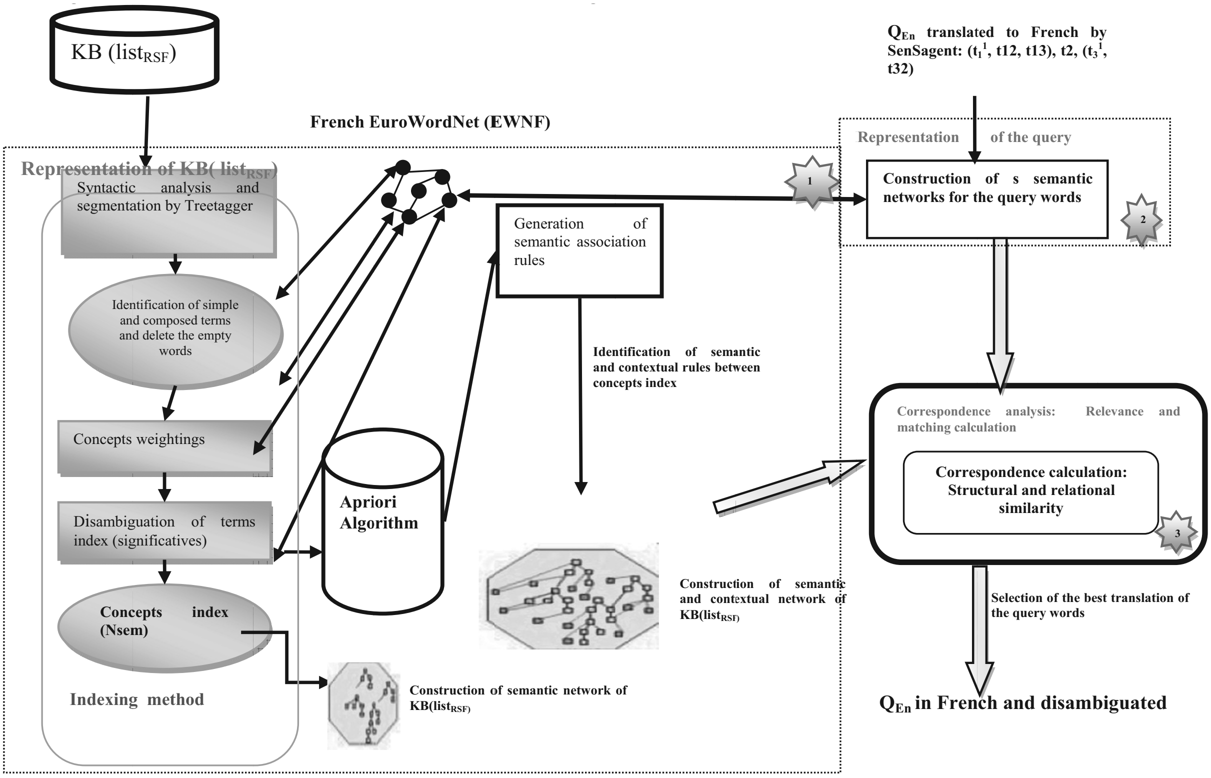

KB (list$_{RSF}$)
French EuroWordNet (EWNF)
Q_{En} translated to French by SenSagent: (t_1^1, t12, t13), t2, (t_3^1, t32)
Representation of KB(list$_{RSF}$)
Syntactic analysis and segmentation by Treetagger
Identification of simple and composed terms and delete the empty words
Concepts weightings
Disambiguation of terms index (significatives)
Concepts index (Nsem)
Indexing method
Apriori Algorithm
Generation of semantic association rules
Identification of semantic and contextual rules between concepts index
Construction of semantic and contextual network of KB(list$_{RSF}$)
Construction of semantic network of KB(list$_{RSF}$)
Representation of the query
Construction of s semantic networks for the query words
1
2
3
Correspondence analysis: Relevance and matching calculation
Correspondence calculation: Structural and relational similarity
Selection of the best translation of the query words
Q_{En} in French and disambiguated

Figure 3. Detailed architecture of our lexical disambiguation method.

scending order, and outputs a collection of French documents. It is based on a function $P(D_{Ar}|D_{Fr})$ expressing the probability of a document in French (D_{Fr}) to be the translation of a document in Arabic (D_{Ar}).

In our case, it is not possible to build a network for the entire contents of relevant documents because a document is too large in terms of words, and does not fully respond to all the terms of the user query. Reducing terms saves time in the calculation of term frequency, and makes the response time of our translation system faster. For this reason, we will choose the sentence as an atomic unit.

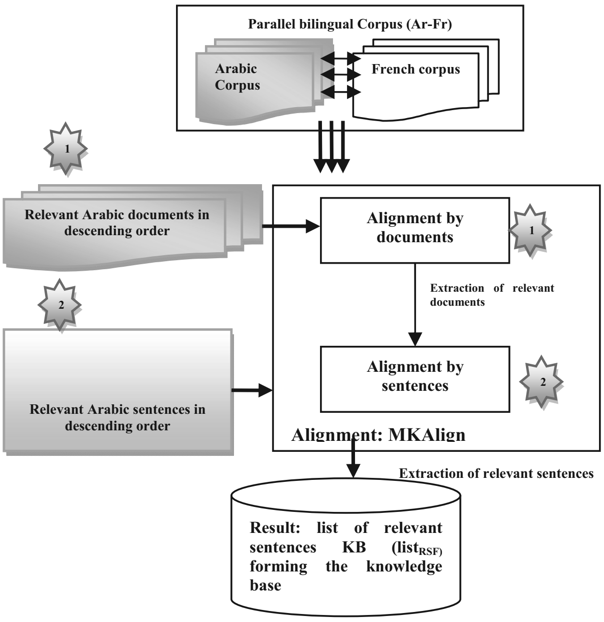

Parallel bilingual Corpus (Ar-Fr)
Arabic Corpus
French corpus
1
Relevant Arabic documents in descending order
2
Relevant Arabic sentences in descending order
Alignment by documents
1
Extraction of relevant documents
Alignment by sentences
2
Alignment: MKAlign
Extraction of relevant sentences
Result: list of relevant sentences KB (list$_{RSF)}$ forming the knowledge base

Figure 4. The construction of the knowledge base.

- Extraction of relevant sentences in French: the alignment at the sentence level is also done by MKalign tool in order to extract French sentences corresponding to the relevant sentences in Arabic. We also consider the similarity between the recovered sentence (and score of a sentence) and the user query. Indeed, the list of returned sentences is ordered according to this. In this case, we proceed with an important step to classify Arabic sentences by their degree of semantic similarity with the query terms.

The result of the alignment is the KB (list$_{RSF}$) list, which is a sort of table of contents of relevant French sentences, that responds contextually and semantically to different significant terms of the query.

2. Representation of the knowledge base KB (list_{RSF}) by a network: Our disambiguation method is based on the analysis of the corpu; we are interested to develop a formal model that describes the KB (list_{RSF}). A number of studies within NLP have exploited networks of lexical collocations (semantic, syntactic and pragmatic) built according to the principles presented by Church [16]. Such networks have the advantage of being easy to construct automatically, therefore we treat the KB (list_{RSF}) without being limited to a particular theme.

Our network is essentially composed of the set of concepts associated with significant terms identified from the KB (list_{RSF}). So, we propose in the first step a method of representation of the KB (list_{RSF}) in the form of a semantic network. This representation is based on an indexing method. After the indexing, we build the network (identification of the nodes and the relations between them). In the second step, we are interested in enriching this semantic network by adding other hidden relations compared to the reference network.

Indexing of KB (list_{RSF})

In creating our indexing method, we are inspired by Baziz work, in order to represent the KB (list_{RSF}) by a list of index concepts. It is based on the combination of semantic and conceptual indexing [5] [54] [27].

In the semantic indexing, the used semantic structure makes possible the extension of the representation of the KB (list_{RSF}) by the relation of synonymy. However, the conceptual indexing may be considered as a generalization of the semantic indexing in so far as the concepts convey senses. Baziz shows that this method ameliorates the quality of the system contrary to an indexing based only on conceptual indexing. He demonstrates that the IR system performs better with this combination, since it has produced less than 30% of disambiguation errors.

For this reason our indexing method ensures the proper functioning of the lexical disambiguation. This method is based on the use of the semantic network French EuroWordNet (EWNF). This network contains nodes (concepts) (see Figure 5), which are composed of a set of synonyms (synsets). A synset is defined by its relations with neighboring senses (for example hypernymy, hyponymy and meronymy relations).

Synsets	**Meaning of words**	**Semantic relations**		
		Hypernomy	Meronymy	Hyponymy
22745	32809	22757	1418	1408

Table 1. Statistics on the number of synsets, senses of words, of semantic relations in EWNF.

The method of indexing is based on three main steps: syntactic analysis and segmentation by Treetagger, concepts weighting and semantic disambiguation of index terms.

1. Extraction of terms from the significant terms (simple and composed) of the semantic content of the KB (list_{RSF}) is done by projection on EWNF. If this projection generates for a given term several corresponding concepts, then this term will be disambiguated. The identification of composed terms in the list is a very interesting way to improve the performance of automatic indexing. The use of these composed terms in the list of sentences considerably reduces the ambiguity of the terms and increases precision (reduces the number of senses of a term). For example, the composed term “North America” takes one sense, with 6 sense for term north, and 3 for America returned by EWNF. Our method for identifying simple and composed terms is based on a symbolic method, which is similar to that presented in [38], in which they define patterns for extracting noun syntagms. Our proposed method requires a morphosyntactic analysis of KB (list_{RSF}), as a preliminary step to extract simple and composed terms. We use an analysis that is obtained by integrating TreeTagger Helmut [46]. The analysis provided by TreeTagger can produce a list of words labeled by their grammatical categories. And like most composed terms consist of combinations of nouns, adjectives and prepositions, we generate a list of n-grams ($2 \leq n \leq 3$), respecting the following patterns:

Noun+adj: example: “champ sémantique” (engl. “Semantic field”), “définition lexicale” (engl. lexical definition)
Noun+Noun: “roi ban“(engl. “King Ban”)
Noun+prep+noun: “partie du discours” (engl. “part of speech”), “dictionnaire de langue” (engl. “language dictionary”)

A set of 1630 simple and composed terms is extracted from the KB (list$_{\text{RSF}}$) of the MD corpus.

Subsequently, identified terms are projected on the lexical database EWNF in order to remove empty terms. First, all non significant words are removed. In practice, elimination of these words can reduce the time of the weighting process; they will not be taken into consideration when calculating the frequency of words distribution.

2. *Terms weighting:* Once the simple and composed terms are extracted from the KB (list$_{\text{RSF}}$), we assign to each one of them a weight in the KB (list$_{\text{RSF}}$). The purpose of this step is to eliminate the least frequent terms and maintain only the most representative terms in the list$_{\text{RSF}}$. In general, IRS uses the weighting method TF.IEF. As TF gives the number of occurrences of a term KB (list$_{\text{RSF}}$), and IEF gives the inverse of the frequency of sentences, it is based on the number of matched sentences by the term in question.

