Summarization at LARIS Laboratory

JAOUA KALLEL Fatma*, JAOUA Maher**,

BELGUITH HADRICH Lamia***, BEN HAMADOU Abdelmajid******

LARIS laboratory

B.P. 1088- 3018 Sfax - Tunisia

(*)Fatma.kallel@component.zzn.com, (**) Maher.jaoua@fsegs.rnu.tn

(***)l.blguith@fsegs.rnu.tn, (****)Abdelmajid.benhamadou@isimsf.rnu.tn

Abstract

We present in this paper the ExtraNews system, currently being developed at Laris Laboratory in the University of Multimedia and Computer Science in Sfax-Tunisia. This system produces a very short and short summary from single/multiple news. The experimented method in ExtraNews is based on a new point of view for the summarization process, which advocates two principle steps: generation and classification. The first step uses the text sentences to produce a population of summaries. The second step concerns the evaluation of each summary according to global criteria in order to select the best one. This method is evaluated in the current year's Document Understanding Conference DUC 2004 and obtains good results for the automatic evaluation.

1 Introduction

To further progress in summarization and enable researchers to participate in large-scale experiments, the NIST[1] is performing an evaluation series in the text summarization field, called the Document Understanding Conference (DUC). The DUC 2004 evaluation corpus includes 2000 document sets from TDT and TREC collections (each set contains 10 documents in average). There are several tasks in the DUC'2004:

- Task 1: concerns very short single-document summaries. The length of the summary doesn't exceed 75 characters.
- Task 2: consists to produce a short-multi document summary (<=665 character). Summaries exceeding the size limit are truncated.
- Task 3: concerns to produce a very short summaries (from each document obtained after an automatic translation (priority 1) and manual translation (priority 2).
- Task 4: concerns to produce a short summaries from multi-document obtained after an automatic translation (priority 1) and manual translation (priority 2). This task explores summarization of noisy input produced by a machine translation (from Arabic to English).
- Task 5: this task proposes to generate a short multi-document summaries focused by question. The question proposed for this task concerns the biography about the described person in the set of articles.

We present in this paper evaluation results of our system called ExtraNews in DUC'2004 conference. The ExtraNews is developed in LARIS Laboratory at the University of Multimedia and Computer Sciences in

[1] National Institute of Standards and Technology

Sfax-Tunisia. The evaluated version in DUC 2004 is the english version after french version (Jaoua, 2002), (Jaoua, 2003a) for news and French version for scientific paper (Jaoua, 2000), (Jaoua, 2003b).

2 The proposed method

The overview of the automatic summarization research shows that the clustering step is a common step in summarization process (Radev, 2003), (Mckeown, 2003). This step is introduced in order to eliminate the information redundancy, which results from the multiplicity of the original documents. However, the redundancy problem cannot be totally solved due to the fact that the clustering process can't maintain the disjunction between clusters. Thus, the quality of the produced summaries depends strongly on the redundancy problems.

To avoid these problems and to enhance the summary quality, we propose, an original summarization process, which operates on the summary on its totality and not on its independent parts (e.g., sentences or phrases) (Jaoua, 2000). Therefore, we consider the summarization process as an optimization problem where the optimal summary is chosen among a set of summaries formed by the conjunction of the original articles sentences (for short summaries) or phrases (for very short summaries).

The experimented approach in ExtraNews system proposes to first generate a population of summaries, then to evaluate and classify them in order to produce the best summary (see figure1). We define the best summary as the set of sentences (or phrases) maximizing the information quantity and covering the important concepts (Jaoua, 2000).

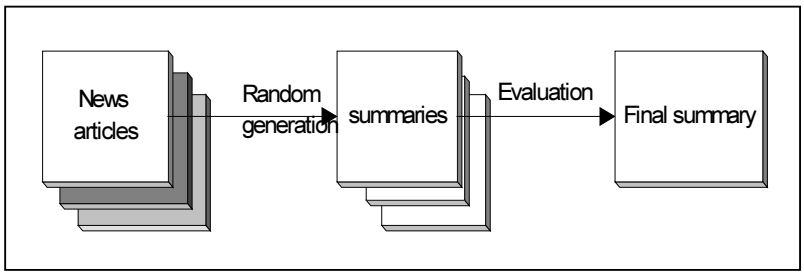

Figure 1. principle of the proposed method

Thus, the final summary represents a sentence set considered to be the best one according to the specific criteria mentioned above. Some criteria are to be maximized (e.g., information coverage) and others minimized (e.g., information redundancy).

