An Agreement Error Correction Method Based on a Multicriteria Approach : An Application to Arabic Language

BELGUITH HADRICH Lamia, BEN HAMADOU Abdelmajid, ALOULOU Chafik
Faculté des Sciences Economiques et de Gestion de Sfax
Laboratoire de recherche LARIS, B.P. 1088, 3018 - Sfax - TUNISIE
Tél. (2164) 278 777, Fax (2164) 279 139

Abstract

Most parsers handling syntactic agreement detect the errors but rarely give enough information on how to correct them. Our interest here is the agreement error correction. Thus, we suggest a multicriteria approach to guide the choice of the best alternative. We propose three main criteria (frequency criterion, morphological criterion and typographic criterion) which we apply to Arabic sentences in order to evaluate the alternatives, and we show the interest of TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) as an aggregation method for the proposed criteria.

Key Words: Agreement error detection, agreement error correction, multicriteria approach, TOPSIS, correction alternatives ranking.

1 Introduction

Many studies have dealt with the problem of agreement errors in written texts. Most of them addressed the detection process rather than the correction one.

The correction process involves the problem of choosing the proper correction among several alternatives. The first parsers left this choice to the user ((Ravin 88), (Coch and Morise 90)) although it can be done automatically and without hesitation in some cases. Therefore, the next parsers opted for the automation of the process of choosing the appropriate correction by using criteria to rank the correction alternatives. Thus the user is guided by the parser in order

to choose the appropriate correction ((Lapalme and Richard 86), (Veronis 91), (Bolioli and al 92)).

The first criterion proposed to classify the possible corrections gives priority to the head of the phrase ((Strube 90), (Genthial and al 90), (Genthial and al 94)) : the idea here is that the writer takes more care to the main words (the governors) than the others (the dependants).

It is clear that this criterion is irrelevant if we deal with competence errors. In French these errors are mainly omission or addition of silent morphological marks (e.g., the mark "s" of the plural) (Veronis 88). In such cases, the governor can't be used for the correction even if the user gives a particular care to the main words.

Moreover, if we consider the case where the agreement marks of the dependants are more frequent than that of the governor, it is unfair to impose the correction according to the governor features only.

These works prove that taking into account one criterion is not appropriate to differing phrases. We think that more one criterion must be considered. The multiplicity of criteria can handle the different causes of errors.

Our study of the Arabic language proves that we can choose three main criteria (the frequency criterion, the typographic criterion and the morphological criterion).

This paper focuses the use of a multicriteria approach to classify the possible corrections in order to choose the best one. This approach can be applied to any language even if in this paper we choose the Arabic language.

We first present a brief overview of the method used to detect the agreement errors in Arabic sentences. Then we propose three main criteria to evaluate the correction alternatives and we present the techniques of scoring them. Finally, we show the interest of using TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) as an aggregation method of these criteria.

2 Agreement error detection method

Most of studies agree with the necessity of parsing the apprehended text in order to detect the agreement errors ((Blache 90), (Blache 91), (Genthial and Courtin 92)). But we think that the accuracy and the robustness of the parser strongly depend on the typology of errors to be

handled. Thus, to detect the past participle agreement errors, a robust parser is required in order to identify the correct syntactic dependencies (Lapalme and Richard 86). Whereas, to handle agreement errors in gender and number, we think that a partial analysis can be sufficient. We proposed in ((Ben Hamadou and Belguith 96a), (Belguith and Ben Hamadou 96b)) a global analysis approach applied to Arabic and termed "Extended Syntagmatic Analysis". This approach aims to group, in the same sets, all the units of the phrase concerned by at least one agreement rule. The resulting sets are termed the "Extended Syntagms".

The proposed approach is based on two main steps which may be summarised as follows :

Step 1: Identification of the initial syntagms

The initial syntagms are mainly identified by the location of the "Function words" (i.e., the particles, the prepositions, etc.). These words are used to identify the syntagm boundaries. We can distinguish three categories of "Function words":

- words which separate two consecutive syntagms and do not belong to any one of them (e.g., prepositions, coordinating conjunctions, etc.)
- words which start a syntagm and belong to it: this is the case of demonstrative pronouns, relative pronouns, etc.
- words which end a syntagm and are referred to previous words which do not belong to this syntagm (i.e., possessive pronouns).

