Meaning representation for automatic indexing of Arabic texts

Bakhouche Abdelali, Yamina Tlili-Guiassa
Laboratory LRI, University Badji Mokhtar Annaba, Algeria

Abstract
The aim of indexing is to identify the words that represent the main idea of a paragraph or a specific text, in the framework of the representation of the meaning in an automatic treatment (NLP) of the Arabic; we propose a model based on conceptual vectors. These vectors try to represent the whole of ideas contained in textual segment (word, expression, texts ...). This model lean on modern linguistic conception the semantic field theory. By basing itself on the semantic relations (synonymy, homonymy...) between the words, we use these fields for the construction of semantic field data base and of a vectorial space then we calculate the meaning of textual segments in the semantic fields. Finally we use this model for indexing the text.

Keywords: Semantic modeling, NLP, Automatic indexing, Arabic language, Semantic field, Conceptual vector

1. Introduction
The semantic has an important position in the treatment of the natural language. It is inconceivable to realize deep treatments with sufficient information about the meaning of the semantic relations between the component words of their texts.
The vectors have been used long-past in the Information Research thus for the meaning representation in a LSA model from the latent semantic analysis studies (LSA) in psycholinguistics [5]. In NLP, proposed formalism for the projection of the linguistic notion of the semantic field in a vectorial space [09], on which our model is based.
The principle problem is axed on the primordial question: What is the vectorial space’s base that allows to represent the meaning of a textual segment in the Arabic language?
In this article, we present the conceptual vectors. We show their characteristics. We illustrate the steps of the construction of semantic field data bases. We use these fields for the construction of a vectorial space and then we construct a vectorial subspace of the semantic traits of each field. Finally we calculate the meaning of textual segments in the semantic fields.

2. Definition and advantages of the vectorial model for the meaning representation
A conceptual vector tries to represent a set of associated ideas to textual segment (word, sentence, syntagm, etc.) [3].
Briefly, we define a conceptual vector as linear combination of elements for meaning. The vectorial space must change as the reference corpus changes [3].
The conceptual vectors model lean on the projection in the mathematical model of information field’s linguistic notion [11]
- Every term (lexical item) and every concept is projected by a conceptual vector.
- We can also, calculate the theme of every text segment (documents, paragraphs, sentences, etc.) guise of conceptual vector: it is meaning of the segment in question.
- So, a vector corresponds to a linear combination of the terms meaning.
- This representation is homogenous as regards to the meaning, whatever the granularity.
- In this conceptual vectorial space, we can define a notion of semantic proximity by calculating angular distance between vectors. This means that we have a representation at the close meanings, without valorising correctly this proximity [4]. We do not know how to well decline this proximity in relation of hyperonymy or hyponymy that are characteristic of ontology [8]. In return, as the synonymy and this space according to the rule stated hereunder.

Be C a finite set of n concepts. A conceptual vector \( V \) is a linear combination of elements from \( c_{ij} \) to \( C \). For a meaning \( A \), the vector \( V_A \) is the description (in extension) of the activations of the C concepts. For example, the meanings “to order” and “to cut” can be projected on the following concept (the concept of c intensity being ordered by the descending intensity):

- **To order**: to sort out, to list, to select, to classify, to distribute, to group, to arrange, to clean, to disentangle, to adjust…

- **To cut**: to clip, to mince, to saw, to cut up, to rim, to intersperse, to crop, to shave, to slaughter, to pollard….
It is desirable to measure the proximity between the meanings represented by two vectors (so, the one of the associated word) \[6\]. Be \(\text{Sim}(X; Y)\) the measure of the similarity, usually used in information research, between two vectors defined by the expression (1) hereunder. We will not suppose here, that vectors components are always positive or null (although it is not necessary the case). Finally, we define an angular distance function \(\text{DA}\) between two vectors \(X\) and \(Y\) according to the expression.

\[
\text{Sim}(X, Y) = \cos(X, Y) / \|X\| \|Y\| \\
\text{DA}(X, Y) = \arccos(\text{Sim}(X, Y))
\]  

(1) (2)

Intuitively, this expression constitutes a thematic proximity evaluation and, it is in practice, the measure of the angle, formed by the two vectors. Generally, we will considerate between the synonymy and vectors a distance \(\pi / 4\). And \(\pi / 4\), the semantic proximity between \(A\) and \(B\) will be considered as weak.

Around \(\pi / 2\), the meanings are without relation. The synonymy (in its largest meaning) is included in the thematic proximity; however, it requires the concordance of the morphosyntactic category. The inverse is not evidently true. \[11\]

3. Outline of the study

We adopted the following steps:

- The first step is preprocessing of text corpus containing two operations (text segmentation, extraction of roots of words) \[2\].

- The second step focuses to the model proposed for the representation of the meaning of concepts.

- The final step is dedicated to indexing texts corpus using our model for representation the meaning of concepts.