   However, the results obtained by this function have not been satisfactory because many of the terms were not recognized due to several ambiguities of lexical and semantic variation. The main disadvantage of this weighting method is that it considers only the occurrences of concepts in the KB (list$_{\text{RSF}}$), ignoring the existing semantic relations between them. To overcome this problem, we proposed a weighting method, which combines statistical and semantic analysis [25], for assigning weight to the terms of KB (list$_{\text{RSF}}$) optimally in terms of the frequency of each with their semantic variations.

   For the statistical analysis: in the step of concepts identification, we are interested in the importance of composed terms but in some cases, the words composing these terms can refer to them even when used alone, after a number of occurrences. This represents a form of simplification or abbreviation used by the author. Let Ti be a term, its frequency depends on the number of occurrences of the term itself, and the words that compose (or sub-term (STi)). Statistical analysis is defined by the conceptual frequency of a term Ti for the KB (list$_{\text{RSF}}$), it is calculated as follows:

$$CF(Ti) = count(Ti) + \sum_{ST \in Ti} \left(\frac{Length(STi)}{Length(Ti)} \cdot count(STi) \right) \quad (1)$$

With Length (STi) represents the number of words in Ti and STi represents the sub-terms (single words) derivatives of Ti.

The semantic analysis is based on the representativeness of a concept, which takes into account the frequency of occurrence of terms, denoting the concept in the KB (list$_{\text{RSF}}$), but also its relations with other concepts in the domain. The more relations with other concepts present in the KB (list$_{\text{RSF}}$) a concept has, the more is this concept a representative of the KB (list$_{\text{RSF}}$). The EWNF resource is used to generate the set of concepts related to these terms in the form of synset taking every defined sense, and its semantic relations. The basic relation between the terms of the same synset is synonymy, but different synsets are otherwise related by various semantic relations such as subsumption, or hyponymy/hypernymy. In our case, we used the weighting method of semantic frequency of the term W_frqsem (Ti), which is calculated for each term in function of:

- the frequency of occurrence of the concepts associated with that term
- the ranks of sentences to which those concepts belong.

The coefficients corresponding to each sentence are assigned as follows: if a term belongs to the first sentence, its coefficient is 10, 9 for the second, and 1 for the tenth and the rest of the sentences in the KB (list$_{\text{RSF}}$). Assuming that the term Ti containing n terms appears p times in the KB (list$_{\text{RSF}}$), Mi, j is the coefficient for the sentences containing the conceptual occurrence j of the term Ti (different senses associated with this term, extracted from a EWNF, and with each sense of this term a synset is associated, as well as all its semantic relations). The weight of semantic frequency W_freqsem of a term Ti in the KB (list_{RSF}) is calculated as follows:

$$W\_freqsem(Ti) = \frac{P(Ti)}{maxp = 1, \ldots, n(P(Tp)) * ns} \quad (2)$$

where $P(Ti) = \sum_{j=1}^{K} (Mi, j)$ is the weight of term Ti, and $Ns = k$ – number $(Mi,\ j = 0)$ with (ns presents the number of possible senses of Ti).

$W(Ti,$ KB (list_{RSF})) represents the global weight of a term Ti in the KB (list_{RSF}), and is defined by the expression:

$$W(Ti, \text{KB } (\text{list}_{\text{RSF}})) = WTi = CF(Ti) * W\_freqsem(Ti) \quad (3)$$

The index of KB (list_{RSF}) noted Index (KB (list_{RSF}))= (Ti, WTi).

3. *Disambiguation of index terms*: The process of disambiguation is introduced to identify the exact sense of a polysemous term in the (KB (list_{RSF})).

   For an ambiguous term Ti belonging to the index (KB (list_{RSF})). Let S_i, the number of senses associated with the term Ti. The principle of the disambiguation method is to select the best concept (sense) in the (KB (list_{RSF})) from several concepts $(C1, C2, Cn)$. In the semantic disambiguation, we are interested in the method used by Baziz [4]. It is based on the calculation of a symmetric similarity weight $(P(c))$ for each concept associated with term Ti of sense j on the list of indexes: the formula is as follows:

$$P(C_i^j) = \sum_{l\epsilon[1,\ldots,m] l \neq i} \ \sum_{k\epsilon[1,\ldots,nl]} Dist(C_i^j, C_l^k) \quad (4)$$

   The concept with the highest weight is considered as the best sense of the term Ti. After extracting the concepts and calculation of their weights, the KB (list_{RSF}) will be represented by m concepts $(m \leq n)$ with their respective weights called list of indexed concepts. This list forms the semantic core, designated by Nsem (KB (list_{RSF})). We pass now to build the semantic network and to identify the relations between the nodes.

   The term Ti in EWNF, $Dist(C_i^j, C_l^k)$ is a measure of proximity between semantic concepts C_i^j and C_l^k [11]. It is calculated by a score based on their mutual distance in the network EWNF [42]. The disadvantage of this method is that it does not take into account the representativeness of terms in the context of KB(list_{RSF}). So the best sense for a term Ti in KB (list_{RSF}) must be strongly correlated to the senses associated with other important terms in KB (list_{RSF}). For this reason, we will integrate the weight of the term in the calculation of conceptual scores, using the following formula:

$$P(C_i^j) = \sum_{l\epsilon[1,\ldots,m] l \neq i} \ \sum_{k\epsilon[1,\ldots,nl]} (WC_i^j, \text{K}B(List_{RSF}) * WC_l^k, \text{KB } (\text{List}_{RSF}) * Dist(C_i^j, C_l^k)) \quad (5)$$

   The concept with the highest weight is considered as the best sense of the term Ti. After extracting the concepts and calculation of their weights, the KB (list_{RSF}) will be represented by m concepts $(m \leq n)$ with their respective weights called list of indexed concepts. This list forms the semantic core, designated by Nsem (KB (list_{RSF})). We pass now to build the semantic network and to identify the relations between the nodes.

Construction of semantic network

The semantic network is composed essentially of the semantic concepts issued from the KB (list_{RSF}) noted Nsem (KB (list_{RSF})). The network is structured in the form of (C domain (C)) by exploiting the lexical database EWNF. With C is the concept (node), and the domain (C) represents all synset S_i of Nsem (KB (list_{RSF})) (C subsumes S_i).

In WordNet, an entry is a concept represented by synset synonyms that can describe this concept. The concepts are defined as a set of lexical units related to specific domains. Let G (KB (list_{RSF}))= $\{(C$ domain $(C))\}$ represents the nodes of the semantic network of the KB (list_{RSF}), in what follows we describe the components of the semantic network, the nodes (concepts) and the semantic arcs.

1. *The concepts nodes:* The nodes represent concepts which are semantically related to different concepts $(C1, C2, C3, \ldots, Ck)$ of the Nsem(KB (list_{RSF})). The basic principles used to create the network nodes associated with the list_{RSF} are, first, a designation for each variable (instance) of Nsem(KB(list_{RSF})) by a corresponding concept from EWNF; Each concept C_i corresponds to C_i^k values in domain $(C_i) = \{C_i^1, C_i^2, C_i^3, \ldots\}$, second, each concept in Domain (C_i) is a concept $C_i^j \epsilon$ Nsem(KB (list_{RSF})) as C_i^k is -a C_i.

   The previous two principles have allowed us to build the set of nodes of the KB (list_{RSF}) nodes (KB (list_{RSF}))$= \{(c1, \text{domain}(c1)), (c2, \text{domain}(c2)), \ldots, (cn, \text{domain}(cn))\}$. The following example presents the C_i nodes, as well as domains domain (C_i) associated with the theme "Military" by the application of the previous two principles.

   Consider the following example from our corpus, which illustrates an indexed list by the following weighted concepts.

   These concepts constitute the semantic core of the KB (list_{RSF}) associated a topic "Military" using the relations of subsumption (is-a) between concepts and properties (relations domain) through EWNF. So we obtain the concepts representing the KB (list_{RSF}) that are the nodes of the network Node (KB (list_{RSF})): {*("Minister of Defence" domain ("Ministry of Defence")), ("military action" domain ("military action")), etc.*}

2. *Semantic relations between nodes (concepts):* Several types of semantics relations are proposed by the EWNF resource, such as generic-specific relations (hypernym-hyponym (is-a)), composition relations (holonymy-meronymy (part-whole)). Figure 5 shows the semantic network that illustrates the concepts with these relations in the "Military" theme.

This network includes only the close relations between semantic concept nodes. However, we observed the absence of relations with other relevant concepts that are close in the same context (victory, military operation, occupation, intervention, etc.). Indeed, the coverage of EWNF is small compared to the list of index concepts (Nsem (KB (list_{RSF}))). Using EWNF mentions the lack of useful contextual relations between relevant concepts. This lexical database contains only limited information on the use of concepts. So the network is obviously insufficient for lexical disambiguation of all existing ambiguous words in the queries. Hence, we need to increase the coverage of this network by contextual enrichment.

Concept	***W***
organisation de défense (engl. defence organisation)	0.55
établissement de défense (engl. defense constitution)	0.5
Action commando (engl. commando action)	0.35
Effort (engl. effort)	0.3
véhicule militaire (engl. Military vehicle)	0.4
véhicule de combattants (engl. vehicle of fighters)	0.25
Entités (engl. Entities)	0.12
victime(engl. victim)	0.25
blessé (engl. injured)	0.25
Panzer (engl. Panzer)	0.1
Pistolet (engl. pistol)	0.12
Tourelle (engl. turret)	0.08
balle (engl. ball)	0.09
indépendance (engl. independence)	0.6
triomphe (engl. triumph)	0.12
Réussite (engl. success)	0.2
sécurité de peuple (engl. people security)	0.4
encadrement (engl. supervision)	0.2
acquérir (engl. acquire)	0.5
obtenir (engl. get)	0.1
opération aérienne (engl. air operation)	0.6
force armée (engl. armed force)	0.8
soldats (engl. soldiers)	0.7
combattants (engl. fighters)	0.6
région montagneuse (engl. mountainous region)	0.5
Forêt (engl. forest)	0.15
ingérence (engl. interference)	0.7
négociation (engl. negotiation)	0.25
imposer (engl. impose)	0.7
Demande (engl. demand)	0.1
massif de soldat (engl. soldiers)	0.7
nombreux (engl. many)	0.2
sécurité de pays (engl. country security)	0.4
sécurité de frontière (engl. border security)	0.5

Table 2. Weighted concepts.

We pass now to the second step of the representation of the semantic network, that consists in

Concept node	Domain(concept)
organisation de défense	{Ministère de défense établissement de défense}
action militaire	{Action commando, Effort}
transporteur militaire	{véhicule militaire, véhicule de combattants, entités}
guerre	{victime, blesse}
Moyen d'attaque	{panzer, pistolet}
munitions	{tourelle, et balle}
autonomie	{indépendance}
victoire	{triomphe, réussite}
occupation	{acquérir, obtenir}
opération	{opération aérienne, encadrement}
militaire	{force armée, soldats, combattants}
nature	{région montagneuse, forêt}
intervention	{ingérence, négociation}
sécurité	{sécurité de pays, sécurité de frontière, sécurité de peuple}
nombre	{massif de soldats, nombreux}
ordre	{demande, imposer}

Table 3. Identification of concept nodes.

the enrichment by contextual relations.