2.1 ExtraNews Architecture

ExtraNews system is designed to summarize single and multi news. It's composed of three main modules (see Figure 2): the pre-processing module, the statistical module, and the selective module.

- **The pre-processing module**: This module consists of the article segmentation into sentences. Thus it works as a tokenizer and mainly uses the punctuation marks to locate the sentence boundaries. For the very short summaries, this module produces a list of phrases using punctuation and stop list word. Phrases are located between stop list word and/or punctuation mark. For the task 3 and 4 the pre-processing module chooses the three first sentences with the highest score produced by the automatic translator system after deleting sentences with unknown translated word.

- **The statistical module:** It concerns the computation of the words frequency in the news articles. For this purpose, it generates the frequent word list. This list will be used to determine the weight and the coverage of the source article sentences or phrases. This weight is considered in the classification process of the generated summaries in order to choose the best extract.

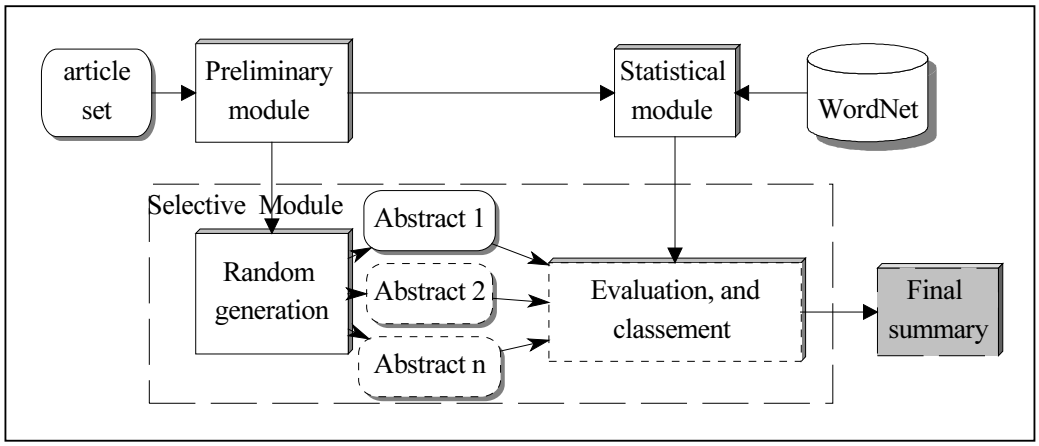

Figure2. ExtraNews atchitecture

- **The summary selection module:** This module operates on two stages: the generation and the classification of the summaries in order to choose the best one. Thus, we define an evaluation function, which we try to maximize (i.e., as an optimization problem). We generate the summaries with a genetic algorithm (see figure 3) that starts from a random solution, and then builds, in each stage, a set of solutions and evaluates them (Goldberg, 1989).