3.1 The preliminary processing of texts corpus

We have two operations in this step

- Segmentation of text: is an important phase for the automatic processing of texts, it consists of dividing the text into lexical units.

- Extracting the roots of lexical units by using an automatic lexicon as \[1\].

Figure 1 shows these operations.

3.2 The stages of the model

To describe the model proposed in a more detailed way, we quoted its various stages. Three stages were defined in this process:

A. The first stage is the construction of a lexical data base made of a set of semantic fields and their contents in words.

B. The second stage is the creation of a vector space the base of which is made by the semantic fields composing the corpus in question.

C. The third stage is representation the sense of the concept in the vector space.
A. Construction automatic of the lexical database

These approaches are illustrated in the following:
- The collection of the common words in the same semantic field (to take into consideration the semantic relations: synonymy, homonymy...).
- The determination of the words possible meanings in a semantic field according to their meaning in the context.
- The determination of the semantic traits for every word in relation to the other words of the same semantic field.
- Put these semantic traits in a board.

A.1. Construction of a data base for a semantic relation (the synonymy)

The construction of a data base for a semantic relation "synonymy" consists of:
- To gather these synonyms from specialized dictionaries.
- To define the semantic field of synonyms.
- To define the semantics traits of these synonyms.

Example: we noted that the following words:
- قتل (to kill), (to slaughter), قتل (to strangle), قتل (to murder), مات (to die), (to execute), شنق (to hang), (to martyr) are synonyms that can be gathered in the same semantic field: الموت ( death ).

We can define the semantic traits of this semantic field as follows [1]:
- قتل (to kill) trait 01: ( death type ) not natural death, (actor)
- قتل (to slaughter) not natural death, intentional (actor), to slaughter the neck with a cutting edge (death type: natural / artificial) also (the method) and (the aspects of the used tool)
- قتل (to strangle ) not natural death, intentional (actor), to strangle a neck.
- قتل (to murder ) not natural death, intentional (actor), but for political motives (death type artificial), and also the motive of the murder.
- مات (to die ) natural death (Death type natural)
- اعدم (to execute) not natural death, intentional (actor), hanged on gallows and with a decision of a court (Death type) so (the method) and (the aspect of the used tool) and (lawfulness)

We can put the semantic field in board

<table>
<thead>
<tr>
<th>Field</th>
<th>Word</th>
<th>semantic traits</th>
<th>death type</th>
<th>Intervention by an actor</th>
</tr>
</thead>
<tbody>
<tr>
<td>الموت</td>
<td>مات</td>
<td>natural</td>
<td>intentional</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>اعدم</td>
<td>artificial</td>
<td>intentional</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>شيح</td>
<td>artificial</td>
<td>intentional</td>
<td>...</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The first column represents the semantic field and in that case we chose "الموت " this field groups together the synonyms of which we put some in the second column (.jpeg...). The rest of the table represents the distinctive features of every word constituting this field.

For example the word "اعدم" is characterized by these features: type of death, Intervention of an actor, a method, a reason and used Tool.

And in the same way we can treat the other semantic fields as: (الانتقال، الموت، الموت)

Table 2: The semantic traits of the semantic fields

<table>
<thead>
<tr>
<th>semantic Fields</th>
<th>Words</th>
<th>Actor</th>
<th>Speed</th>
<th>Direction</th>
<th>Means of Transport</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>سقط</td>
<td>-</td>
<td>normal</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>أخر</td>
<td>-</td>
<td>normal</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>شيخ</td>
<td>human</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>....</td>
<td>....</td>
<td>....</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ساري</td>
<td>normal</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>نصب</td>
<td>-</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>الشيش</td>
<td>normal</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>....</td>
<td>....</td>
<td>normal</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>....</td>
<td>....</td>
<td>....</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 from the book of the semantic fields "المميزات الداخلية العربية للمفاهيم الفعلية "

SLIMAN FIAD
A.2. Determination of Meanings to the semantic trait

We can determine the meaning of the words semantic traits as follows:

"مات": for all the living beings, natural death
"أعدم": for human being, artificial death, method: strangling, reason: a crime, used object: gallows, it is lawful
"استشهد": for the human being, artificial death reason: the war, used object: war weapons, unlawful.

<table>
<thead>
<tr>
<th>concept</th>
<th>Another</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ممات</td>
<td>human</td>
<td>yes</td>
</tr>
<tr>
<td>أعدم</td>
<td>human</td>
<td>yes</td>
</tr>
<tr>
<td>استشهد</td>
<td>human</td>
<td>yes</td>
</tr>
</tbody>
</table>

The first column represents the most used words in this meaning for this concept, for the second column there is precision of all the synonyms that represent the meaning, the third column represents the paragraph numbers of the themes presenting the use of synonyms, and the last column represents the precise meaning in the previous paragraph.