Step 2: Constitution of the "Extended Syntagms"

The constitution of the "Extended Syntagms" is guided by a rule set which aims to extend the initial syntagms by all the units of the phrase (function words or initial syntagms) that have a dependency relationship.

The result is a list of independent syntagms in which we can, separately, apply the process of agreement error detection. The detection process can be reduced to a simple unification process of the morphological features of all the constituents of the extended syntagm.

Example :

Let us consider the sentence : 'إعتنى الممرّضة المتربّصات بالمريض وأعطته الدّواء'

↑3 ↑2 ↑1

(The $\text{trainee}_{\text{fem.plu.}}$ $\text{nurse}_{\text{fem.sing.}}$ took care $_{\text{masc.sing.}}$ of the $\text{patient}_{\text{masc.sing.}}$ and gave $_{\text{fem.sing.}}$ $\text{him}_{\text{masc.sing}}$ some medicines)

The location of the function words: "ب" (1) , "و" (2) ," ه" (3) entails the decomposition of the sentence into the following initial syntagms :

SI_1={ المتربّصات (trainee) , الممرّضة (nurse), إعتنى (took care) }

SI_2= { المريض (patient) }

SI_3= { أعطت (gave) }

SI_4= { ه (him) }

SI_5= { الدّواء (medicines) }

The result of the "Extended Syntagmatic Analysis" is given by :

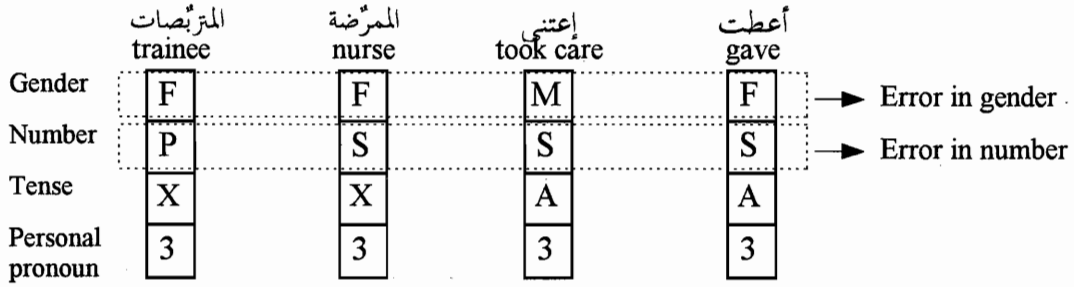
SE_1={ أعطت (gave), المتربّصات (trainee) , الممرّضة (nurse), إعتنى (took care) }

SE_2={ ه (him), المريض (patient) }

SE_3= { الدّواء (medicines) }

The detection process can be done separately in each extended syntagm[1].

Let us consider SE_1. The units of SE_1 and their features can be represented by the following figure[2] :



The unification process of the unit features of SE_1 fails in terms of gender and number.

3 Agreement error correction method

Upon many correction alternatives, the choice of the best alternative may be obvious : this is generally the case of phrases involving few errors since the best solution is the same in terms of all points of view. For instance, the sentence 'الأولاد الذين يلعب في الحديقة تلامذتي' (the

1 SE_3 is a singleton, so it is not concerned by the detection process.

2 Gender : F(Feminine), M(Masculine).
Number : S(Singular), P(Plural), D(Duel)
Where Duel refers to two persons and plural refers to more than two persons.
Tense : P(Present). F(Future), A(Past).
Personal pronoun : 1.2,3.

children$_{masc.pl}$ who plays$_{masc.sing}$ in the garden are$_{masc.pl}$ my students $_{masc.pl}$) has two possible corrections : the first one aims to line up the sentence with the singular, however the second one favors the plural.

To classify these corrections, a first point of view consists of minimizing the number of errors and therefore aims to favor the correction which features are the most frequent. An other point of view may favor the correction that minimizes the number of typographic transformations.

We can remark that the second alternative (plural) is the best one according to the two points of view (three word in the plural and only one in the singular; addition of two letters versus an omission of five letters and a substitution of one letter).

