Human judgment as a parameter in evaluation campaigns

Jean-Baptiste Berthelin and **Cyril Grouin** and **Martine Hurault-Plantet** and **Patrick Paroubek**
LIMSI-CNRS
BP 133
F-91403 Orsay Cedex
firstname.lastname@limsi.fr

Abstract

The relevance of human judgment in an evaluation campaign is illustrated here through the DEFT text mining campaigns.

In a first step, testing a topic for a campaign among a limited number of human evaluators informs us about the feasibility of a task. This information comes from the results obtained by the judges, as well as from their personal impressions after passing the test.

In a second step, results from individual judges, as well as their pairwise matching, are used in order to adjust the task (choice of a marking scale for DEFT'07 and selection of topical categories for DEFT'08).

Finally, the mutual comparison of competitors' results, at the end of the evaluation campaign, confirms the choices we made at its starting point, and provides means to redefine the task when we shall launch a future campaign based on the same topic.

1 Introduction

For the past four years, the DEFT[1] (*Défi Fouille de Texte*) campaigns have been aiming to evaluate methods and software developed by several research teams in French text mining, on a variety of topics.

The different editions concerned, in this order, the identification of speakers in political speeches (2005), the topical segmentation of political, scientific and juridical corpora (2006), the automatic affectation of opinion values to texts developing an argumented judgment (2007), and the identification of the genre and topic of a document (2008).

Human judgment was used during the preparation of the last two campaigns, to assess the difficulty of the task, and to see which parameters could be modified. To do this, before the participants start competing via their software, we put human judges in front of versions of the task with various sets of parameters. This allows us to adjust the definition of the task according to which difficulties were encountered, and how judges agree together. These human judges are in small number, and belong to our team. However, results of the campaign are automatically evaluated with reference to results attached to the corpus from the start. This is because the evaluation of a campaign's results by human judges is expensive. For instance, TREC[2] international evaluation campaigns are supported by the NIST institute and funded by state agencies. In Europe, on the same domains, the CLEF[3] campaigns are funded by the European Commission, and in France, evaluation campaigns are also funded by projects, such as Technolangue[4]. DEFT campaigns, however, are conducted with small budgets. That means for us to have selected corpora that contain the desired results. For instance, in a campaign for topical categorization, we must start with a topically tagged corpus. By so doing, we also can, at the end of a campaign, compare results from human judges with results from competitors, using an identical

[1] See http://deft.limsi.fr/ for a presentation in French.

[2] http://trec.nist.gov

[3] http://www.clef-campaign.org

[4] http://www.technolangue.net

common reference.

In this paper, we describe experiments we performed with human judgments when preparing DEFT campaigns. We survey the various steps in the preparation of the last two campaigns, and we go through the detail of how human evaluation, performed during these steps, led us to the parametrization of these two campaigns. We also present a comparative analysis of results found by human judges and results submitted by competitors in the challenge. We conclude about the relevance of the human evaluation of a task, prior to evaluating software dedicated to this task.

2 Parametrization of the campaign

We were competitors in the 2005 and 2006 editions, and became organisators for the 2007 and 2008 campaigns. For both challenges that we organized, we went through the classical steps of the evaluation paradigm (Adda et al., 1999), to which we systematically added a step of human test of the task, in order to adjust those parameters that could be modified. The steps, therefore, are following:

1. thinking about potential topics;
2. choice of a task and collection of corpora;
3. choice of measurements;
4. test of the task by human judges on an extract of the corpus in order to precisely define its parameters;
5. launching the task, recruiting participants;
6. testing period;
7. adjudication: possibility of complaints about the results;
8. workshop that closes the campaign.

Whenever human judges have to evaluate the results of participants in a campaign, the main problems are about correctly defining the judging criteria to be applied by judges, and that judges be in sufficient number to vote on judging each document. Hovy et al. (2002) describe work toward formalization of software evaluation methodology in NLP, developed in the EAGLES[5] and ISLE[6] projects. For cost-efficiency reasons, automatic evaluation is relevant, and its results have sometimes been compared to results from human judges. For instance, Eck and Hori (2005) compare results of evaluation measurements used in automatic translation with human judgments on the same corpus. In (Burstein and Wolska, 2003), the authors describe an experiment in the evaluation of writing style and find a better agreement between the automatic evaluation system and one human judge, than between two human judges.

