PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 95 APRIL 2011 19–32

Towards a New Approach for Disambiguation in NLP by Multiple Criterian Decision-Aid

Youssef Hoceini[a], Mohamed A. Cheragui[a], Moncef Abbas[b]

[a] High school of Computer Science, Algeria
[b] USTHB University, Algeria

Abstract

The aim of this paper is to present a combination of NLP and Multiple Criteria Decision-Aid (MCDA) in order to reach an effective analysis when dealing with linguistic data from various sources. The coexistence of these two concepts has allowed us, based on a set of actions and criteria, to develop a coherent system that integrates the entire process of textual data analysis (no-voweled Arabic texts) into decision making in case of ambiguity. Our solution is based on decision theory and an MCDA approach with a TOPSIS technique. This method allows the multi-scenario classification of morphosyntactical ambiguity cases in order to come out with the best performance and reduce the number of candidate scenarios.

1. Introduction

In the Arabic language, the duality between the word and vowels[1] implies a large increase in tidal volume of the tongue, knowing that a word can sometimes take more than twenty forms depending on the configuration that accompanies it. In fact, it leads to the most complex problems in understanding humans and machines Hoceini and Abbas (2009a). The phenomenon that arises from this multiplicity is called ambiguity. The determination of a unique morphosyntactic category for each word in the text of a treaty, for instance, is necessary for vowels in the text, and resolves most issues related to automatic processing of Arabic. The specific context of Arabic emphasizes

[1] Consider a set of codes that provide a number of functions have diacritical marks placed above or below the letters appear in some texts as: the Quraan, Hadith, poetry and textbooks in particular.

Corresponding author: y_hoceini@yahoo.fr
Cite as: Youssef Hoceini, Mohamed A. Cheragui, Moncef Abbas. Towards a New Approach for Disambiguation in NLP by Multiple Criterian Decision-Aid. The Prague Bulletin of Mathematical Linguistics No. 95, 2011, pp. 19–32. doi: 10.2478/v10108-011-0002-5.

the presence of a multitude of criteria that reflect the function of several constraints (e.g., grammar, semantics, logic and statistics). Therefore, a proper parsing system is required to be robust, fast, and most importantly less ambiguous.

This paper is organized as follows. First, an overall presentation of our morphological analyzer is given with a brief and comprehensive description of the phenomenon of ambiguity. The second part, we deals with the approaches for ambiguity removal or disambiguation. Next, the proposed model is presented along with the aggregation method known as "TOPSIS"[2] and the weighting method called "Entropy". Then, we show the implementation of our model. Finally, we summarize our findings in the conclusion.

Contrary to probabilistic and constraint based rules models, the proposed model of morphosyntactic disambiguation of Arabic implements an original method base on decision theory as an approach to categorize multi scenarios disambiguation in order to bring out the best. This approach has the advantage of reducing dominated scenarios and ranking the rest by different criteria evaluation.

2. Morphological Analysis

The morphological processing of the morpheme is based on two key concepts; The synthesis step that generates words or phrases based on a set of derivation rules, and inflectional adaptations, and the analysis step that associates a word graph to a set of information that describe the morphological and grammatical units of their composition (proclitic, prefix, basic, suffix, enclitic). This information allows the morphological analysis phase to determine the morphological properties of a word, such as: category (or part of speech: verb, noun or article), gender (male or female), number (singular or plural), voice (active or passive), time of action (accomplished or fulfilled), mode of the verb (indicative, subjunctive), and person (first, second or third person).

At this stage, the morphological ambiguity occurs when the analysis assigns a word more than one set of information (or the vice versa), which generates a combinatorial notion. Thus, prior to parsing, we must remove the ambiguity of many morphological labels that are associated to one word.

3. Disambiguation

Disambiguation is a crucial step in the process of morphological analysis. The morphological ambiguity in Arabic is mainly caused by the absence of vowels. According to Debili et al. (2002), 43.03 % of words are ambiguous in the Arabic voweled text. This proportion increases to 72.03 % when the text is not voweled. To sum up, the absence of these signs generates more cases of morphological ambiguity; for instance, the word with no vowels كتب (writing) may have 16 possible vowels, which