Nevertheless, the best correction is not usually the same according to all points of view. For example, if we consider the sentence 'البنت الصغير النشيط نجحوا في الإمتحان' (The little $_{masc..sing.}$ and dynamic $_{masc..sing.}$ girl $_{fem.sing.}$ succeeded $_{masc.pl.}$ in the exam), the best correction given by the frequency criterion (first point of view) is the masculine-singular. Whereas the typographic criterion (second point of view) favors the correction with the feminine-singular. Consequently, the choice of the best alternative requires a careful analysis and, inevitably, needs a negotiation between the considered criteria. In the following section, we present the basic concepts of a multicriteria approach and we show how it is appropriate to this kind of problems.

3.1 Basic concepts of the multicriteria approach

Our multicriteria decision problem can be defined as follows :

Let $X=\{x_1, \ldots, x_n\}$ a set of correction alternatives and let $F = \{f_1, \ldots, f_q\}$ a set of criteria. The evaluation function of an alternative x_i according to criterion f_j is denoted by :

$$f_j : X \rightarrow IR$$
$$x \rightarrow f_j(x)$$

Each criterion has to be maximized. The problem can be written as follows :

$$\left| \begin{array}{l} \text{"Max" } f(x)= (f_1(x), f_2(x), \ldots, f_q(x)) \\ \\ \text{subject to } x \in X \end{array} \right.$$

We say that x_1 dominates x_2 if $\forall$ i, $f_i(x_1) \geq f_i(x_2)$ $(f(x_1) \neq f(x_2))$.

A correction alternative in X is said efficient if it is not dominated.

In order to determine all efficient correction alternatives, we can use one of the following methods :

1. $P(\alpha) = \text{Max} \sum_j \alpha_j f_j(x) \quad \alpha_j > 0 \ , \quad \sum_{j=1}^{q} \alpha_j = 1$, where α_j is the associated weight of the criterion f_j. The solution of $P(\alpha)$ is an efficient correction alternative.

2. The ideal correction alternative (x^+) is the point in IRq whose cordinates are :

$$(y_1^+, \ldots, y_q^+) \text{ where } y_j^+ = \underset{x}{\text{Max}}\ f_j(x) \quad j = 1,\ldots,q$$

$P(d) = \underset{x}{\text{Min}}\ (d(f(x^+), f(x)))$, where d is a distance (e.g., Euclidean distance).

The solution of P(d) is generally an efficient one.

In our work we will use TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) which is advocated to cardinal criteria and uses a combination of $P(\alpha)$ and P(d) to rank the efficient correction alternatives.

3.2 Main criteria to evaluate the correction alternatives

When the detection of agreement errors involves many correction alternatives, choosing the best one is not usually a simple task since it requires the ranking of the alternatives according to many criteria.

We propose three main criteria to evaluate the correction alternatives in Arabic sentences : the frequency criterion, the morphological criterion and the typographic criterion.

3.2.1 Frequency criterion

The frequency criterion is measured by the occurrence of the alternative features in the sentence. This criterion favors the alternative whose features are more frequent in order to minimize the agreement error number. For example, the sentence :

'هـذه البنـت الاتـي تلعبن في الحديقـة جميلـة جـدا' (This$_{\text{fem. sing.}}$ girl$_{\text{fem.sing.}}$ who$_{\text{fem.pl.}}$ play$_{\text{fem.pl.}}$ in the garden is very beautiful$_{\text{fem.sing.}}$) is corrected in the singular (only two words corrected) rather than the plural (three words corrected).

To determine the score of an alternative x according to this criterion, we sum the occurrence of the pair (gender, number) with the occurrence of the tense and that of the personal pronoun. We obtain the following :

$f_1(x)$ = Occ(gender, number) + Occ(tense) + Occ(personal pronoun)

where Occ is the occurrence of the specified feature[3].

Let us consider the sentence presented above, the units concerned with the agreement and their features are represented in the following figure :

	هذه (This)	البنت (girl)	اللاتي (who)	تلعبن (play)	جميلة is beautiful
Gender	F	F	F	F	F
Number	S	S	P	P	S
Tense	X	X	X	P	X
Personal pronoun	3	3	3	3	3

There are two possible corrections :

x_1= feminine, singular, present, 3

x_2= feminine, plural, present, 3

The outcomes of these alternatives are respectively :

$f_1(x_1) = Occ(F,S) = 3$

$f_1(x_2) = Occ(F,P) = 2$

Thus in terms of this criterion, we choose x_1 as the best correction since $f_1(x_1)$ is greater than $f_1(x_2)$.