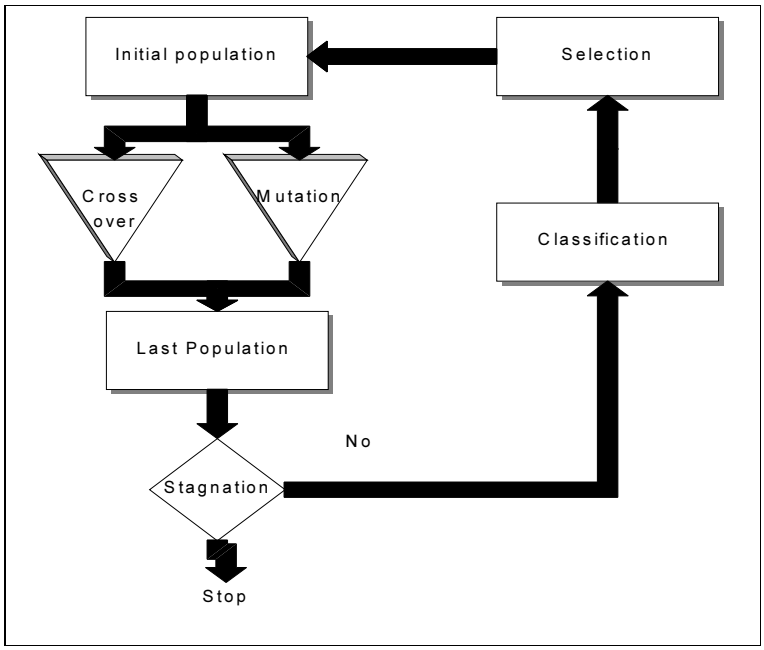

Figure 3. Selective module

The genetic algorithm produces, for each iteration, a population of summaries while combining, randomly, sentences (for the multi-documents task) or phrases (for the single document summarization) of the various articles and applying crossover and mutation operators. It assigns, for each generated summary, an objective value, which depends on the dominance of the summary according to the statistical criteria: length criterion, weight criterion, and coverage criterion. The solution is obtained when there is a stagnation of the selected solution. Figure 4 presents an example of summary obtained after applying the genetic algorithm for the multi-document summarization (Task 2).

```
<MULTI DOCSET="d1003t">
More than ten thousand people were killed while 13 thousand are
considered missing, according to official sources in the five
inflicted countries, which are Guatemala, Honduras, Nicaragua,
Salvador and Costa Rica, which suffered the least damages.
Various sources announced today, Thursday, that at least 25
villages or towns on the Caribbean Coast in Honduras disappeared
under the waters of the Iguana river, whose level had risen
significantly due to the heavy rains cause by Hurricane "Mitch".
Throughout Central America, but particularly in Honduras and
Nicaragua, tens of towns were erased from the map and tens of
thousands of homes were destroyed.
</MULTI>
```

Fig 4. Example of ExtraNews output

2.2 Fitness value

After applying the genetic operators with random values, we obtain a number of intermediary solutions. The classification of this intermediary population is based on statistical criteria. Dominated summary (according to Pareto sense) gets a low fitness value and then is removed from the population [SRI 93]. The process is repeated until the entire population is classified.

Note that summary Ei dominates summary Ej if all features $f_x i$ of the summary Ei are greater or equal to the same features $f_x j$ of the summary Ej and there is at least x where $f_x i$ is greater then the same feature $f_x j$. In (Jaoua, 2000) and (Enrique, 2003) the fitness value is calculated by an aggregation function that combines features but there is any method available to determine feature weight. But when we consider the dominance criteria we are able to classify the importance of each feature.

We present in the following, the different criteria, in order of their importance, used in the classification process and their associated coefficients.

- The coefficient ω_1 associated to the "length" criterion:

If $0,9 * L_E \leq \sum_{i=1}^{m} L(ph_i) \leq L_E$ Then $\omega_1 = 1$ ($L_E = 75$ for very short summarization task)

If $\sum_{i=1}^{m} L(ph_i) < 0,9 \times L_E$ Then $\omega_1 = \frac{\sum_{i=1}^{m} L(ph_i)}{L_E}$ ($L_E = 665$ for very short summarization task)

If $\sum_{i=1}^{m} L(ph_i) > L_E$ Then $\omega_1 = 0$

Where: $L(ph_i)$: The length of the sentence I;

L_E: The summary length fixed by the user;

m: The sentence number in the summary.

- The coefficient ω_2 associated to the "Coverage" criterion:

$$\omega_2 = \frac{\sum M_{ext}}{\sum M_{doc}}$$

Where: M_{ext}: keywords of the summary;

M_{doc}: keywords of the articles set.

Keywords are selected from the list of words frequency in the documents set.

- The coefficient ω_3 associated to the weight criterion:

$$\omega_3 = \frac{\sum P_{ext}}{Max(P_{pop})}$$

Where: P_{ext}: Weight of a sentence of the summary;

Max(P_{pop}) Maximal summary weight in the population.

3 ExtraNews results

3.1 Automatic Evaluation using ROUGE system

The ExtraNews results for task 1,2,3,4 and 5 are compared to human summaries. The comparison is done using ROUGE system (Lin, 2003) and presented in the Annex 1 and Annex 2. Preliminary results obtained indicate that ExtraNews system does fine in the DUC 2004 evaluation. It ranks in a good position for short summaries (< 665 character). For example, ExtraNews obtain the 1st rank in the task 4 in Rouge 1, ROUGE 2, ROUGE L and ROUGE W1-2, the 2nd rank in the same task for ROUGE 3 and ROUGE 4.
For the task 5 our system obtains the 2nd position in ROUGE 1, ROUGE L and ROUGE W1-2 . For the task 2 our system obtains the 4th position for ROUGE 1, ROUGE 2, ROUGE L, and ROUGE W1-2.
But ExtraNews is does badly for very short summaries (less than 75 character) (rank 14 in the task 1, rank 4 in the task 3 for automatic translated news, and rank 8 for the task 3 which deals with manual translated news). One explanation of this badly result could be the fact that the phrases considered in the segmentation process are not very well suitable for very short summaries.