A.3. the Enrichment of the Data Base:

The relation between synonyms, homonyms and their meanings form a matrix:

The cell \( W_{ij} \) means that the form \( F_i \) is used to express the meaning \( M_j \). If there is several cells in the same line, then the both of expressions are synonyms, however if two cells are in the same column it means that they are homonyms example: \( W_{1,1} \) \( W_{1,2} \) \( W_{2,1} \) \( W_{2,2} \) represent two words that have the same form, but not the same meaning).

We can represent the elements of the first line as follows:

Table 3: The concepts semantic traits

<table>
<thead>
<tr>
<th>hyponym</th>
<th>synonyms</th>
<th>paragraph</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ممات</td>
<td>قتل</td>
<td>160</td>
<td>30</td>
</tr>
<tr>
<td>أعدم</td>
<td>161</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>استشهد</td>
<td>162</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>.....</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>

The concepts are not clearly independent

The construction of the semantic fields is based on the semantic relations (synonymy, homonymy etc.)

The concepts are not clearly independent

The too exhaustive number of the words and concepts in the Arabic language makes difficult their stake (placing) as a base for the vectorial space.

B. Creation of Vectorial Space

In this second stage we use the semantic fields constituting the database which we built in the first stage as axes of a vectorial space (the base of a space). The representation of the senses of the concepts (Arabic concepts in our study) in this space is made by vectors. The constituents of vectors are the distances between the concept and the semantic fields [7].

The graph (Fig. 2) illustrates an example of a vectorial space.

We choose the semantic fields as a base for the vectorial space for the following reasons:

- The construction of the semantic fields is based on the semantic relations (synonymy, homonymy etc.)
- The concepts are not clearly independent
- The too exhaustive number of the words and concepts in the Arabic language makes difficult their stake (placing) as a base for the vectorial space.

C. Representation of the sense of the concept in the vectorial space

The concept "سقط" is represented by the vector \( V_\text{sقاط} \) such as:

\[
V_{\text{sقاط}} = \alpha V_{\text{السقط}} + \beta V_{\text{السقط}} + \delta V_{\text{السقط}} \]

\( \alpha, \beta, \delta \) are the constituents of the vector \( V_{\text{sقاط}} \) in the vectorial space.
C.1 Calculation of the vector component

We use the method on base of measure of similarity. The constituents in this case are represented by the distances between the concept and the semantic fields; For this we use the method of the average link. The distance between the concept and the semantic field is calculated by the equation

\[
D(C, H_K) = \frac{1}{N_K} \sum_{C_j \in H_K} d(C, C_j) \quad (3)
\]

Where:
- \( N_K \): is the number of concepts in the semantic field \( H_K \).
- \( C \): is the concerned concept (example: السطح (السطح)).
- \( H_K \): is the semantic field which is an axis in the vectorial space (السطح (السطح), السطح (السطح), ..).
- \( d(C, C_j) \): is the distance between both concepts \( C \) and \( C_j \).

To calculate the distance between two concepts, we use as measure of dissimilarity the coefficient of Jaccard. If we have two concepts \( C_i \) and \( C_j \), the coefficient of Jaccard, \( d(C_i, C_j) \) is calculated thus:

\[
d(C_i, C_j) = \frac{b + e}{a + b + e} \quad (4)
\]

Where: \( a \) represents the features shared by both concepts, \( b \) the features of the concept \( i \) not appearing in the concept \( j \) and \( e \) the features of the concept \( j \) not appearing in the concept \( i \).

Table 5 shows some components of the vectors that represent concepts. So for the concept "سطح" (سطح) is represented by the vector \( V \) (سطح) (0.27, 1, 0.75, ..) where the value '0.27' is the constituent of the vector on the axis (السطح)، the value '1' is the constituent of the vector on the axis (السطح) and the value '0.75' is the constituent of the vector on the axis (السطح) in the vectorial space, same thing for the other concepts.

### 3.3 Text indexing

- Indexation of text is a process that allows the representation of text with a word or group of words (index) to be used in the research process [7][10].
- Calculating the cardinalities of vector for common traits and calculating the angular distance between this vector and other vectors.
- Put this word and the corresponding text in the table. Figure 4 shows these steps.

![Figure 3 Example of vector space](image)

![Figure 4: scheme illustrates indexing automatic for Arabic text](image)

### 4. Conclusion

This article establishes data bases intended to create a vectorial space to represent the meaning of a textual segment in Arabic by considering the semantics which is the biggest challenges, because of that the semantic and...
the context are strongly bound, which is hard to be planned by a computer.

The aim of our model is to improve the application of automatic processing of the Arabic language where semantic is important in automatic translation. It can concern to find the vector corresponding to the closest equivalent in another language, in automatic summary of texts. We may chose to privilege a part from the text that represent the principle idea of general speech better than another part, in categorization we may regroup the closest texts according to the method based on the angular distance…

RÉFÉRENCES