3.2.2 Typographic criterion

This criterion is devoted to agreement errors which have a lexical origin. These errors can not be detected by the lexical analysis given they belong to the lexicon; however they are detected by the parser as agreement errors since they didn't fire the agreement rules. For instance, in the sentence : ' البنت الصغيرات تلعبن بدراجتهن في الحديقة ' (The $little_{fem.pl.}$ $girl_{fem.sing.}$ $play_{fem.pl.}$ with $their_{fem.pl.}$ bicycle in the garden), it is clear that the agreement error in number between the noun 'البنـت' (girl) and the other units of the phrase has a lexical origin. So, in order to write 'البنـات' (girls), the user can omit the letter 'ا' and then writes the word 'البنت' (girl) which belongs to the lexicon.

To carry on this kind of errors, the typographic criterion favors the correction which minimizes the typographic transformations of the erroneous words. In most cases these transformations can be either an omission of a letter (e.g., ' بنات' (girls) → ' بنت' (girl)), an

[3] If the specified feature doesn't fail in terms of the unification process, we attribute zero to its associated occurrence.

addition of a letter (e.g., 'طالب' (student)$_{masc.sing.}$ → 'طالبة' (student)$_{fem.sing.}$) or a substitution of a letter by another (e.g., 'مهندسون' (engineers)$_{masc.pl.}$ → 'مهندسان' (engineers)$_{masc.duel}$). The permutation errors occur in some cases and they are generally followed by one of the errors presented above (e.g., 'مهندسات' (engineers)$_{fem.pl.}$ → 'مهندستان' (engineers)$_{fem.duel}$: permutation of two letters and omission of a letter).

To take into account the transformations necessary to correct an erroneous word, we assign the following ordinal scores to the different kind of errors (Ben Hamadou 93) :

- score of a letter omission ($W_o = 0.25$)
- score of substitution of a letter by another ($W_s = 0.5$)
- score of permutation of two letters ($W_p = 0.75$)
- score of a letter addition ($W_a = 1$)

These scores are chosen according to the frequency of each kind of error. For instance, the writer can omit a letter where it is necessary, but rarely adds one where it is not. Consequently, the score of omission of a letter is strictly lower than that of the addition.

The outcome $f_2(x)$ of an alternative x in terms of this criterion is the sum of the scores of the different typographic transformations which affect the erroneous words. For instance, to change the word 'طالب' (student$_{masc.sing.}$) by the word 'طلبة' (students $_{masc.pl.}$) there is a total score of 1.25 (addition of a letter and omission of a letter).

3.2.3 Morphological criterion

Generally, the correction of an erroneous word requires the change of some of its letters. Thus, in English, to conjugate a verb in the present with a plural personal pronoun, we omit the letter 's' from the singular form (e.g., he eats → they eat). However, in French, we must add the letters 'ent' (the plural mark) (e.g., il mange $_{masc. sing.}$ → ils mangent $_{masc. pl.}$). The same thing is used for Arabic since we add the plural mark which is generally represented by one or many letters (e.g., هو يأكل → هم يأكلون).

We can say that the change of the tense or the personal pronoun of a verb requires the addition/omission of some of its letters. This is not usually the case for the nouns and adjectives since many words have restricted morphological features. For example, the noun 'رجل' (man) can be only masculine and the adjective 'حامل' (pregnant) can be only feminine. In these cases, to change the morphological features of a word, we must change the root of the

word. For instance, the masculine form of the word 'بنت' (girl) is 'ولد' (boy) (the second word is not derived from the first one since they didn't share the same root).