3.2 Manual Evaluation

The manual evaluation concerns only the task 2 and 5 in DUC'2004. ExtraNews results for task 2 are detailed in annex 3 and those for task 5 are in annex 4. In those tables Q1.. Q7 represent the average score obtained for the several question asked for the human evaluator. In the following table we indicate the linguistic subject of the questions used in this evaluation. The main coverage of task 2 is 0.23 and ExtraNews is ranked 9/16 systems. In this task, our system obtains the highest average score in the question 5 concerning the readability of summary when entities are re-mentioned in overly explicit way. Extranews obtains good result for the question 3 dealing with redundancy. This result is due to the utilization of the coverage criteria in the classification step.

The main coverage for task 5 is 0.21, and our system is ranked in the second position. In this task, we obtain the best average score in the question 3 and the question 5 like the task 2. The results for other question are evident thus our system doesn't use any linguistic resources.

4 Conclusion and future work

We presented in this paper results of ExtraNews system in the DUC'2004 conference evaluation. The applied method in our system is based only on statistical criteria to select sentences (or phrases) that are most likely to be relevant to the cluster topic or document topic.
The evaluation of ExtraNews shows that it produces high quality of short summaries covering the most important information in the source articles, without any linguistic resources. But for the very short summaries we must improve results of our system by developing other components. In order to enhance the summary quality, we prospect to add new techniques to handle the coherence and the cohesion problems. We suggest developing "revising" module. This module will be used for the first ten best summaries of the generated summary population, and then they will be ranked again in order to select the best summary.

References

DUC 2003. Document Understanding Conference, *Workshop on Text Summarization 2003*, USA. http://www-nlpir.nist.gov/projects/duc/

Enrique A. Pilar R. 2003. "Description of the UAM system for generating very short summaries at DUC-2003". *Workshop on Text Summarization (DUC 2003)*. Edmonton, Canada.

Goldberg D.E., 1989, Genetic algorithms in search, optimization, and machine learning, Addison-Wesley, New York, 1989.

Jaoua K.F., Jaoua M., Ben Hamadou A., 2003a. "Une méthode de condensation automatique des documents multiples : cas des articles de presse", *in the CIDE.6 International Conference on Electronique Document.* Caen, France, Novembre 2003.

Jaoua M., A., Ben Hamadou, 2003b, "Automatic Text Summarization of Scientific Articles Based on Classification of Extract's Population", *in proceeding of Cicling'03 : Conference on Intelligent Text Processing and Computational Linguistics*, Mexico, 16-22 February 2003.

Jaoua K.F., 2002. Une méthode de condensation automatique des dépêches multiples , mémoire de DEA, Faculté des sciences Economiques et de Gestion de Sfax, Tunisie

Jaoua M., Ben Hamadou A., 2000. "Automatic Text Extraction Based on Classification of Extract Population", *in proceeding of ACIDCA'2000*, pp. 117-123, 22-24 Mars, Monastir, Tunisia, 2000.

Lin, Chin-Yew and E.H. Hovy 2003. "Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics ». *In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, May 27 - June 1, 2003.

McKeown K. and al. 2003. "Columbia at the Document Understanding Conference 2003" , *in Workshop on Text Summarization (DUC 2003)*, Edmonton, Canada, 2003.

Radev R., Otterbacher J., Qi H., Tam D. 2003. "MEAD ReDUCs: Michigan at DUC 2003". *Workshop on Text Summarization (DUC 2003)*. Edmonton, Canada.

Srivinas, N., Deb, K., 1993. "Multiobjective optimization using nondominated sorting in genetic Algorithms". Technical report, Department of Mechanical Engineering, Institute of Technology India, 1993.