Contextual enrichment of the semantic network

Our goal is to identify existing implicit contextual relations between nodes (concepts) representing Nsem (KB (list_{RSF})). We used a method based on the technique of semantic association rules which are extracted by the Apriori algorithm, for more details see [1]. The principle of association rules discovery can be presented as follows: Let $I = i_1, i_2, i_n$ a set of items and D a set of transactions where each transaction T is a set of items such that $T \subset I$.

A set of items is called an itemset. An association rule is an implication of the form $X \rightarrow Y$,

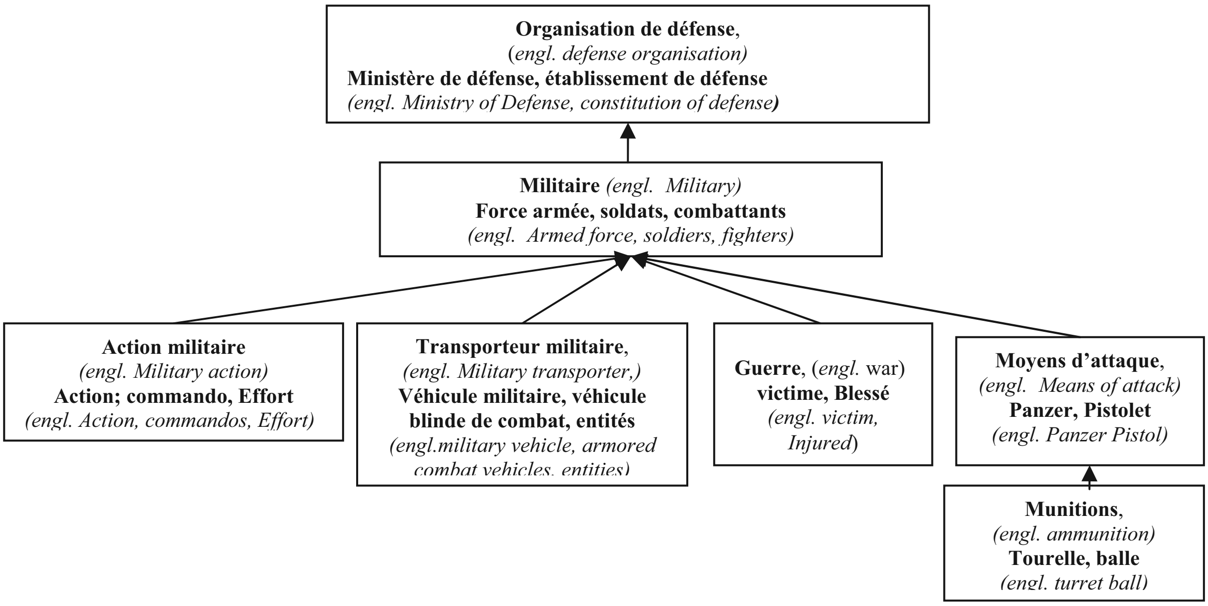

Organisation de défense,
(engl. defense organisation)
Ministère de défense, établissement de défense
(engl. Ministry of Defense, constitution of defense)
Militaire (engl. Military)
Force armée, soldats, combattants
(engl. Armed force, soldiers, fighters)
Action militaire
(engl. Military action)
Action; commando, Effort
(engl. Action, commandos, Effort)
Transporteur militaire,
(engl. Military transporter,)
Véhicule militaire, véhicule
blinde de combat, entités
(engl.military vehicle, armored
combat vehicles, entities)
Guerre, (engl. war)
victime, Blessé
(engl. victim,
Injured)
Moyens d'attaque,
(engl. Means of attack)
Panzer, Pistolet
(engl. Panzer Pistol)
Munitions,
(engl. ammunition)
Tourelle, balle
(engl. turret ball)

Figure 5. A semantic network corresponding to the theme 'Military' in EWNF.

where $X, Y \subset I$ and $X \cap Y = \emptyset$. Generally, X is called the antecedent and Y the consequent.

We apply the technique of semantic association rules in order to identify the contextual relations between nodes (concepts). In this step, we use the Apriori algorithm to extract relations (arcs). This algorithm has two steps; the first is to extract all frequent itemsets of the KB (list_{RSF}). The second step is generation of the association rules between frequent itemsets discovered during the first step. They are detailed as follows:

1. Generation of frequent itemsets is composed of these three phases:

- construction of the group E1 1-itemsets which are the most frequent concepts in the KB(list_{RSF}), which have a weight P1-itemsets greater than a given threshold.
- From the E1 of 1-itemsets frequent calculated in the previous step, we generate the set 2-itemsets of candidate in order to construct E2, which have a weight greater than a given threshold $\text{P}_{\text{2-itemset}}$. With $\text{P}_{\text{2-items}} = \text{min}_{\text{items}} = \min(\text{P}_{\text{1-itemsets}}(\text{1-itemset1}), \text{P}_{\text{1-itemsets}}\ (\text{1-itemset2}))$ (see Table 4).
- The stop condition of the algorithm is when there is no more generation of new itemsets candidate in order to return the set $E = E1 \cup E2$ of all frequent itemsets in the KB(list_{RSF}).

2. Generation of semantic association rules: after the construction of the set E corresponding to all significant itemsets in the KB (list_{RSF}), we generate the semantic association rules [1].

A semantic association rule between C and S is noted $C \rightarrow \text{sem}(S)$, and defined: $C \rightarrow \text{sem}(S_i)$, exist $C_i \in \text{Dom}(C)$, exist $Sj \in \text{Dom}(S)/C_i \rightarrow Sj$.

The rule $C_i \rightarrow S_i$ means if the KB (list_{RSF}) is linked to the concept C by the semantic relation is-a (is, about), it must have the same type (is-about of) with the concept S. So, the rule $R : C_i \rightarrow S_i$: represents the probability that the semantic content of KB (list_{RSF}) covers S_i, knowing that it also covers C_i. This semantic interpretation is based on two metrics which are: the confidence(conf) and the support (Sup).

The confidence associated to the rule R : $\text{conf}(R : C_i \rightarrow S_j) = P(C_i/S_j)$ is based on the degree of importance of S_j in the KB (list_{RSF}), knowing the degree of importance of C_i in the KB (list_{RSF}). It is defined as: $\text{Conf}(C \rightarrow \text{sem}S) = \max_{i,j}(\text{conf}(R : C_i \rightarrow S_j))$ with $C_i \in \text{Dom}(C)$, $S_j \in \text{Dom}(S)$

$$\text{Conf}(R) = \frac{\min\left(WC_{i_{\text{KB (list}_{\text{RSF}})}}, WS_{j_{\text{KB (list}_{\text{RSF}})}}\right)}{WC_{i_{\text{KB (list}_{\text{RSF}})}}} \tag{6}$$

The support (Sup) is associated with a semantic association rule between entities $\text{Sup}(C_i \rightarrow \text{sem}(Sj)) = P(C_i \rightarrow S_j)$ (probability of simultaneous occurrence of C_i and S_j). It is based on the number of rules of groups $C_i \in (\text{Domain}(C_i))$ and $S_j \in \text{Domain}(S_j)$, having a support greater

2-itemset	$\text{P}_{\text{2-itemsets}}$	**Association rules:** $C_i \rightarrow S_j$
{indépendance, triomphe} (engl. independence, triumph)	0.4	R1: indépendance → triomphe R2: triomphe → indépendance
{indépendance, acquérir} (engl. independence, acquire)	0.4	R3: indépendance → acquérir R4: acquérir → indépendance
{indépendance, région montagneuse} (engl. independence, mountainous region)	0.4	R5: indépendance → région montagneuse R6: région montagneuse → indépendance
{indépendance, opération aérienne} (engl. independence; air operation)	0.4	R7: indépendance → opération aérienne R8: opération aérienne → indépendance
{indépendance, ingérence} (engl. independence; interférence)	0.4	R9: indépendance → ingérence R10: ingérence → indépendance
⋮	⋮	⋮
{triomphe, acquérir} (engl. triumph, acquire)	0.5	R: triomphe → acquérir R: acquérir → triomphe
{acquérir, opération aérienne} (engl. acquire; air operation)	0.5	R: acquérir → opération aérienne R: opération aérienne → acquérir

Table 4. Generation of association rules from 2-itemset.

than or equal to the threshold sup_{min} (minimal support). The support of a rule is as follows:

$$\text{sup}(R) = \frac{|\{C_i \rightarrow S_j/\text{conf}(C_i \rightarrow S_j) \geq \text{conf}\,\text{min}\}|}{|\{C_i \rightarrow S_j, (C_i, S_j) \in \text{Dom}(C) \times \text{Dom}(S)\}|} \tag{7}$$

However, some problems may occur during the discovery of rules such as redundancy: $C \rightarrow {}_{\text{sem}}S$, $S \rightarrow {}_{\text{sem}}X$ et $C \rightarrow {}_{\text{sem}}X$. To eliminate it, we must build a minimum coverage of all the extracted rules. There are also rules of semantic association of type: $C \rightarrow {}_{\text{sem}}S$, $S \rightarrow {}_{\text{sem}}C$ et $C \rightarrow {}_{\text{sem}}S$, $S \rightarrow {}_{\text{sem}}X$ et $X \rightarrow {}_{\text{sem}}C$. To solve this problem, we eliminate the rule with low support.

From the rules of semantic associations discussed above, we build a semantic and contextual network of indexed concepts. This network represents the contents (subject) of the (Nsem $(\text{KB}(\text{list}_{\text{RSF}}))$), and the contextual relations between them. An arc oriented from concept – node C to the concept – node S. C is the parent node of S in the network.

3.2.2. Illustrative example

Returning to the previous example, we apply the Apriori algorithm to extract the contextual relations. We only keep the concepts having a weight P1-itemsets above the threshold (0.2), for the reasons of reliability and limitation of vocabulary for network construction so the terms with weight ≤ 0.2 will be eliminated. The terms such as: “forêt”, “réussite”, “encadrement”, “négociation”, “demande”, “nombreux” (“forest”, “success”, “leadership”, “negotiation”, “application”, “numerous”) are the less frequent in the KB (list_{RSF}). The other terms are the most frequent 1-itemset in the KB (list_{RSF}), which are used to construct the 2-itemset (set of two terms). We calculate subsequently the weight of each 2-itemsets: P2-itemsets ({independence triumph}) = min(0.4, 0.6) = 0.4 . . . etc.

Similarly, we only keep 2-itemset that has P2-itemsets a higher of 0.2, which allows us to construct a set of association rules (R: $C_i \rightarrow S_j$).

We calculate the confidence for each association rule R_i: Conf (R_i: $C_i \rightarrow S_j$), we obtain the following Table 5.

We retain only the rules that have a confidence $\geq$ threshold of minConf = 1. These association rules are used to construct the semantic rules forming the basis for the identification of relation between concepts nodes KB (list_{RSF}).

The next step is the calculation of support for each semantic rule (sup (Rsemk: $C_i \rightarrow$ sem S_j)), with $k = 1, \ldots, n$ (number of semantic rules), we obtain the following Table 6.

We retain from Table 6 the rules of semantic association whose support ≥ 0.5, like Rsem2 etc. Finally, the selected rules enable the selection of semantic relations between concept nodes of the KB (list_{RSF}), in order to construct the semantic and contextual network presented by Figure 6.