[2]TOPSIS: Technique for Order Preference by Similarity to Ideal Solutions

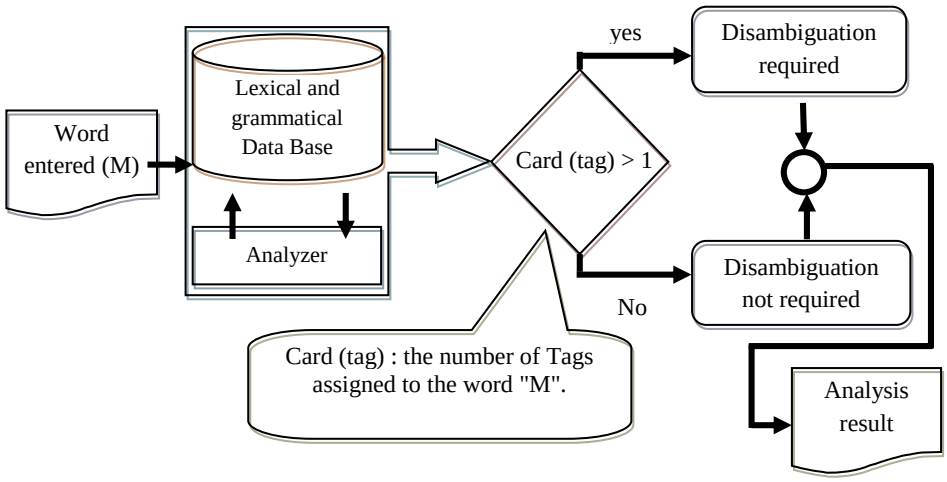

Figure 1. Different disambiguation techniques

leads to 9 different grammatical categories Aloulou et al. (2004). However the phase of disambiguation is not always required in the analysis process. The disambiguation module intervenes if the word receives more than one tag, which generates a situation of confusion or ambiguity (see Figure 1).

3.1. Existing Approaches to Disambiguation

Current analyzers are classified according to their mode of disambiguation. Yet, they all fall into two model classes; the probabilistic models that are meant grammatical labeling, and the constraint models Hoceini and Abbas (2009c). A summary of the different disambiguation techniques is given in Figure 2.

3.1.1. The Constraint Approach

This approach is based on a model that involves a linguist, which will allow the establishment of list of rules per class or category in order to be able to disambiguate. These categories can be: grammatical, structural, semantic, logical, etc… The grammatical constraints are mainly used for removing the ambiguity due to the simultane-

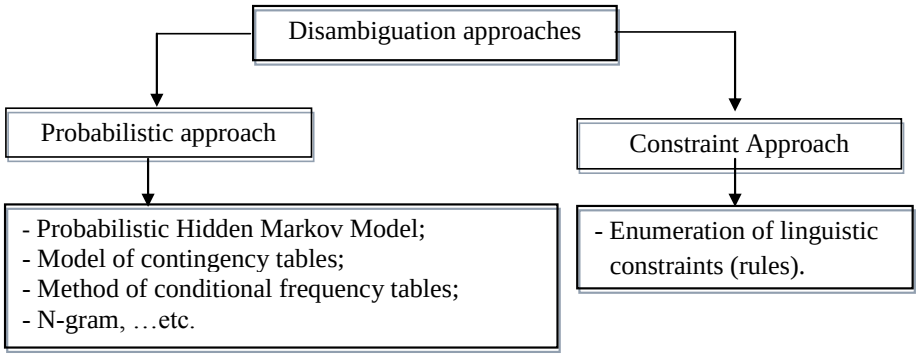

Figure 2. Different disambiguation techniques

ous membership of the semantic unit to more than one grammatical model. The use of grammatical constraints may be sufficient by itself, but sometimes other semantic constraints are imposed.

3.1.2. The Probabilistic Approach

In this approach, the probabilistic and statistical factor classifies constraints based on their redundancy. This is done on the basis of the highest rate of presence of a language constraint which can be lexical, morphological, syntactic, morphosyntactic or semantic. The statistical and probabilistic constraint are determined by searching in the language (corpus) to assess the rate of occurrence of each constraint in relation to other constraints. This rate is estimated using complex arithmetic. The removal of ambiguity is performed using two types of information: the words label and the contextual syntax. Then one proceeds to a combination of both information and learning[3] on their corpus annotated on hand. The Markov technique is a probabilistic model commonly used due to its efficiency Merialdo (1994).

[3]The technique of learning and classification: A set of examples is stored in memory, each set contains a word or its lexical representation, its context (anterior and posterior) and its grammatical category that is related to the context. The analysis is done as follows: for each word in the sentence, the Tager will look for a stored similar example (in memory) and deduce its grammatical category.