To handle this kind of errors, the morphological criterion favors the corrections that do not change the root of a word. Consider the sentence : 'البنت ياكلون تفاحة' (The $girl_{fem.sing.}$ $eat_{masc.pl.}$ an apple) which includes an agreement error in gender between the subject ' بنت' (girl) and the verb 'ياكلون' (eat). The best alternative in terms of this criterion is the feminine-singular: it is more simple to correct the verb by changing some of its letters (a substitution of a letter by another and a deletion of two words) than to change the noun by another which is not derived from it. The idea here is that the user may omit or add some letters by mistake rather replaces a word by another.

This criterion, which we have to minimize, is measured by the occurrence of such words in the sentence. Each alternative x is evaluated by a score $f_3(x)$ that represents the number of words with restricted morphological features to be corrected.

4 Criteria aggregation by the TOPSIS method

4.1 Main steps of TOPSIS

TOPSIS is a multiple criteria decision making method (MCDM) devoted to cardinal criteria.

According to this method, the best solutions are defined to be those which are farthest from the negative-ideal point (the alternative with worst scores on all criteria) as well as closest to the ideal point (the alternative which has the best scores on all criteria). The ideal point and the negative-ideal point can be two artificial (not feasible) alternatives.

The various steps of TOPSIS may be summarised as follows ((Yoon and Hwang 81), (Hwang and Yoon 85)) :

step0: Construction of the decision matrix

Let $Y=(y_{ij})$, $i = 1 \dots n$, $j = 1 \dots q$ be the decision matrix such that $y_{ij} = f_j(x_i)$.

y_{ij} is the outcome of the alternative x_i with respect to the criterion fj and Y represents the outcomes of each alternative in terms of all criteria.

step1: Construction of the normalised decision matrix

This step tries to transform the various attribute dimensions into non-dimensional attribute in order to allow comparison across the attributes.

The corresponding element of the normalised decision matrix can be calculated as :

$$r_{ij} = \frac{y_{ij}}{\left\{\sum_{i=1}^{n} y_{ij}^2\right\}^{1/2}} \qquad i = 1,\ldots,n \qquad j = 1,\ldots,q$$

step2: Construction of the weighted normalised decision matrix

This matrix is obtained by multiplying each column of the normalised decision matrix with its associated weight (α_j). An element of the new matrix will be :

$$v_{ij} = \alpha_j r_{ij} \qquad i = 1,\ldots,n; \quad j = 1,\ldots,q$$

step3: Determination of ideal and negative-ideal solutions

The ideal solution (x^+) is defined as :

$$x^+ = \left\{v_1^+, v_2^+, \ldots, v_j^+, \ldots, v_q^+\right\}$$

$$\text{where } v_j^+ = \left\{\max_i v_{ij}, j \in J, \min_i v_{ij} \quad j \in J'\right\}$$

J is the set of criteria to be maximized (frequency criterion) and J' is the set of criteria to be minimized (typographic criterion and morphological criterion).

The negative-ideal solution (x^-) is defined as :

$$x^- = \left\{v_1^-, v_2^-, \ldots, v_j^-, \ldots, v_q^-\right\}$$

$$\text{where } v_j^- = \left\{\min_i v_{ij}\, j \in J, \max_i v_{ij} \quad j \in J'\right\}$$

step4: Calculation of the separation measure

This step tries to measure the separation (in terms of Euclidean distance) of each alternative from the ideal solution as follows:

$$S_i^+ = \left\{\sum_{j=1}^{q} (v_{ij} - v_j^+)^2\right\}^{1/2} \qquad i = 1,\ldots,n$$

Similarly, the separation from the negative-ideal solution is given by :

$$S_i^- = \left\{\sum_{j=1}^{q} (v_{ij} - v_j^-)^2\right\}^{1/2} \qquad i = 1,\ldots,n$$

step5: Calculation of the relative closeness to the ideal solution

The relative closeness of an alternative x_i with respect to x^+ is defined by :

$$C_i^+ = \frac{S_i^-}{\left(S_i^+ + S_i^-\right)} \qquad 0 < C_i^+ < 1 \qquad i = 1, \ldots, n$$

Then the preference order can be obtained according to the descending order of C_i^+ and the best alternative will be defined as the one which is closer to x^+ than to x^-.

4.2 Weighting the different criteria

In the following we present subjective weighing criteria experimented on a variety of real sentences.

Determination of the weight of the frequency criterion (α_1)

The frequency criterion is more important when the difference in terms of score between the best alternative and the other ones is very important.