Task 1 (Sys id = 18)								
N_gram	**Average**	**95% CI Lower**	**95% CI Upper**	**Rank[2] /18**	**Best Human**	**Worst Human**	**Best system**	**Worst System**
Rouge1	0,15322	0,14516	0,16128	14	0,31478	0,25017	0,22101	0,12062
Rouge 2	0,03661	0,03224	0,04098	14	0,10144	0,06563	0,06377	0,00735
Rouge 3	0,00911	0,00678	0,01144	11	0,03893	0,01955	0,02134	0,00017
Rouge 4	0,00204	0,00105	0,00303	14	0,0142	0,0061	0,0073	0,00007
Rouge L	0,13470	0,12756	0,14184	14	0,2663	0,21402	0,19443	0,10647
Rouge W1-2	0,08218	0,07816	0,0862	15	0,15369	0,12397	0,11748	0,06537
Task 3 Priority 1 (Sys id = 20)								
N_gram	**Average**	**95% CI Lower**	**95% CI Upper**	**Rank[2] /11**				
Rouge1	0,20166	0,18765	0,21567	4				
Rouge 2	0,04441	0,03649	0,05233	3				
Rouge 3	0,00862	0,00439	0,01285	5				
Rouge 4	0,00128	-0,00002	0,00258	5				
Rouge L	0,16715	0,15564	0,17866	4				
Rouge W1-2	0,10254	0,09606	0,10902	3				
Task 3 priority 2 (Sys id =21)								
N_gram	**Average**	**95% CI Lower**	**95% CI Upper**	**Rank[2] /11**				
Rouge1	0,20713	0,19282	0,22144	8				
Rouge 2	0,05771	0,04922	0,0662	5				
Rouge 3	0,01692	0,01268	0,02116	7				
Rouge 4	0,0029	0,00132	0,00448	7				
Rouge L	0,18637	0,17316	0,19958	7				
Rouge W1-2	0,1147	0,10728	0,12212	7				
Task 3 priority 2 without Headline (Sys id =21)								
N_gram	**Average**	**95% CI Lower**	**95% CI Upper**	**Rank[2] /11**				
Rouge1	0,23114	0,21665	0,24563	8				
Rouge 2	0,06508	0,05646	0,07370	5				
Rouge 3	0,01842	0,01375	0,02309	5				
Rouge 4	0,00377	0,00176	0,00578	7				
Rouge L	0,19452	0,18310	0,20594	7				
Rouge W1-2	0,11905	0,11242	0,12568	7				

Annex 1 : ExtraNews Results for tasks dealing with a single document

[2] This rank is obtained by comparing systems having the same priority as ExtraNews.

Task 2 : Multidocument summarization								
N_gram	Average	95% CI Lower	95% CI Upper	Rank /16	Best Human	Worst Human	Best system	Worst System
Rouge1	0.37386	0.36080	0.38692	**4**	0,41828	0,38902	0,38224	0,24190
Rouge 2	0.08026	0.07119	0.08933	**6**	0,10654	0,08595	0,09216	0,01876
Rouge 3	0.02564	0.02031	0.03097	**7**	0,03574	0,02427	0,03529	0,00277
Rouge 4	0.01016	0.00691	0.01341	**10**	0,01280	0,00783	0,01658	0,00078
Rouge L	0.38368	0.36888	0.39848	**4**	0,43380	0,40631	0,38950	0,27630
Rouge W1-2	0.13146	0.12629	0.13663	**4**	0,14804	0,13805	0,13378	0,09358
Task 4 Priority 1: Automatic translation								
N_gram	Average	95% CI Lower	95% CI Upper	Rank /11				
Rouge1	0.38654	0.36352	0.40956	**2**				
Rouge 2	0.09063	0.07794	0.10332	**6**				
Rouge 3	0.02393	0.01706	0.03080	**6**				
Rouge 4	0.00762	0.00388	0.01136	**6**				
Rouge L	0.34591	0.32795	0.36387	**3**				
Rouge W1-2	0.11621	0.10980	0.12262	**4**				
Task 4 Priority 2: Manual translation sys id =23								
N_gram	Average	95% CI Lower	95% CI Upper	Rank /11				
Rouge1	0.41577	0.39333	0.43821	**1**				
Rouge 2	0.12828	0.10994	0.14662	**2**				
Rouge 3	0.04881	0.03674	0.06088	**3**				
Rouge 4	0.02199	0.01502	0.02896	**4**				
Rouge L	0.40749	0.38636	0.42862	**1**				
Rouge W1-2	0.13823	0.12995	0.14651	**1**				
Task 4 Priority 2 manual translation without headline sys id =23								
N_gram	Average	95% CI Lower	95% CI Upper	Rank /11				
Rouge1	0,41265	0,38840	0,4369	**1**				
Rouge 2	0,13214	0,11520	0,14908	**1**				
Rouge 3	0,05329	0,04161	0,06497	**2**				
Rouge 4	0,02587	0,01777	0,03397	**2**				
Rouge L	0,40716	0,38532	0,4290	**1**				
Rouge W1-2	0,13897	0,13059	0,14735	**1**				
Task 5 : Question focused summaries								
N_gram	Average	95% CI Lower	95% CI Upper	Rank /14	Best Human	Worst Human	Best system	Worst System
Rouge1	0.34918	0.33543	0.36293	**2**	0,48990	0,37333	0,35495	0,26285
Rouge 2	0.07980	0.07064	0.08896	**3**	0,17486	0,10030	0,08571	0,04868
Rouge 3	0.02927	0.02289	0.03565	**5**	0,08149	0,03637	0,03280	0,01515
Rouge 4	0.01346	0.00980	0.01712	**7**	0,04812	0,01685	0,01641	0,00672
Rouge L	0.37188	0.35915	0.38461	**2**	0,49488	0,39716	0,37330	0,28435
Rouge W1-2	0.12554	0.12124	0.12984	**2**	0,16903	0,13429	0,12674	0,09843