3.1.3. Comparison

Many researchers have found that constraint analyzers are faster and easier to implement than the stochastic parsers. In addition, they are more reliable and efficient in terms of analysis. Allotti and Ponsard (2005); Chanod and Tapanainen (1995). A third class of analyzers that combines the two previous approaches is added to increase performance and analysis suitability.

4. Proposed Approach : Multi-criteria Analysis Model

The NLP has frequent decision-making practices that meet a series of choices. Knowing the context of a specific language such as Arabic emphasizes the presence of criteria that reflect the function of several constraints (e.g., grammatical inflectional, structural, semantic, logical and statistics). So, the use of decision tools that support Multi-criteria is very effective Hoceini and Abbas (2009b).

Our goal is to propose a new model of disambiguation based on a mathematical approach called MCDA. The basis of this method is to involve the collection of many criteria from various sources to form a mega rule that guides a parsing process. The advantage of this approach is to reduce the number of disambiguation scenarios discarding the dominated scenarios (i.e., scenarios with no better assessment and dominated by all used criteria) and classifying the effective scenarios (i.e., the ones that are not dominated) by a calculated overall score. All this is based on a clear definition of assessment criteria.

4.1. Main phases of Proposed Model

The establishment of a morphosyntactic disambiguation process based on multiple criteria decision requires us to follow a number of steps shown in Figure 3.

4.2. Description of the Approach

Our approach is summarized in the following steps:

- **Step 1:** *Compilation of a list of potential actions.*
  The establishment of a set of all possible solutions or actions. In our case, these solutions are the ambiguous tags. So, let A is the set $(a_1, a_2, \ldots, a_n)$, where a_i is considered like a candidate label, then a set of morphosyntactic information is generated.
- **Step 2:** *Constructing of a coherent family of criteria* $F = \{f1, f2, \ldots, fp\}$.
  Proper application of a multi-criteria approach requires a good choice for the applied criteria. These criteria are defined on the based of different concepts such as consistency, indifference, strict preference and comparability. However, developing a test that influences the choice of scenario i compared to another scenario is not an easy task.

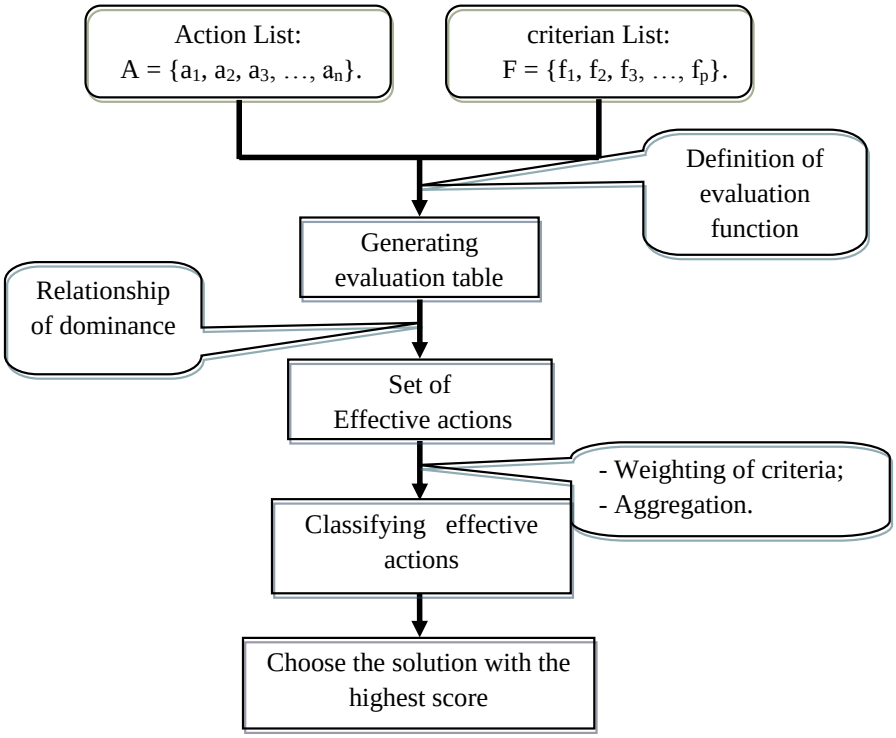

Figure 3. Different disambiguation techniques

But most importantly in defining a criterion is its power of discrimination between scenarios. In fact, discrimination becomes easier when the appropriate scenario is selected. However, a test that is discriminatory in some situations may not be so in other cases. Therefore, we need to construct a set of criteria that must meet three conditions namely: comprehensiveness, coherence and non-redundancy.