Then, α_1 may be defined by :

$$\alpha_1 = \sum_{i=1}^{n} \left(\operatorname{Max}_i f_1(x_i) - f_1(x_i)\right)$$

Determination of the weight of the morphological criterion (α_2)

Our experimental study of test sentences shows that α_2 depends on α_1 if ($\alpha_1 \neq 0$) otherwise it depends on the frequency of the alternatives :

If $\operatorname{Max} f_1(x_i) > 5 \operatorname{Max} f_2(x_i)$ then α_1 is more important than α_2.

If $\operatorname{Max} f_1(x_i) \leq 5 \operatorname{Max} f_2(x_i)$ then α_2 is more important than α_1.

α_2 may be defined as follows :

if $\alpha_1 \neq 0$ then

$$\alpha_2 = \begin{cases} \frac{1}{5}\alpha_1 & \text{if } \operatorname{Max} f_1(x_i) > 5 \operatorname{Max} f_2(x_i) \\ \frac{1}{\alpha_1} & \text{if } \operatorname{Max} f_1(x_i) \leq 5 \operatorname{Max} f_2(x_i) \end{cases}$$

if $\alpha_1 = 0$ then

$$\alpha_2 = \begin{cases} \frac{1}{5} \operatorname{Max} f_1(x_i) & \text{if } \operatorname{Max} f_1(x_i) > 5 \operatorname{Max} f_2(x_i) \\ \frac{1}{\operatorname{Max} f_1(x_i)} & \text{if } \operatorname{Max} f_1(x_i) \leq 5 \operatorname{Max} f_2(x_i) \end{cases}$$

Note that (1/5) represents the trade-off between the morphological criterion and the frequency criterion.

Determination of the weight of the typographic criterion (α_3)

α_3 depends on the typographic gaps ($\sum_{i=1}^{n} (\max_i f_3(x_i) - f_3(x_i))$) and conversely depends on the frequency gaps ($\sum_{i=1}^{n} (\max_i f_1(x_i) - f_1(x_i))$). α_3 may be done by :

$$\alpha_3 = \begin{cases} \dfrac{\sum_{i=1}^{n} (\max_i f_3(x_i) - f_3(x_i))}{\alpha_1} & \text{If } \alpha_1 \neq 0 \\ \dfrac{\sum_{i=1}^{n} (\max_i f_3(x_i) - f_3(x_i))}{\max_i f_1(x_i)} & \text{If } \alpha_1 = 0 \end{cases}$$

Note that the weights (α_1 , α_2 , α_3) are calculated on the basis of the normalised decision matrix and since they must satisfy the constraints :

$0 \leq \alpha_j \leq 1$ and $\sum_{j=1}^{q} \alpha_j = 1$, we will normalise them by : $\alpha_j = \dfrac{\alpha_j}{\sum_{j=1}^{q} \alpha_j}$

4.3 An illustrative example

Let us consider the sentence : 'هذا الرجل الغنية إستأجر حانوتين ووضعا متاعهما فيهما ' (this $_{masc.sing.}$ rich $_{fem.sing.}$ man $_{masc.sing.}$ rented$_{masc.sing.}$ two shops and put $_{masc.duel}$ their $_{masc.duel}$ goods in them)

The result of the "Extended Syntagmatic Analysis" is given by the following extended syntagms :

SE_1= { هما (their), وضعا (put), إستأجر (rented), الغنية (rich), الرجل (man), هذا (this) }

SE_2= { هما (them), حانوتين (two shops) }

Let us consider SE_1. The features of SE_1 may be represented by the following figure which shows the errors in gender and number.