In the following section we present the second phase of the lexical disambiguation which is the representation of the query term by networks.

3.2.3. Representation of the query

The objective of this phase is to represent the query in the form of a semantic network; this

R_i	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	. . .	R	R	R	R
Conf(Ri)	1	0.8	1	0.8	1	0.8	1	0.66	1	0.57	. . .	0.83	1	1	0.83

Table 5. Calculation of confidences for the generated association rules.

$\text{R}_{\text{sem}k}$	R_{sem1}	R_{sem2}	R_{sem3}	R_{sem4}	R_{sem5}	. . .	R_{sem}	R_{sem}
Sup($\text{R}_{\text{sem)}}$	1	0.5	0.5	0.5	0.5	. . .	1	1/6

Table 6. Calculation of confidence for the semantic association rules.

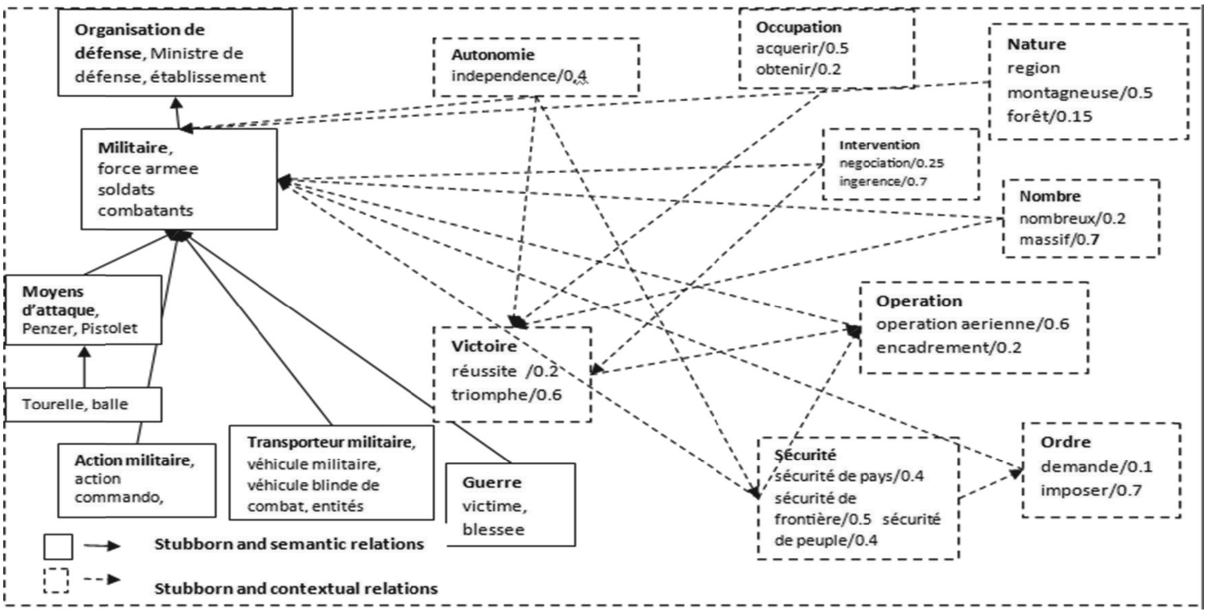


Figure 6. Example of semantic and contextual network (nodes, arcs) from the KB (list$_{\mathrm{RSF}}$).

mechanism allows expressing every word of the enriched query in the form of the semantic network based on the resource EWNF. The following example (Figure 7) presents a semantic network generated by EWNF for an ambiguous word of the Arabic query "تدخل", translated by SenSagent dictionary in "intervention" and "mediation".

With ***gt1***$^{\mathbf{1}}$ and ***gt1***$^{\mathbf{2}}$ are the two networks associated to the two senses "intervention", and "mediation" of the ambiguous term "تدخل".

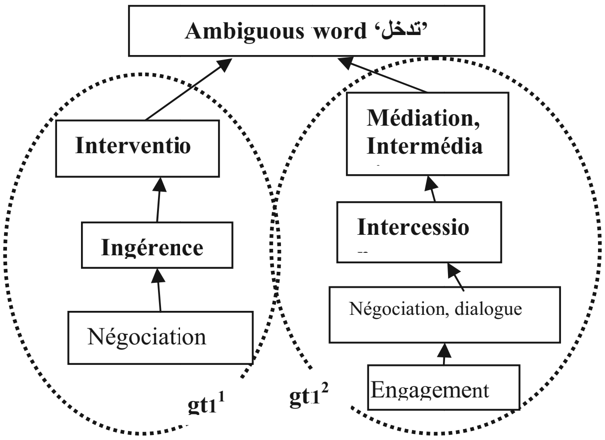


Figure 7. Representation by semantic network based EWNF of the ambiguous term "تدخل".

3.2.4. Relevance and matching between KB (list$_{\mathrm{RSF}}$) and the query

Once the networks are associated with the terms of the query, and KB (list$_{\mathrm{RSF}}$) is built, the last step of the disambiguation is to measure the relevance and matching between them. The relevance aims to select the best sense associated with an ambiguous term of the query compared with KB (list$_{\mathrm{RSF}}$).

It is based on a matching function that performs a comparison between the representatives of the KB (list$_{\mathrm{RSF}}$) by a network, and different concepts associated with query's terms built in the previous phase. Comparing means calculating the relevance of the senses of each ambiguous term according to the KB (list$_{\mathrm{RSF}}$). This value is calculated by a function of similarity noted Sim$G$ ($G_{\mathrm{KB\ (list_{RSF})}}, G_{ti}$). With $G_{\mathrm{KB\ (list_{RSF})}}$ and G_{ti} are two networks describing respectively the KB (list$_{\mathrm{RSF}}$) and a terms *ti* in the query.

The network *Gti* is represented by the set of sub-networks: $\{G_{t1}, G_{t2}, \ldots, G_{tn}\}$, with n being the number of query terms, and gti^j is a network for an ambiguous term (ti) of the j^{th} sense. It is represented by the entire sub-network $\{gti^1, gti^2, \ldots, gti^k\}$ with k being the number of concepts (sense) corresponding to the term *ti*.

In this work, the matching process relies on the comparison between the network of KB $(\text{list}_{\text{RSF}})$ and the ambiguous terms of the query; the idea is to proceed to a comparison between each network of a sense in order to select the networks which are maximally similar to the context of KB $(\text{list}_{\text{RSF}})$.

Consider the two networks gti^2 and gti^j, for us gti^2 is more similar than gti^j to the KB $(\text{list}_{\text{RSF}})$ denoted by $gti^j \supset_{\text{KB }(\text{list}_{\text{RSF}})} gti^p$ with $j \in [1,\ldots,k]$ and $p = 2$, only if the following condition is verified: $i \in [1,\ldots,n]$, $j \in [1,\ldots,k]$, and $\exists p \in [1,\ldots,k]$ with $p = 2 \neq j$ such that $\{gti^1, gti^3, \ldots, gti^k, \ldots\}$ is the set of candidates networks for each ambiguous term ti of the query:

$$\text{Sim}_G(gti^j, G_{\text{KB }(\text{list}_{\text{RSF}})}) < \text{Sim}_G(gti^p, G_{\text{KB }(\text{list}_{\text{RSF}})}).$$

More simply, the relation $gti^j \supset_{\text{KB}(\text{list}_{\text{RSF}})} gti^p$ verifies that gti^p is potentially more relevant than $gti^j \forall j \in [1,\ldots,k]$ and it is the answer network has selected and associated with the best sense. We apply the same condition for all the ambiguous terms of the query, to return all the dominant networks which define the best sense: $\{gt1^p, gt2^p, gt3^p, gtn^p\}$ according to the KB $(\text{list}_{\text{RSF}})$.

In this work, we propose a technique of matching which distinguishes nodes by their importance in the structure of the networks. This technique puts at first in correspondence the important nodes of an ambiguous term compared with the nodes of network of the KB $(\text{list}_{\text{RSF}})$. Several models were proposed [12] [13] [50] to measure the similarity between graphs, so, by analogy, the same measure is applied to networks.

In our work, we used two similarity measures adopted for the two tasks of selection; the best translation and the best synonym

1. The first task, the selection of the best translation (exact sense) according to the semantic context in which the term appears. It aims to resolve polysemy and homonyms of ambiguity, so we use structural similarity function. Let $G_{\text{KB}(\text{list}_{\text{RSF}})} \cap gti^j$ generate the Concept-Nodes (CNm), with $m \in [1,\ldots,N], N$: the number of concepts-Nodes containing common instances respectively to the $\text{KB}(\text{list}_{\text{RSF}})$ and the network of the j^{th} sense for an ambiguous term ti of the query CN, we consider the domains $(CN, \text{KB }(\text{list}_{\text{RSF}}))$, and domain (CN, Q) instances associated to existent CN respectively in $G_{\text{KB }(\text{list}_{\text{RSF}})}$ and gti^j. The structural similarity is defined by the following function:

$$\text{Sim}_{Str}^{CNm} = \frac{|\text{Domain}(CNm, \text{KB }(\text{list}_{\text{RSF}})) \cap \text{Domain}(CNm, Q)|}{|\text{Domain}(CNm, \text{KB }(\text{list}_{\text{RSF}})) \cup \text{Domain}(CNm, Q)|} = \text{Sim}_{str}(G_{\text{KB }(\text{list}_{\text{RSF}})}, gti^j) \quad (8)$$

to every j^{th} sense.

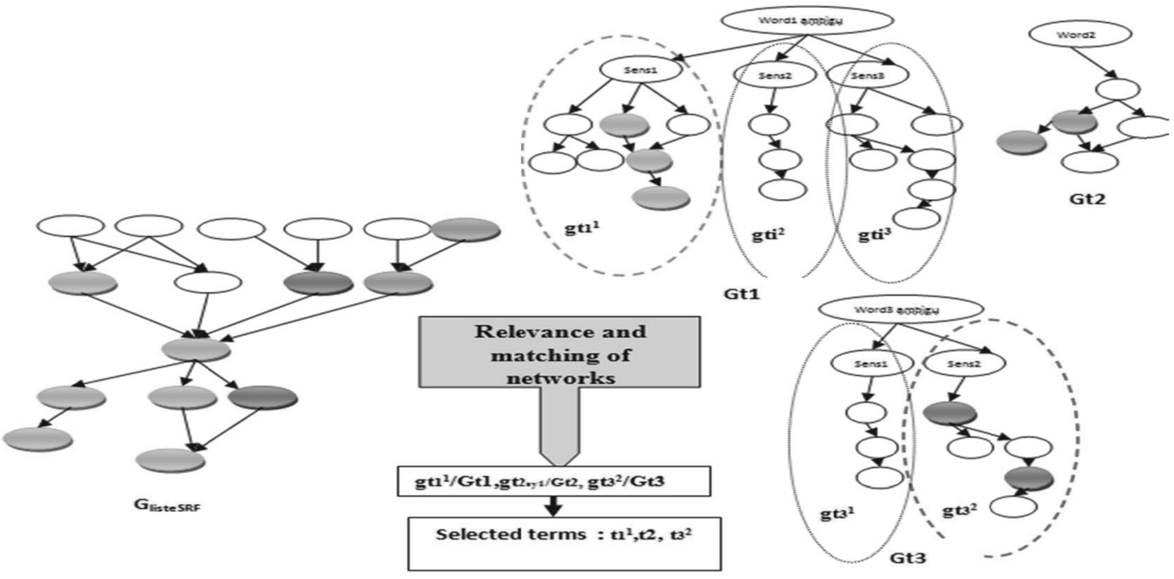


Figure 8. Matching between the network of KB $(\text{list}_{\text{RSF}})$ and those of the query terms.