- **Step 3:** *Defining an evaluation function and an array of performance.*
  For each criterion we must generate an evaluation function that must be maximized or minimized depending on the type of the test used. The result of this function is a scorecard called the evaluation matrix. This later contains all the

evaluation results of each potential action when criteria are applied. Evaluation matrix rows correspond to the potential actions and the columns correspond to criteria. The matrix elements are the calculated estimates.

- **Step 4:** *Aggregation and criteria weighting*
  - **a) Aggregation:** it reduces the number of labels, and classifies them according to their overall scores. Choosing a method of aggregation will help standardize the evaluation table for better reading. To aggregate the different evaluations of a scenario calculated by the criteria, we propose to apply the TOPSIS aggregation method.
  - **b) Weighting:** it determines the weight of each criterion according to its importance[4]. So, weighting generates a vector of weights α, where each coordinate corresponds to a criterion. In our model, and to weigh the different criteria we adopt the Entropy weighting method.
- **Step 5:** *Selecting the label with the highest score*
  In order to obtain the scenario with the highest score, a classification of labels is performed decreasingly.

4.3. Aggregation Method : TOPSIS

4.3.1. Principle

The basis of the method is to choose a solution that is closest to the ideal solution, based on the relationship of dominance resulting from the distance to the ideal (the best on all criteria) and to leave the most of the worst possible solution (which degrades all criteria). TOPSIS is a multi-criteria method developed by Hwang and Yoon (1981). It reduces the number of disambiguation scenarios discarding the dominated ones, and ranking them according to their effective overall scores. In case of a tie, the closest scenario to the ideal, based on segregation measurements, is chosen.

4.3.2. Algorithm

- **Step 1:** Standardizing the performance (i.e., calculation of the normalized decision matrix); The normalized values e_{ij} are calculated as follows:

$$e'_{ij} = \frac{f_j(a_i)}{\sqrt{\sum_{i=1}^{m} [f_j(a_i)]}} \tag{1}$$

  With $i = 1, \ldots, m$, $j = 1, \ldots, n$, where $f_j\ (a_i)$ are the deterministic values of share(s) i for criterion j.

[4]The important criteria are able to discriminate between the solutions; and these criteria will have significant weights.

- **Step 2:** Calculation of the normalized and weighted decision matrix (i.e., calculating the product performance standard by the coefficients of relative importance of attributes). The matrix elements are calculated as follows:

$$e''_{ij} = \pi_j \cdot e'_{ij} \tag{2}$$

With $i = 1, \ldots, m$, $j = 1, \ldots, n$. π_j is the weight of j^{th} criterion.

- **Step 3:** Determination of ideal solutions (a^+) and anti-ideal solutions (a^-):

$$\begin{aligned} a^+ &= \{\max_i e"_{ij}, i = 1, \ldots, m; \ \& \ j = 1, \ldots, n\}; \quad e^*_j = \max_i\{e''_{ij}\} \\ a^+ &= \{e^*_j, j = 1, \ldots, n\} = \{e^*_1, e^*_2, \ldots, e^*_n\}; \\ a_- &= \{\min_i e"_{ij}, i = 1, \ldots, m; \ \& \ j = 1, \ldots, n\}; \quad e_{j*} = \min_i\{e''_{ij}\} \\ a_- &= \{e_j*, j = 1, \ldots, n\} = \{e_1*, e2*, \ldots, e_n*\}; \end{aligned} \tag{3}$$

- **Step 4:** Calculation of removal (i.e., calculate the Euclidean distance compared to the profiles a^+ and a^-). The distance between the alternatives is measured by Euclidean distance of dimension n. The remoteness of the alternative i with respect to the ideal (a^+) can be assimilated to the extent of exposure to risk and is given by:

$$D^*_i = \sqrt{\sum_{j=1}^{n} (e''_{ij} - e^*_j)^2} \tag{4}$$

$$D_{i*} = \sqrt{\sum_{j=1}^{n} (e''_{ij} - e_{j*})^2} \tag{5}$$

- **Step 5:** Calculating a coefficient that measures closeness to the ideal profile:

$$C^*_i = \frac{D_{i*}}{D^*_i + D_{i*}} \tag{6}$$

- **Step 6:** Storage of shares following their order of preferences (i.e., according to decreasing values of C^*_i; i is better than j if $C^*_i > C^*_j$).