	هذا This	الرجل man	الغنية rich	إستأجر rented	وضعا put	هما their	
Gender	M	M	F	M	M	M	→ Error in gender
Number	S	S	S	S	D	D	→ Error in number
Tense	X	X	X	P	P	X	
Presonal pronoun	3	3	3	3	3	3	

Clearly, there are three correction alternatives :

x_1: Masculine, singular

x_2: Masculine, duel

x_3: Feminine singular

The scoring of these alternatives in terms of the criteria is given in the following decision matrix :

	x_1	x_2	x_3
Frequency Criterion	3	2	1
Morphological Criterion	0	0	1
Typographic Criterion	4	1,75	7,25

According to this matrix, we can conclude that x_3 is dominated by x_1 and x_2. x_1 and x_2 are two efficient solutions.

step1: Construction of the normalised decision matrix

	x_1	x_2	x_3
Frequency Criterion	0,80	0,534	0,267
Morphological Criterion	0	0	1
Typographic Criterion	0,472	0,206	0,856

step2: Construction of the weighted normalised decision matrix

The weights of each criterion are respectively :

$$\alpha_1 = (0,8 - 0,53) + (0,8 - 0,267) = 0,8 \qquad \alpha_1 = 0,239$$

$$\alpha_2 = \frac{1}{0,8} = 1,25 \xrightarrow{\text{Normalised weights}} \alpha_2 = 0,373$$

$$\alpha_3 = \frac{(0,856 - 0,472) + (0,856 - 0,206)}{0,8} = 1,29 \qquad \alpha_3 = 0,386$$

The normalised decision matrix is the following :

	x_1	x_2	x_3
Frequency Criterion	0,191	0,127	0,063
Morphological Criterion	0	0	0,373
Typographic Criterion	0,182	0,079	0,331

step3: Determination of ideal and negative-ideal solutions

Ideal solution V^+	***Anti-ideal solution*** V^-
0,191	0,063
0	0,373
0,079	0,331

step4: Calculation of the separation measure

	x_1	x_2	x_3
S+	0,102	0,063	0,468
S-	0,422	0,455	0

step5: Calculation of the relative closeness to the ideal solution

	S+	*S-*	*C+*	***Ranking***
x_1	0,102	0,422	0,804	2
x_2	0,063	0,455	0,876	1
x_3	0,468	0	0	3

x_2 has the best ranking, thus the best correction is obtained by lining up all the words of the erroneous sentence by the Masculine- Duel :

"هـذان الرجـلان الغنيـان إسـتأجرا حـانوتين ووضعـا متاعهمـا فيهمـا (these masc. Duel rich masc. Duel men masc. Duel rented masc. Duel two shops and put masc.Duel their masc.Duel goods in them).

5 Preliminary Experiment

A prototype implementation of the proposed method called 'DECORA' is developed using the C++ programming language with WINDOWS environment.

In order to evaluate the DECORA performance, 300 sentences are chosen from real texts written by secondary school students. The sentences are various : they contain from 1 to 4 extended syntagms and each syntagm contains a maximum of 9 words.

The sentences are corrected by a human expert whom we ask to classify them in the three following classes :

Class 1 : sentences which corrections are obvious.

Class 2 : sentences which corrections are not obvious and are somewhat challenging.

Class 3 : sentences which corrections need a very careful analysis to choose among the possible ones.

The same sentences are processed by DECORA. The results of the comparison between DECORA and the expert corrections are given in the following table :

(A) \ (B)	1	2	≥3
1	100 %	0 %	0 %
2	85 %	13 %	1 %
3	74 %	21 %	5 %

(A) : Sentences classes

(B) :Ranks of the corrections proposed by DECORA which are similar to the expert corrections.

As shown in this table, for the first sentences class, all corrections proposed by DECORA as the best ones are similar to those given by the human expert. 85 % of sentences of the second class are corrected in the same way by DECORA and the expert. This percentage decreases to 74 % for the third class. DECORA's best proposed corrections of the reminding sentences are different from those of the expert. Generally, these sentences have more than one plausible correction (i.e., they have very close scores). The expert may in some cases be hesitating between two or more possible corrections and then the choice of the best one is almost made at random. However, DECORA can make distinction between these alternatives by ranking them according to their scores.

Note that in some cases of the third class, different human experts may have different opinions about the best correction. In fact they may disagree with the relative importance of each criteria.

We think that the obtained results are very satisfying and we hope to obtain better results by studying more real sentences in order to improve the criteria weights.

6 Conclusion

As the correction process of agreement errors is not usually a simple task since the choice of the best correction alternative requires a careful analysis, we think that the use of a multicriteria approach to guide the correction process is very interesting.