Remark: if $G_{\text{KB (list}_{\text{RSF}})} \cap gti^j = \emptyset$, then $\text{Sim}_{str}(G_{\text{KB (list}_{\text{RSF}})}, gti^j) = \emptyset$.

The high weights of similarity of the j^{th} sense corresponding to ambiguous terms are returned.

2. The second task concerns selection of the best synonym because synonyms are words with very close meaning. Sometimes they share exactly the same sense; however, there are few perfect synonyms that replace one another in any context. In fact, there is generally a nuance of senses between the synonyms; there is even a difference in the construction of the sentences according to the selected synonym. A relational similarity function is supported. This function is denoted by Sim_{rel}^{Syi} (with Syi representing the synonym) and indicates the degree of representativeness of S_i corresponding to its importance in the hierarchy of the semantic network: measured in function of the depth associated to the synonym.

   Each synonym is defined by a degree Deg_{syi}, Q, as the importance level of Syi in the network. The maximum depth of the network is n. The root of the network has level 1 and the degree of elements of this level is 1. Its direct descendants are elements of level 2 with degrees equal to $1/2$ etc. The elements of the level n have the degree $1/n$. The relational similarity function is as follows:

   $$\text{Sim}_{rel}^{Syi} = W_{Syi,\text{KB (list}_{\text{RSF}})} * \text{Deg}_{syi,Q} \quad (9)$$

   With $W_{Syi,list}$, the weight associated to the synonym Syi in the list. $\text{Deg}_{syi,Q}$ is the degree of importance in the network of the selected sense. The degrees of high importance associated with synonymous of the query's terms are returned. At the end of this phase, the selected terms, allows building the query in French language.

Let's consider the following example for the matching measure between an ambiguous term and the KB (list_{RSF}) using structural and relational similarity. This measure is a mapping function between two networks:

- the semantic and contextual network $G_{\text{KB (iste}_{\text{RSF}})}$ associated with the KB (list_{RSF}) corresponding to the theme "Military" (see Figure 5),
- the semantic network associated with the ambiguous term تدخل: $g^j_{\text{تدخل}} = \{g^{\text{intervention}}_{\text{تدخل}}, g^{\text{médiation}}_{\text{تدخل}}\}$ (see Figure 7).

So $G_{\text{KB (list}_{\text{RSF}})} \cap gti^j = \{$"ingérences", "négociation"$\}$, with the concept (engl. "interference"), "ingérence" $\in$ Domain "intervention" (engl. "interference"), and (engl. "negotiation") "négociation" $\in$ Domain "intervention" and "médiation" (engl. "interference", and "mediation").

Hence, the two concepts-nodes are $NC1 =$ "intervention" and $NC2 =$ "mediation".

$$\begin{aligned}\text{Sim}_{Str}^{CN1} &= \frac{|\text{Domain}(CN1, \text{KB (list}_{\text{RSF}})) \cap \text{Domain}(CN1, Q)|}{|\text{Domain}(CN1, \text{KB (list}_{\text{RSF}})) \cup \text{Domain}(CN1, Q)|} \\ &= \frac{2}{3}\end{aligned}$$

and

$$\begin{aligned}\text{Sim}_{Str}^{CN2} &= \frac{|\text{Domain}(CN2, \text{KB (list}_{\text{RSF}})) \cap \text{Domain}(CN2, Q)|}{|\text{Domain}(CN2, \text{KB (list}_{\text{RSF}})) \cup \text{Domain}(CN2, Q)|} \\ &= \frac{1}{4}\end{aligned}$$

$\text{Sim}_{Str}^{CN1} > \text{Sim}_{Str}^{CN2}$, then the sense "intervention" (engl. interference) is selected as the best sense. The next step is selecting the best synonym of the sense "intervention" (engl. interference): among the three {"intervenir", "ingérence", "négociation"} (engl. "involving", "interference", "negotiation").

$$\begin{aligned}\text{Sim}_{rel}^{\text{intervenir}} &= W_{\text{intervenir, KB (list}_{\text{RSF}})} \\ &\quad * \text{Deg}_{\text{intervenir}} = 0,\end{aligned}$$

$$\begin{aligned}\text{Sim}_{rel}^{\text{ingérence}} &= W_{\text{ingérence, KB (list}_{\text{RSF}})} \\ &\quad * \text{Deg}_{\text{ingérence}} = 0.7 * \frac{1}{2} \\ &\quad = 0.35,\end{aligned}$$

$$\begin{aligned}\text{Sim}_{rel}^{\text{négociation}} &= W_{\text{négociation, KB (list}_{\text{RSF}})} \\ &\quad * \text{Deg}_{\text{négociation}} = 0.25 * \frac{1}{3} \\ &\quad = 0.083.\end{aligned}$$

The best synonym coresspond of sense "intervention" is "ingérence" (engl. interference). So the translation of the Arabic ambiguous term "تدخل" is "ingérence" (engl. interference).

4. Experimentation and Evaluation

4.1. Description of the Corpus and the Training Queries

In our experiment, we used the Monde Diplomatic corpus composed of newspaper articles from the web [15]. MD treats a variety of topics (geopolitics, international relations, economics, social issue, culture, etc.). This corpus is published in eight languages Arabic, French, English, Russian, Greek, Persian, Japanese, Chinese, and contains 414 articles. In our work, the languages used for the training are Arabic and French. The Extract of the corpus contains 150 articles aligned in both languages. In these articles, we selected a training corpus of 200 pairs of bilingual aligned documents of size 0.6MB. These documents represent the knowledge base which contains objects and their properties in order to build the semantic network of the selected sentences. In addition, this knowledge base provides the application of association rules between the concepts in order to extract the contextual relations, which are used to enrich the semantic network. This is presented in the following table.

	Arabic Corpus	**French Corpus**
Number of sentences	1650	1650
Number of words	45350	46350
Number of terms	36 280	37 650
Number of words/ sentences (avg)	24	30
Number of words/ sentences (avg)	19	22

Table 7. Characteristics of the training corpus (Ar-Fr).

4.2. Evaluation and Comparison between the Networks

The evaluation of the two networks (the semantic network and a semantic network with contextual enrichment) is established by comparing them to our reference network.

- Evaluation of a semantic network according to the reference network. Our evaluation relies on the comparison between the concepts of the semantic network of the Nsem (KB (list$_{\text{RSF}}$)) and those of the reference network [37]. Our reference network is developed by an expert, we asked him to develop the best representation for the KB (list$_{\text{RSF}}$). This reference network (Rref) is composed of about more than 600 nodes and 790 relations.

$$\text{Precision} = \frac{\text{number of returned relevant concepts}}{\text{number of returned concepts}} \tag{10}$$

$$\text{Recall}{=}\frac{\text{number of returned relevant concepts}}{\text{number of concepts in the network of reference}} \tag{11}$$

Global evaluation of the lexical coverage of the network built over a network of reference is based on F-measure which combines the two measures (precision and recall). This measure is defined as follows:

$$\text{F-measure} = \frac{2 * \text{Precision recall}}{\text{Precision} + \text{recall}} \tag{12}$$

Table 8 presents the different metrics of recall, precision and F-measure for a number of sentences from 10 to the complete KB (list$_{\text{RSF}}$). These sentences are classified by their order of relevance.

We can conclude that the results of recall, precision and F-measure are relatively close to the size of 40 to 60 sentences and we note a decrease of results for the 100 sentences and same for the whole KB (list$_{\text{RSF}}$). Reducing the size of the KB (list$_{\text{RSF}}$) to 60 sentences

	10S	20S	30S	40S	60S	100S	list$_{\text{RSF}}$
Recall	52.7	56.8	60.6	72.3	72.8	70.3	65.2
Precision	51.7	55.9	60.4	69.4	70.8	69.8	68.3
F-measure	52.2	56.3	60.5	70.82	71.78	70.04	66.71

Table 8. Recall, Precision, F-measure for the first evaluation.

shows a good quality of semantic core in terms of lexical coverage. As it reached a recall of 72.8%, a precision of 70.8% and an F-measure of 71.78%, the evaluation of the quality of the semantic core Nsem (KB (list_{RSF})) in function of the number of sentences in the list (60) improves the response time of the machine translation system.

- Evaluation of a semantic network with contextual enrichment according to a reference network: This evaluation is also based on the same standard measurements. The table below shows the impact of the contextual enrichment of the semantic network in terms of global coverage. Our interest is to have good score of recall and precision that allow us to fix a threshold of support and confidence. So, which number of sentences in the KB (list_{RSF}) allows us to get the best result. Table 9 presents the metrics of recall, precision and F-measure for 10 sentences to the complete KB (list_{RSF}) These sentences are classified by their semantic relevance.

We notice that the contextual enrichment of the semantic network allows adding hidden contextual links related with the variation of two parameters (support, confidence). These entities were absent in the network without enrichment. We reach in this case an F-measure from 67 to 82%, that corresponds to two threshold values of the support = 0.5, and confidence = 1. This improvement is shown in the Table 9. We note that the use of contextual links in the semantic network with these two threshold values increases the precision (80,3) and recall (81,9). This means that the contextual enrichment of this network covers almost the whole contents of the KB (list_{RSF}).

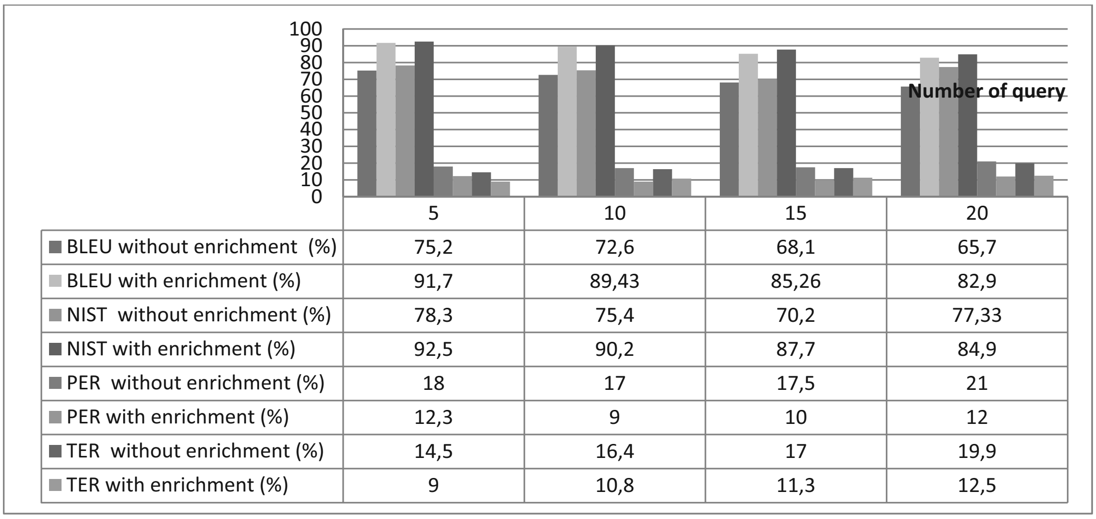
Regarding the size of the KB (list_{RSF}), it may be an obstacle for its representation by a semantic network (contextually enriched). This treatment would require a lot of memory and computation time. Indeed, several sentences are classified last in the KB (list_{RSF}), so we assume that their discrimination power is low. From the table below, we notice that with the first 90 sentences of the KB (list_{RSF}) we obtain the best precision. This reduces the noise and increases the response time of the translation system. In the training corpus, we used a set of 20 queries with key terms for experimentation. These queries include a large lexical variability and their themes are varied, giving rise to ambiguity. These queries are translated into French by our machine translation system and by our expert to get a collection of reference.