4.4. Weighting Method : Entropy

4.4.1. Principle

The Entropy method is an objective technique for the weighting of criteria. The idea is that a criterion j is more important than the dispersion of stock valuations. Thus the most important criteria are those that discriminate most between actions (in our case actions are labels).

4.4.2. Algorithm

The entropy of a criterion j is calculated by the next formula Pomerol and Barba-Romero (1993):

$$E_j = -K \cdot \sum_{i=1}^{n} X_{ij} \cdot Log(X_{ij}).$$

where K is a constant chosen so that for all j, such as $0 \leq E_j \leq 1$, and $K = 1/\log(n)$ (n is the number of scenarios disambiguation). The entropy E_j is much larger than the values of e_j which are close. Thus, the weights are calculated according to the D_j (opposite of entropy):

$$D_j = 1 - E_j.$$

The weights are then normalized:

$$W_j = \frac{D_j}{\sum_j D_j}.$$

5. Implementation of the Proposed Solution

To better understand the proposed solution, we will keep the same approach mentioned above.

Let P ="رجع الوطنِ إلَى المغترِب", presented to our analyzer.

After segmenting the sentence into words, the analysis is done without any problem for units 2, 3 and 4. However, unit 1 "رجع" presents a typical morphological ambiguity. To remove this ambiguity we will apply our approach called multicriteria disambiguation as follows:

- **Step 1: Building a List of Analysis Scenarios:**
  The list (the set A) is obtained directly after the process of morphological analysis.

Verb	Scenario	Root
رجـع	فَـعَـلَ	رجـع
	فَـعِـلَ	رجـع
	فَـعُـلَ	رجـع
	فُـعِـلَ	رجـع

Table 1. Example of ambiguity generated when analyzing the verb "رجـع".

- **Step 2: Application of Criteria** To build a coherent family of criteria F, we propose two basic criteria to discriminate between the scenarios of the analysis: the test of vowel consistency, and the occurrence frequency test.

a) Criterion 1: Concordance of Vowels
This test will verify the correlation between the vowels of the lexical unit and the vowels of each candidate scenario. This test maximizes the function of assessment that goes with it is the addition (+).

a) Criterion 2: The Frequency of Occurrence.
This criterion is based on a statistical calculation on the basis of an annotated corpus so that the scenario that occurs most frequently will always score the highest. (Each appearance is one (1), so this is a test and to maximize the evaluation function that goes with it is the addition (+)). The results of applying this criterion are made on the basis of an annotated corpus is composed of 300 units spread over 10 arbitrarily selected paragraphs that are selected from (the books school school) an Algerian school textbook.

- **Step 3: Application of the Evaluation Function**
For both criteria (Concordance of vowels and frequency of appearance) the evaluation function is addition (+).
- **Step 4: Generating a Score Table (or score matrix)**

Scenario→ Criteria↓	S1"فَـعَـلَ"	S2"فَـعِـلَ "	S3"فَـعُـلَ"	S4"فُـعِـلَ"
Vowel Concordance	3	2	2	1
Appearance Frequency	16	5	2	1

Table 2. Evaluation Table (matrix).

- **Step 5: Aggregation and Weighting of Performance Criteria.**
Normalization of the scorecard is made by applying the formula (1) of the TOPSIS method.

Scenario→ Criteria↓	S1"فَـعَـلَ"	S2"فَـعِـلَ"	S3"فَـعُـلَ"	S4"فُـعِـلَ"
Vowel Concordance	0.71	0.47	0.47	0.24
Appearance Frequency	0.95	0.30	0.12	0.06

Table 3. Normalization of the Score Table.

a) Weighting of Criteria
In order to weight the criteria we use the entropy method, with respect of the initial condition mentioned in TOPSIS, i.e., the sum of the weights must be equal to 1. The following table shows the calculation Entropy values (E_j), the opposite of entropy (D_j) and normalization of weight (W_j) of the two criteria.

E_j	D_j	W_j
0.24	0.76	0.47
0.15	0.85	0.53

Table 4. Weighting the criteria

Note:
Checking the Status of weighting:

$$\sum_{j=1}^{p} W_j = W_1 + W_2 = 0.47 + 0.53 = 1.$$

(Condition tested).

b) Weighting of Evaluation Table (standard): This weighting is done using the formula (2) of the TOPSIS method.