Annex 2: ExtraNews Results for tasks dealing with a single document

SYS id	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Unmarked peer units	Mean Coverage
11	2,60	2,60	2,18	1,42	1,62	1,22	1,20	0,07	0,22
19	**3,29**	**2,90**	**1,49**	**2,59**	**1,04**	**1,65**	**1,73**	**0,69**	**0,23**
27	3,28	2,70	1,36	2,34	1,08	1,22	1,36	0,48	0,17
34	3,24	2,58	1,52	2,42	1,14	1,38	1,52	0,61	0,22
44	2,56	2,36	1,58	1,60	1,18	1,30	1,32	0,35	0,26
55	3,06	2,68	1,40	2,38	1,18	1,58	1,42	0,54	0,24
65	2,86	2,52	1,62	1,60	1,44	1,40	1,34	0,42	0,30
81	2,76	2,74	1,66	1,92	1,30	1,38	1,34	0,49	0,25
93	2,98	2,52	1,70	1,46	1,38	1,54	2,70	0,22	0,26
102	2,68	2,66	1,52	1,64	1,20	1,42	1,40	0,36	0,24
111	4,74	4,44	1,40	4,66	1,12	1,68	2,30	0,42	0,05
117	4,82	4,54	2,06	4,24	1,50	4,30	2,00	0,53	0,12
120	2,32	2,08	1,56	1,20	1,46	1,22	1,38	0,21	0,24
123	3,34	2,84	1,66	2,76	1,16	1,28	1,46	0,56	0,17
124	2,82	2,56	2,26	1,64	1,58	1,42	1,44	0,43	0,26
138	2,42	2,68	1,76	1,54	1,08	3,36	2,52	0,02	0,16

Annex 3: Manual evaluation of task N°2

SYS id	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Mean Coverage	Responsiveness score	Average Responsiveness
16	3,12	2,70	1,6	2,04	1,7	1,4	1,5	0,15	65	1,3
24	**3,86**	**3,40**	**1,4**	**2,98**	**1,30**	**1,9**	**1,9**	**0,21**	**58**	**1,16**
30	2,90	2,42	1,4	2,00	1,46	1,3	1,3	0,19	71	1,42
43	3,78	2,94	1,3	2,84	1,46	1,8	1,5	0,19	65	1,3
49	3,34	2,88	1,5	2,12	1,62	1,8	1,4	0,20	70	1,4
62	3,62	3,12	2,0	1,90	2,14	1,8	1,6	0,20	82	1,64
71	3,32	2,96	1,8	1,44	2,42	1,6	1,7	0,21	75	1,5
86	3,86	3,22	1,2	2,96	1,18	2,2	1,7	0,14	53	1,06
96	3,30	3,06	1,6	1,82	1,66	1,7	3,5	0,21	72	1,44
109	3,14	2,76	1,9	1,74	1,96	1,4	1,5	0,24	88	1,76
116	4,52	4,04	1,5	3,4	1,76	2,1	2,1	0,17	50	1
122	2,94	2,32	1,7	1,24	2,08	1,2	1,4	0,18	63	1,26
125	3,52	3,22	1,6	2,12	1,94	2,8	2,2	0,18	70	1,4
147	3,18	2,64	1,8	1,92	1,80	1,2	1,4	0,21	77	1,54

Annex 4: Manual evaluation of task N°5