4.3. Evaluation Translation Metrics and Comparison

Most of automatic evaluation methods operate by comparing the produced query by our system (Q_s) with one or two reference queries (Q_{ref}).

		10 S	**20 S**	**30 S**	**40 S**	**60 S**	**80 S**	**90 S**	**100 S**	**KB (list_{RSF})**
Support = 0.5 Confidence = 0.5	**Recall**	55	61.4	66.7	79.5	80.2	80.9	81.4	81.4	81.4
	Precision	53	58.3	63.4	74	77.2	78.7	70.5	70.5	59
	F-measure	53.98	59.8	65	76.6	78.67	79.7	75.55	75.55	68.41
Support = 0.5 Confidence = 1	**Recall**	54	60.1	64	77.8	80	81.9	81.9	81.9	81.9
	Precision	53	59.2	65	75.4	77.6	79.4	80.3	72	61
	F-measure	53.3	59.7	64.5	76.6	78.78	80.6	81.09	76.6	70
Support = 0.4 Confidence = 0.5	**Recall**	55.7	60.8	66	78.8	81.2	82.7	82.5	82.5	82.6
	Precision	53	58	63	70.1	71.5	72.7	69.5	62	54.5
	F-measure	54	59.3	64.46	74.2	76.04	77.37	73.8	70.8	65.6
Support = 0.4 Confidence = 1	**Recall**	55	59.6	65.2	77.5	80.6	81.8	82.2	82.2	82.2
	Precision	53.7	60	65	72	74.7	76.7	77.8	65	58.8
	F-measure	54.34	59.8	65.1	74.64	77.53	79.1	80	72.7	68.55
Support = 0.3 Confidence = 0.5	**Recall**	54.1	59.6	63.5	74.6	77	79.6	79.6	79.6	79.6
	Precision	53	57.8	62.4	71.5	76.6	78.3	77	75.2	61.5
	F-measure	53.54	58.68	62.94	73	76.8	78.94	78.27	77.3	69.39
Support = 0.3 Confidence = 1	**Recall**	54	59	63.2	74.2	76.5	78.8	78.8	78.8	87.8
	Precision	53	57.8	62.8	73	78.4	80.2	78	77.3	65
	F-measure	53.5	58.39	62.6	73.6	77.43	79.5	78.4	78.04	71.23

Table 9. Recall, Precision, F-measure for the second evaluation.



	5	10	15	20
■ BLEU without enrichment (%)	75,2	72,6	68,1	65,7
■ BLEU with enrichment (%)	91,7	89,43	85,26	82,9
■ NIST without enrichment (%)	78,3	75,4	70,2	77,33
■ NIST with enrichment (%)	92,5	90,2	87,7	84,9
■ PER without enrichment (%)	18	17	17,5	21
■ PER with enrichment (%)	12,3	9	10	12
■ TER without enrichment (%)	14,5	16,4	17	19,9
■ TER with enrichment (%)	9	10,8	11,3	12,5

Figure 9. Results of automatic evaluation metrics applied on the translation system with and without contextual enrichment using manual alignment.

The used metrics in our experiments are metric of similarity with references such as BLEU measure (Bilingual Evaluation Understudy) [39] of 2-grammes (in our case two words because most queries include composed terms). The second one is the measure of Nist (National Institute of Standards and Technology).

For the metrics based on rate of erroneous words: we are interested in both metrics PER (Position-Independent Word Error Rate) [48] it does not take into account the order of words, and TER (Translation Error Rate) [47] count the minimum number of operations (insert, delete and substitution) performed on Q_s to transform it into Q_{ref}. Our results are given in the Figure 9.

Discussion: We can conclude from the figure (Figure 9) that contextual enrichment strengthens the ability to provide a correct translation in French of the queries w*i*th an important measure of precision (an improvement of about 16%, and rates error -8%).

We used the same experimental context, but using automatic alignment by MKAlign. The new evaluation was performed on the same collection of queries that served in the previous evaluation.

Discussion: The Figure 10 demonstrates the same impact of the contextual enrichment on the disambiguation in the improvement of the translation system. We have an improvement of about 17%, and rates error -9%.

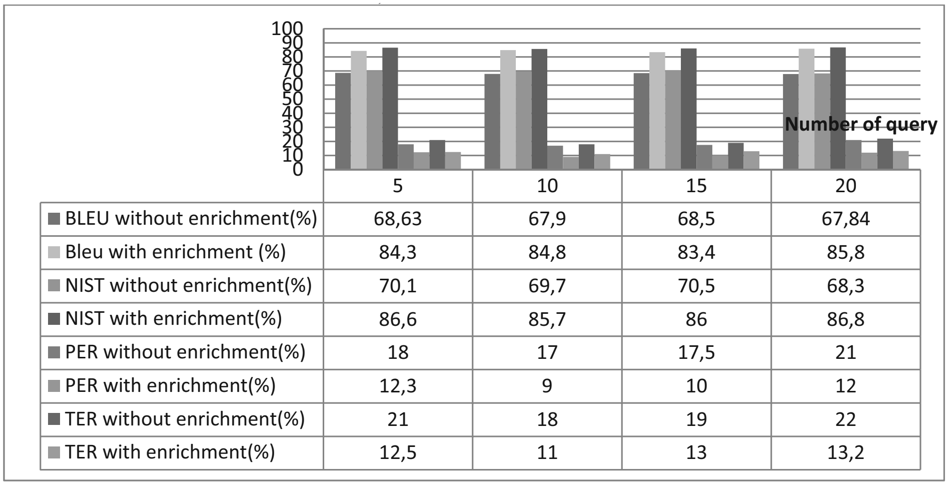
The comparative evaluation results presented in Figure 9 and Figure 10 show that the manual alignment makes the translation system more effective (as we attain an improvement of 7.5% and a decrease in error rate of 10%) compared to the translation system based on automatic alignment.

In general manner, improving the quality of translation shows that our method of lexical disambiguation overcomes the potential gaps of ambiguous words in Arabic queries. We were interested in the primary evaluation on the impact of the semantic network contextually enriched on the performance of the translation system. Second evaluation was conducted on the performance of the automatic alignment tool on the results. It can be concluded that a low lexical coverage and the use of a less efficient automatic alignment tool causes a loss in performance of the translation system. So we can deduce that precision of the translation system $T_{\text{Pécision}}$ is a linear function as follows:

$$T_{\text{Pecision}}(a) = \text{minconf} * \text{minsup} * a \qquad (13)$$

with minconf and minsup are two constants specified by user and (a) the precision of the alignment tool at the sentence level.

According to our experiments, we noted that increasing the values of the thresholds $(conf_{\text{min}} * \text{sup}_{\text{min}})$ makes the network size smaller, but we do loose in the exactitude of the network representation of the KB $(\text{list}_{\text{RSF}})$ compared to the reference network. That is why we must choose



	5	10	15	20
■ BLEU without enrichment(%)	68,63	67,9	68,5	67,84
■ Bleu with enrichment (%)	84,3	84,8	83,4	85,8
■ NIST without enrichment(%)	70,1	69,7	70,5	68,3
■ NIST with enrichment(%)	86,6	85,7	86	86,8
■ PER without enrichment(%)	18	17	17,5	21
■ PER with enrichment(%)	12,3	9	10	12
■ TER without enrichment(%)	21	18	19	22
■ TER with enrichment(%)	12,5	11	13	13,2

Figure 10. Results of automatic evaluation metrics applied on the translation system with and without contextual enrichment using MKalign.

the appropriate thresholds for each reconstruction. These thresholds ensure a large lexical coverage of KB (list_{RSF}), so not to lose too much information (to ensure an important recall). In our case $\text{sup}_{\min} = 0.5$, $conf_{\min} = 1$ achieves a maximum recall. See Table 9.

4.4. Impact of Automatic Alignment on Lexical Disambiguation in the Translation System

Conducted evaluation showed that the quality of translation system depends on the effectiveness of the alignment process. Indeed, in some cases, the KB (list_{RSF}) generated by the alignment technique, is not equivalent to KB(list_{RSA}). This results in several possible cases:

1. case: several Arabic sentences are aligned with the same sentence in French, which creates a redundancy, influencing the indexing method (assigns a high weight to less significant terms).
2. case: several Arabic sentences are aligned with French sentences whose content is completely different from their content.

5. Comparison with the Translation System of Google

The comparison with Google translate (which is a translation system based on statistical models) aims to demonstrate the impact of our lexical disambiguation method that we integrated in our automatic queries translation system. For a first comparative evaluation, we did a test with 20 queries containing lexical ambiguities. Those queries were translated by our system and by Google translate (August 2013). The translated queries were compared to the reference queries.

Table 10 presents the results of our system and a comparison with Google translate in terms of Blue Score and the rate of words out of vocabulary for the French language (OOV out of vocabulary). This rate of OOV increases the rate of error of word in the system. In fact, not

Number of queries	**Our system**		**Google translate**	
	Blue score (%)	OOV (%)	Blue score(%)	OOV(%)
5 queries	84.23	0.16	73.62	0.08
10 queries	82.9	0.20	64.36	0.15
15 queries	78.4	0.22	56.27	0.18
20 queries	75.8	0.27	44.75	0.22
Average	80.33	0.21	58.89	0.15

Table 10. Comparison between our system and Google translate (August 2013).

every OOV word is known but it influences the recognition of its neighbors' words.

An example of output of our system and Google translate is presented in the following Table 11.