Scenario→ Criteria↓	S1"فَعَلَ"	S2"فَعِلَ"	S3"فَعُلَ"	S4"فُعِلَ"
Vowel concordance	0.33	0.22	0.22	0.11
Frequency of appearance	0.50	0.16	0.06	0.03

Table 5. Weighting of Score Table

c) Calculation of Removal Measures
After applying formulas (3), (4) and (5), TOPSIS method reacts with different measures of distance for each scenario as illustrated in Table 6:

	S1"فَـعَـلَ"	S2"فَـعِـلَ"	S3"فَـعُـلَ"	S4"فُـعِـلَ"
D^*	0.33	0.22	0.22	0.11
D_*	0.50	0.16	0.06	0.03

Table 6. Weighting of Score Table

d) Calculation of the Measure of Closeness to Ideal Profile
To calculate coefficients C_i^*, we use the formula (6) of the TOPSIS method, and then establish a decreasing ranking of the factors. The scenario with the highest score is elected. So, these are the values obtained:

$$C_1^* = 1 > C_2^* = 0.32 > C_3^* = 0.24 > C_4^* = 0.$$

In our method the solution 1 "فَـعَـلَ" will be selected by the system, so the following morphological information will be generated.

	Information
Root	رجـع
Pattern	فَـعَـلَ
Tag	VAA3PMSIA
Designation in French	Verbe Accompli Actif 3e Personne Masculin Singulier Invariable Accusatif
Designation in English	Accomplished Verb Active 3rd Person Masculine Singular Invariable Accusative
Designation in Arabic	الفتح. على مبني الغائب، المذكر للمفرد للمعلوم مبني ماضي عل
Verb vowelized	رَجَـعَ

Table 7. Information generated by tagging the verb "رجع".

6. Conclusion

Using multiple criteria decision is a methodology that provides decision makers with tools to solve a decision making problem, taking into account several points of view. This paper attempts to present a new mathematical approach based on MCDA in order to categorize Multi scenarios of disambiguation and extract the best. This

method has the advantage of reducing dominated scenarios and ranking the rest by different evaluation criteria. Even though this technique is not widely used, it shows that the path of a multi-criteria analysis in NLP (based on recurrent common phenomena and to texts in all languages combined,) is very interesting. This technique offers an alternative and crucial complement method compared to systems that are based on a probabilistic approach and can be an indispensable complement to the model by contextual constraint.

Bibliography

Allotti, D. and C. Ponsard. Exposé sur l'étiqueteurs Statistiques et étiqueteurs par contraintes, 2005.

Aloulou, C., L. H. Belguith, A. H. Kacem, and A. Ben Hamadou. Conception et développement du système MASPAR d'analyse de l'arabe selon une approche agent. *RFIA*, 2004.

Chanod, J. P. and P. Tapanainen. Les étiqueteur statistiques et les étiqueteurs par contraintes, 1995.

Debili, F., H. Achour, and E. Souissi. La lague arabe et l'ordinateur: de l'étiquetage grammatical à la voyellation automatique. *IRMC*, 2002.

Hoceini, Y. and M. Abbas. Morphosyntactical Disambiguation Model of Arabic Based on a Multi-criteria Approach. In Arabnia, Hamid R. and David de la Fuenteand Jose A. Olivas, editors, *International Conference on Artificial Intelligence, ICAI 2009*, volume 2, pages 756–762, Las Vegas Nevada, USA, 2009a. CSREA Press.

Hoceini, Y. and M. Abbas. Une analyse multicritère de l'arabe. In *Journées d'étude du FSP France-Maghreb 'Pratiques langagière au Maghreb : corpus et applications*, Paris, France, Septembre 2009b. CERTAL - INALCO.

Hoceini, Y. and M. Abbas. Méthodologie Multicritère de Désambiguïsation Morphosyntaxique de la langue Arabe. In *3rd International Conference on Arabic Language Processing, CITALA'09*, pages 89–95, Rabat Morocco, May 2009c.

Hwang, C. R. and K. Yoon. *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag Berlin Heildelberg, New York, 1981.

Merialdo, B. Tagging english text with a probabilistic model. *Computational linguistics*, 1994.

Pomerol, J.C. and S. Barba-Romero. *Choix multicritère dans l'entreprise: principes et pratique*. Hermes, 1993.

Address for correspondence:
Youssef Hoceini
y_hoceini@yahoo.fr
Computer Science Institute,
Bechar University, P.B. 417,
Bechar 08000, Algeria