In this paper, we proposed three main criteria (the frequency criterion, the morphological criterion and the typographic criterion) to rank the correction alternatives of Arabic sentences. We presented the techniques of scoring the correction alternatives in terms of the considered criteria. Finally, we showed that using TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) as an aggregation method of the considered criteria is well appropriated for our problem and the obtained results are very satisfying.

Acknowledgements

The authors are grateful to Mr Foued Ben Abdelaziz : professor at the management school "Institut Superieur de Gestion de Tunis" for his interest to this work and his comments on the preliminary version of the paper.

References

BELGUITH L. and BEN HAMADOU A., "Un vérificateur et correcteur des accords pour l'arabe non voyellé", Actes de la 3ème conférence annuelle sur le traitement automatique du langage naturel, 22-24 Mai, Marseille, 1996.

BEN HAMADOU A., "Vérification et correction automatique par analyse affixale des textes écrits en langage naturel : le cas de l'arabe non voyellé", Thèse d'Etat en informatique, Faculté des Sciences de Tunis, 1993.

BEN HAMADOU A. and BELGITH L., "Une méthode de vérification et de correction des accords appliquée à l'arabe non voyellé", Actes de la 1ère conférence internationale NLP+IA, Moncton, Canada, 4-6 Juin 1996.

BLACHE P., "L'analyse par filtrage ascendant: Une stratégie efficace pour les grammaires syntagmatiques généralisées", Actes de la 10ème conférence internationale sur le traitement du langage naturel et ses applications, Avignon, 1990.

BLACHE P., "Problèmes d'analyse syntaxique pour les correcteurs grammaticaux", Actes de la 11ème conférence internationale sur le traitement du langage naturel et ses applications, Avignon, 1991.

BOLIOLI A., DINI L. and MALNATI G., "Parsing Italian with a robust constraint grammar", International Conference on Computational Linguistics, Coling 1992.

COCH J.and MORIZE G., "Un analyseur conçu pour le traitement d'erreurs et ambiguïtés", Actes de la 10ème conférence internationale sur le traitement du langage naturel et ses applications, Avignon, 1990.

GENTHIAL D., COURTIN J. and KOWARSKI I., "Contribution of a category hierarchy to the robustness of syntactic parsing", Coling' 90, pp. 139-144, Helsinki, 1990.

GENTHIAL D and COURTIN J., "From Detection / Correction to computer Aided Writing", Coling'92, pp. 1013-1018, Nantes, 1992.

GENTHIAL D., COURTIN J. and MENEZO J., "Towards a More User-Friendly Correction", CoLing'94, pp. 1083-1088, 1994.

HWANG C.L. and YOON K., "Multiple attribute decision Making- Methods and applications : A State of the art survey", Springer-Verlag, New York, 1981.

LAPALME G. and RICHARD D., "Un système de correction automatique des accords des participes passés", Technique et Science Informatique, N°4, 1986.

LAPALME G.and LACOURTE R., "Une implantation informatique du français fondamental", Technique et science Informatique, N° 5, 1988.

PO-LUNG YU, "Multiple-criteria Decision Making concepts, Techniques, and extensions", PLENUM Press, NEW YORK, 1985.

RAVIN Y., "Grammar errors and style weakness in a text-critiquing system", IEEE Transactions on Professional Communication, Vol. 31, N° 3, September 1988.

STRUBE DE LUMA V.L., "Contribution à l'étude du traitement des erreurs au niveau lexico-syntaxique dans un texte écrit en français", Thèse de doctorat, institut IMAG, Université JOSEPH FOURIER, 1990.

VERONIS J., "Morphosyntactic correction in natural language interfaces", 12th International Conference on Computational Linguistics Coling, 1988.

VERONIS J., "Error in natural language dialogue between man and machine", Internatonal Journal Man - Machine Studies, Vol. 35, pp. 187-217, 1991.

VOSSE T., "Detecting and correcting morpho-syntactic errors in real texts", Proceedings of the third conference on applied natural language processing, Trento Italy, 31 March - 3 April, pp. 111-118, New York ACL, 1992.

YOON K. and HWANG C., "Manufacturing plant location analysis by multiple attribute decision making", International journal of production research, Vol. 23, N°2, pp. 345-359, 1985.