The best performance of our translation system is reached. This clearly shows the interest of the lexical disambiguation method in choosing the exact meaning depending on the context (for example the word إقامة) (engl. foundation), and

Q1 Arabic: تأسيس صندوق مالي جديدا لدعم التنمية وتطويرالتكنولجية **Reference Translation Ar-Fr:** Fondation d'un nouveau fonds pour supporter et développer la technologie (*engl. Establishment of a new fund to support technological development*) **Our translation system (Ar-Fr):** Fondation d'un nouveau fonds monétaire appuyer évolution et le développement technologie (*engl. Foundation of a new Monetary Fund support evolution and development technology*) **Google translate:** La mise en place d'un nouveau fonds pour soutenir le développement et Ttoeraltknulgih (*engl. the establishment of a new fund to support the development and Ttoeraltknulgih*)
Q2 Arabic:: فيإقامةالسلامفيالعراقالدورالأوروبيوالأمريكي **Reference Translation Ar-Fr:** le rôle européen et américain dans l'instauration de la paix en Irak (*engl. European and American roles in Middle East peace instauration*) **Google Translate:** Dans l'établissement de la paix en Irak et le rôle de l'American européenne **Our system translation (Ar-Fr):** Rôle europe et America dans instauration paix en Irak
Q3 Arabic: الخبراء يشاهدون تناقضات في مظاهر العولمة **Reference Translation Ar-Fr:** les experts observent des contradictions dans la mondialisation (*engl. Experts observe contradictions in the globalization*) **Google translate:** Experts consultez la sexualité et la violence aveugle (*engl. experts at sexuality and mindless violence*) **Our system translation (Ar-Fr):** Experts voir contradiction dans des comportements mondialisation (*engl. Experts much contradiction in globalization behavior*)
Q4 Arabic: مفاوضات سريّة بين الولايات المتحدة وحركة طالبان **Reference Translation Ar-Fr:** Des négociations secrètes entre les Etats-Unis et talibanii (*engl. Secret negotiations between the United States and talibanii*) **Google translate:** Des négociations secrètes entre les Etats-Unis et le mouvement de talibanii (*engl. Secret negotiations between the United States and the movement of talibanii*) **Our system translation (Ar-Fr):** Négocier secrètes entre Et à tous unis et le mouvement étudiant (*engl. Negotiate secret between And all united and the student movement*)
Q5 Arabic: خطاب باراك أوباما وبعض ردود الأفعال االمختلفة **Reference Translation Ar-Fr:** Le discours de Barack Obama et certaines réactions différentes (*engl. The speech of Barack Obama and some different reactions*) **Google translate:** Le discours de Barack Obama et certaines des réactions Aalmokhtlfah (*engl. The speech of Barack Obama and some reactions Aalmokhtlfah*) **Our system translation (Ar-Fr):** Le discours de Arak Obama et certaines réactions actes différentes (*engl. Arak Obama's speech acts and some acts reactions different*)
Q6 Arabic: قتل أسامة بن لادن بعد سنوات من الملاحقة والعمليات الاستخباراتية **Reference Translation Ar-Fr:** L'assassinat d'Oussama Ben Laden après des années de poursuite et d'opérations de renseignement (*engl. killing Osama bin Laden after years of pursuit and inquiries operations*) **Google translate:** La mort d'Oussama ben Laden après des années d'opérations de poursuite et de renseignement (*engl. The death of Osama bin Laden after years of pursuit operations and inquiries*) **Our system translation (Ar-Fr):** L'assassinat Oussama cafier Laden après année de poursuite et opération renseignement (*engl. The assassination of Osama coffee Laden after year of tracking operation andinquiries*)
Q7 Arabic: حركة حماس يحظى بالدعم من قبل بعض الفلسطينيين **Reference Translation Ar-Fr:** Hames est soutenu par certains palestiniens (*engl. Hames is supported by some Palestinian*) **Google translate:** Hamas est soutenu par certains Palestiniens (*engl. Hames is supported by some Palestinian*) **Our system translation (Ar-Fr):** mouvement aideur ayant supporter de la part certains Palestinien (*engl. movement helping having support from some Palestinian*)

Table 11. Example of Arabic queries translation to French by our system and Google translate(August 2013).

the choice of the best synonym for example the word (قتل) (engl. Assassination). This is due to the use of all the lexical data of the KB (list_{RSF}) represented by lexical network, and lemmatization removes the ambiguity caused by the phenomenon of agglutination in the Arabic query words.

5.1. Discussion of the Results

A comparison of our machine translation system with "Google Translate" tool shows that our results are better in terms of BLEU score (80,33 compared to 58,89). This is because the MT system Google is very general and it has been trained on many heterogeneous data (it may contain a lot of data of the MD corpus). For the translation of unrecognized or incorrect words "Google Translate" performs their transliteration. However it is translated by our system. Google Translate produces less of out of vocabulary words (OOV) in French language with a gap of about 0,06 compared to our system. This decline in performance of our system is caused by the translation by word of named entities (which are composed of multiple words) after the lemmatization (see examples of queries Rq5, Rq6, Rq7 in Table 11). That's why a named entity composed of multiple words should not be separated in the middle, which imposes the following question: How can we improve the translation of named entities?

5.2. Proposed Solution

Several studies have been conducted to evaluate and improve the translation of named entities (NE) [43], [44], Ling et al. (2011) and Rauf (2012). The basic idea of our approach is quite similar to that of Hálek et al. (2011) who conducted a preliminary study to improve the automatic translation of named entities in Czech with Wikipedia. Our solution is to detect the Named Entities, then to translate them with Sensagent dictionary using Wikipedia, without going through the lemmatization process. For the detection of basic Named Entities we used NERAr (Named Entity Recognition for Arabic). The latter gives better performance for place names, names of people and organizations. The evaluation of NERAr showed an improvement for our translation system, it gave us a total F-measure equal to 81.96 [23]. Figure 11 illustrates the process.

6. Conclusion and Future Work

In this work, we compared two methods of disambiguation based on a semantic network with and without contextual enrichment for machine translation. For this, we selected Arabic queries containing several ambiguous words (polysemous, ambiguous words which belong to the context of the query). Then we evaluated and compared the performance of these two methods. The evaluation purposes are, first, to measure the impact of contextual enrichment on the translation. Second, to find a translation quality improvement approach based on the combination of two thresholds (minsupp, minconf). This threshold is associated with the Apriori algorithm to generate the association rules. These rules are used to extract useful hidden contextual links, and to fix the number of the relevant sentences forming KB (list_{RSF}).

Obtained results are satisfying, but could be improved in different ways

- Performance of the Arabic queries enrichment process: precisely the analysis phase and the pretreatment of queries in order to improve the results obtained by the method of lexical disambiguation and achieve a high coverage

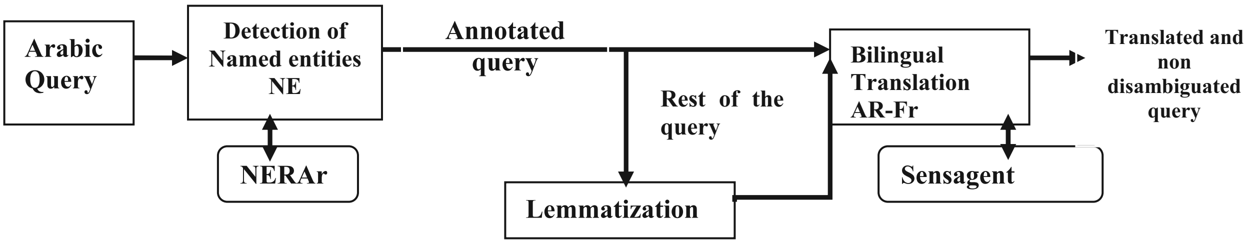

Arabic Query
Detection of Named entities NE
NERAr
Annotated query
Rest of the query
Lemmatization
Bilingual Translation AR-Fr
Sensagent
Translated and non disambiguated query

Figure 11. Identification and translation of named entities.

- The precision of alignment at the sentence level is an interesting idea. In order to improve the precision, we will try to use other alignment tool such as GIZA++ of Moses toolkit.
- The problem of translation by word of composed terms. A composed term that has two words (eg. ”الأفعال المختلفة” and translated only by one word (reactions)), is translated by word (different acts) and brings the noise. Also, there may be a composed word ex “العالم الثالث” (engl. Third World) which will be translated by word (“monde troisième”) (“world third”). The only way to solve this problem is to identify composed query terms before the translation.
- In future work, we will try to find a compromise between the size of the generic basis rules and the response time provided by our translation system.
- We intend to apply the process of our automatic translation system on other languages, such as English, German, and Spanish. We also intend to evaluate our translation system in the domain of cross-language information retrieval.

References

[1] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (1993) Washington D.C., pp. 207–216. `http://rakesh.agrawal-family.com/papers/sigmod93assoc.pdf`

[2] L. Ballesteros, W. Croft, Dictionary methods for cross-language information retrieval. In *Proceedings of DEXA'96*, (1996), pp. 791–801.

[3] C. Banea, A. Moschitti, S. Somasundaran, M. Zanzotto, TextGraphs-5 Workshop. Uppsala, Suède, 2010.

[4] M. Baziz, M. Boughanem, Nathalie Aussenac-Gilles. The Use of Ontology for Semantic Representation of Documents. *Dans, The 2nd Semantic Web and Information Retrieval Workshop(SWIR)* (Y. Ding, K. van Rijsbergen, I. Ounis, J. Jose, Eds.), (2004) pp. 38–45, Sheffield, UK.

[5] M. Baziz, M. Boughane, N. Aussenac-gilles, Conceptual Indexing Based on Document Content Representation. *Information Context, Nature, Impact, and Role, 5th International Conference on Conceptions of Library and Information Sciences, CoLIS 2005* (I. R. F. Crestani, Ed.), (2005), vol. 3507, Lecture Notes in Computer Science, pp. 171–186.

[6] O. Ben Khiroun, B. Elayeb, I. Bounhas, F. Evrard, N. Bellamine, A possibilistic approach for automatic word sense disambiguation. In *Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012)*, (2012) Chung-Li, Taiwan, China, pp. 261–275.

[7] W. Ben Romdhane, B. Elayeb, B. Ibrahim, F. Evrard, N. Bellamine-Bensaoud, A Possibilistic Query Translation Approach for Cross-Language Information Retrieval. In *Proceedings of International Conference on Intelligent Computing (ICIC 2013)*, (2013) Nanning, China, pp. 73–82.

[8] L. Ben Ghezaiel, C. Latiri, M. Ben Ahmed, N. Gouider-Khouja, Enrichissement d'ontologie par une base générique minimale de règles associatives – application aux maladies neurologies: les dystonies. In *Internationale Conference CORIA 2010*, (2010), pp. 289–300.

[9] L. Ben Ghezaiel, C. Latiri, M. Ben Ahmed, N. Gouider-Khouja, Un réseau proxémique pour la recherche d'information: Application à la maladie des dystonies. In *Proceedings of international Conférence coria*, (2011), pp. 26–33.

[10] S. Boucham, Une approche basée Ontologies pour l'indexing automatique et la recherche d'information Multilingue. Master thesis, University of M'hamed Bougara Boumerdes, 2009.

[11] M. Boughanem, C. Julien, J. Mothe, C. Soule-Dupuy, Mercure at TREC-8. In *Proceedings of TREC-8*, (2000). `http://trec.nist.gov/publications`

[12] H. Bunke, On a relation between graphs edits distance and maximum common subgraph. *Pattern Recogn. Letters*, **18**(9) (1997), 689–697.

[13] H. Bunke, K. Shearer, A graph distance metric based on the maximal common subgraph. *Pattern Recogn. Letters*, **19**(3-4) (1998), 255–259.

[14] C. Brouard, Construction et Exploitation de Réseaux Sémantiques Flous pour l'Extraction d'Information Pertinente le système RELIEFS, 2010. `http://www.risc.cnrs.fr/mem_theses_pdf/2000_Brouard.pdf`

[15] Y. Chiao, O. Kraif, D. Laurent, T. Nguyen, N. Semmar, F. Stuck, J. Véronis, W. Zaghouani, Evaluation of multilingual text alignment systems. *The ARCADE II project*, Actes de LREC-2006.

[16] K. W. Church, P. Hanks, Word Association Norms, Mutual Information, And Lexicography. *Computational Linguistics*, **16**(1) (1990).

[17] H. Dinh, Théorie de sens et la traduction des facteurs culturels: Synergies Pays riverains du Mékong *n*°1, (2010), pp. 142–143. `http://ressources-cla.univfcomte.fr/gerflint/Mekong1/dinh_hong_van.pdf`

[18] T. Delbecque, P. Jacquemart, P. Zweigenbaum, Repérage de phrases, et étiquetage par les relations sémantiques. *Utilisation du réseau sémantique de l'UMLS pour la définition de types d'entités nommées médicales*, (2005). `http://mrim.imag.fr/ARIA/2005/28.pdf`

[19] B. Espinasse, Représentation des Connaissances. *Introduction aux Réseaux Sémantiques*, (2008). `http://www.lsis.org/espinasseb/Supports/MR-TC/ReseauxSemantiques-oct08-4p.pdf`

[20] S. Fernández, P. Velázquez, Les systèmes de résumé automatique sont-ils vraiment des mauvais élèves. *JADT 2008, 9es Journées internationales d'Analyse statistique des Données Textuelles*, (2008). `http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/fernandez-velasquez-mandin-sanjuan-moreno.pdf`

[21] C. Fellbaum, *WordNet: an electronic lexical database, Language, Speech and Communication*. The MIT Press, Cambridge, Massachusetts, 1998.

[22] S. Fleury, M. Zimina, Exploring Translation Corpora with mkAlign. *Translation Journal*, **11**(1) (2002). `http://accurapid.com/journal/39mk.htm`

[23] S. Gahbiche, F. Yves, Amélioration des systèmes de traduction par analyse linguistique et thématique Application à la traduction depuis l'arabe. PhD thesis, University Paris Sud, LIMS, 2013.

[24] F. Harrathi, C. Roussey, S. Calabretto, Une approche d'indexation sémantique des documents multilingues guide par une ontology. In RISE (Recherche d'Information SEmantique). *In Proceedings of the Conference INFORISO*, (2009).

[25] X. Huang, S. E. Robertson Comparisons of Probabilistic Compound Unit Weighting Methods. *Proc. of the ICDM'01 Workshop on Text Mining*, (2001), San Jose, USA.

[26] Y. Kadri, Recherche d'information translinguistique sur les documents en arabe. Predoctoral report, DIRO, University of Montreal, 2003.

[27] F. Kboubi, A. Habacha Chabi, M. Ben Ahmed, L'exploitation des relations d'association de termes pour l'enrichissement de l'indexation de documents textuels. In *Proceedings of 10th International Conference of 'Journées d'Analyse statistique des Données Textuelles'*, (2010) Sapienza, University of Rome, pp. 9–11.

[28] J. Léon, La traduction automatique I. Les premières tentatives jusqu'au rapport ALPAC. In *History of the Language sciences* (W. de Gruyter, Ed.), (2002), Histoire des Sciences du Langage, pp. 2774–2780.

[29] A. Mercier, M. Beigbeder, Calcul de pertinence basée sur la proximité pour la recherche d'information. *Numeric document 2006/1*, **9**, (2006), 43–60.

[30] M. Manser, État de l'art sur l'acquisition de relations sémantiques entre termes. Contextualisation des relations de synonymie. In *Proceedings of International Conference JEP-TALN-RECITAL 2012*, (2012) Grenoble, pp. 163–175.

[31] S. Mallat, A. Zouaghi, M. Zrigui, E. Hkiri, Method of Enriching Queries in Process of Information Retrieval. In *Proceedings of Arabic International Conference on Agents and Artificial Intelligence (ICAART)*, (2013). `websiteamta2012.amtaweb.org/AMTA2012Files/papers/wor-far-01.pdf)`

[32] S. Mallat, A. Zouaghi, E. Hkiri, M. Zrigui, Method of lexical enrichment in information retrieval system in Arabic. In *Proceedings of International Journal of Information Retrieval Research (IJIRR)*, (2014) China.

[33] S. Mallat, Z. Zouaghi, M. Zrigui, Proposal of a method of enriching queries by statistical analysis of information retrieval in Arabic. Association for Machine Translation in the Americas (AMTA): *Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4)*, (2012) California, USA, pp. 80–86.

[34] R. Mihalcea, Using Wikipedia for Automatic Word Sense Disambiguation. *Actes de NAACL*, (2007), 196–203.

[35] A. Meillet, Linguistique historique et linguistique générale. *Honoré Champion*, (1926) Paris.

[36] N. Nassr, M. Boughanem, Croisement de langues en recherche d'information. Traduction et désambiguisation de requêtes par phrases alignées INFORSID 2002. In *Proceedings of the 20th International Congress INFORSID IRIN*, (2002) Polytech'Nantes, pp. 135–142.

[37] T. Oibeau, D. Dutoit, S. Bizouard, Évaluer l'acquisition semi-automatique de classes sémantiques. In *Proceedings of the (TALN 2002) Conference*, (2002) Nancy, France, pp. 37–39.

[38] M. T. Guerra, A. Graham, A. Way, Multi-word expression sensitive word alignment. In *Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010*, (2010), pp. 26–34.

[39] W. Oard, B. Dorr, A Survey of Multilingual Text Retrieval. Report UMIACS-TR-96-19 CS-TR-3615, 1996. `http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps`

[40] K. Papineni, S. Roukos, T. Ward, V. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation. *Proc. ACL-02*, (2002) Philadelphia, USA, pp. 311–318.

[41] A. Pirkola, The effect of query structure and Dictionary setups in Dictionary-Based cross-language information retrieval. In *Proceedings of 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, (1998), pp. 55–63.

[42] P. Resnik, Semantic Similarity in a Taxonomy, An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research (JAIR)*, **11** (1999), 95–130.

[43] E. HKIRI, S. MALLAT, M. ZRIGUI, Events automatic extraction from Arabic texts, In *Proceedings of International Journal of Information Retrieval Research(IJIRR)*, **3**(1) (2014).

[44] P. SANTANU, K. SUDIP, P. PAVEL, B. SIVAJI, W. ANDY, Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation. In *Proceedings of the COLING 2010 Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, (2010) Beijing, China, pp. 46–54.

[45] J. SAVOY, Recherche multilingue d'informations. Report of the Institute of Computer Science, University of Neuchâ-tel, 2001.

[46] M. SAYAH, R. NAGEM, La langue arabe, histoire et controversies, Synergies Espagne *n*°2, (2009), pp. 63–78.

[47] H. SCHMID, *New Methods in Language Processing*. Studies in Computational Linguistics, chapter Probabilistic part-of-speech tagging using decision trees, UCL Press, London, 1997. http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/

[48] M. SNOVER, B. DORR, R. SCHWARTZ, L. MICCIULLA, J. MAKHOUL, A Study of Translation Edit Rate with Targeted Human Annotation. In *Procedings of AMTA*, (2006) Cambridge, MA, USA, pp. 223.

[49] C. TILLMANN, S. VOGEL, H. NEY, A. ZUBIAGA, H. SAWAF, Accelerated DP-based search for statistical translation. In *Proceedings of Eurospeech*, (1997).

[50] P. VOSSEN, W. PETERS, ET AL., The Multilingual design of the EuroWordNet Database, of Lexical Semantic Resources for NLP Applications, (1997) http://citeseer.nj.nec.com/cache/papers/cs/343/http:zSzzSzwww.let.uva.nlz.SzewnzSzdocszSzP013.pdf/vossen97multilingual.pdf

[51] W. D. WALLIS, P. SHOUBRIDGE, M. KRAETZ, D. RAY, Graph distances using graph union. *Pattern Recogn. Letters*, **22**(6-7) (2001), 701–704.

[52] K. YAMABANA, S. MURAKI, S. DOI, S. KAMEI, A Language Conversion Front-end for Cross-Linguistic Information Retrieval. In *Cross-Language Information Retrieval, Chapter 8* (G. GREFENSTETTE, Ed.), (1998) pp. 202–210. Kluwer Academic Publisher, Boston.

[53] M. ZRIGUI, R. AYADI, M. MARS, M. MARAOUI, Arabic Text Classification Framework Based on Latent Dirichlet Allocation. *CIT. Journal of Computing and Information Technology*, **20**(2) (2012), 125–140.

[54] M. ZRIGUI, M. CHAHAD, A. ZOUAGHI, M. MARAOUI, A Framework of Indexation and Document Video Retrieval based of the Conceptual Graphs. *CIT. Journal of Computing and Information Technology*, **18**(3) (2010), 245–256.

Received: October, 2013
Revised: August, 2014
Accepted: September, 2014

Contact addresses:

Souheyl Mallat
Department of computer sciences
University of Monastir
Tunisia
LATICE Laboratory Research
e-mail: Mallatsou@hotmail.com

Mohamed Achraf Ben Mohamed
Department of computer sciences
University of Monastir
Tunisia
LATICE Laboratory Research
e-mail: mohamedAchraf@gmail.com

Emna Hkiri
Department of computer sciences
University of Monastir
Tunisia
LATICE Laboratory Research
e-mail: emna.hkiri@gmail.com

Anis Zouaghi
Department of computer sciences
Higher Institute of Applied Science and Technologies Sousse
Tunisia
LATICE Laboratory Research
e-mail: anis.zouaghi@gmail.com

Mounir Zrigui
Department of computer sciences
University of Monastir
Tunisia
LATICE Laboratory Research
e-mail: mounir.zrigui@fsm.rnu.tn

SOUHEYL MALLAT is a PhD student in Computer Sciences at the Faculty of Economics and Management of Sfax, under the guidance of Dr. Mounir Zrigui since 2011. His main research interests are in information retrieval in multilingual context (Arabic, French, English).

MOHAMED ACHRAF BEN MOHAMED is a Phd student at the Faculty of Economic Sciences and Management of Sfax, Tunisia. He is member of LaTICE Laboratory, Monastir unity (Tunisia). His areas of interest include natural language processing (NLP), computer-assisted language learning (CALL) and machine learning.

EMNA HKIRI is a PhD student in Computer Sciences at the Faculty of Economics and Management of Sfax, under the guidance of Dr. Mounir Zrigui. She is a member of LaTiCe Laboratory. Her main research interests are in natural language processing (Arabic language); text translation and ontologies.

ANIS ZOUAGHI received his PhD from the Manouba University, Tunisia in 2008 and his master from the INPG, Grenoble, France in 2000. Since 2009, he is a Computer Sciences Professor at Sousse University, Tunisia. He has started his research, focused on the processing of the Arabic Language, in RIADI Laboratory and later in LATICE Laboratory.

MOUNIR ZRIGUI received his PhD from the Paul Sabatier University, Toulouse, France in 1987and his HDR from the Stendhal University, Grenoble, France in 2008. Since 1986, he is a Computer Sciences Professor at Brest University, France, then at the Faculty of Science of Monastir, Tunisia. He has started his research, focused on all aspects of automatic processing of natural language (written and oral), in RIADI Laboratory and later in LaTiCe Laboratory.