Returning to the DEFT campaign, once the task is chosen, the corpora are collected, and evaluation measurements are defined, there can remain some necessity of adjusting parameters, according to the expected difficulty of the task. This could be, for instance, the level of granularity in a task of topical segmentation, or which categories should be relevant in a task of categorization. To get this adjusting done, we submit the task to human judges.

In 2007, the challenge was about the automatic affectation of opinion values to texts developing an argumented judgment (Grouin et al., 2007). We collected opinion texts already tagged by an opinion value, such as film reviews that, in addition to a text giving the judgment of the critic on the film, also feature a mark in the shape of a variable number of stars. The adjustable parameter of the task, therefore, is the scale of opinion values. The task will be more or less difficult, according to the range of this scale.

The 2008 campaign was about classifying a set of documents by genre and topic (Hurault-Plantet et al., 2008). The choice of genres and topics is a crucial one. Some pairs of topics or genres are more difficult to separate than other ones. We also had to find different genres sharing a set of topical categories, while corpora in French are not so very abundant. So we selected two genres, encyclopedia and daily newspaper, and about ten general topical categories. The parameter we had to adjust was the set of categories to be matched against each other.

3 Assessing the difficulty of a task

3.1 Calibration of an opinion value scale

In 2007, the challenge was about the automatic affectation of opinion values to texts developing an argumented judgment. In view of that, we collected four corpora that covered various domains:

[5] http://www.ilc.cnr.it/EAGLES96/home.html

[6] http://www.ilc.cnr.it/EAGLES96/isle/

reviews of films and books, of video games and of scientific papers, as well as parliamentary debates about a draft law.

Each corpus had the interesting feature of combining a mark or opinion with a descriptive text, as the mark was used to sum up the judment in the argumentative part of this text. Due to the diversity of sources, we found as many marking scales as involved copora:

- 2 values for parliamentary debates[7] (the representative who took part in the debate was either in favour or in disfavour of the draft law) ;
- 4 values for scientific paper reviews (*accepted as it stands – accepted with minor changes – accepted with major changes and second overall reviewing –rejected*), based on a set of criteria including interestingness, relevance and originality of the paper's content ;
- 5 values for film and book reviews[8] (a mark between 0 and 4, from bad to excellent) ;
- 20 values for video game reviews[9] (a global mark calculated from a set of advices about various aspects of the game: graphics, playability, life span, sound track and scenario).

In order to, first, assess the feasibility of the task, and to, secondly, define the scale of values to be used in the evaluation campaign, we submitted human judges to several tests (Paek, 2001): they were instructed to assign a mark on two kinds of scale, a wide one with the original values, and a restricted one with 2 or 3 values, depending on the corpus it was applying to. The results from various judges were evaluated in terms of precision and recall, and matched to each other by way of the Kappa coefficient (Carletta, 1996) (Cohen, 1960).

We present hereunder the values of the κ coefficient between pairs of human judges, and with the reference, on the video game corpus. The wide scale (Table 1) uses the original values (marks between 0 and 20), while the restricted scale (Table 2) relies upon 3 values with following definitions: class 0 for original marks between 0 and 10, class 1 for marks between 11 and 14, and class 2 for marks between 15 and 20.

Judge	**Ref.**	**1**	**2**	**3**
Ref.		0.17	0.12	0.07
1	0.17		0.03	0.05
2	0.12	0.03		0.07
3	0.07	0.05	0.07	

Table 1: Video game corpus: wide scale, marks from 0 to 20.

Judge	**Ref.**	**1**	**2**	**3**
Ref.		0.74	0.79	0.69
1	0.74		0.74	0.54
2	0.79	0.74		0.69
3	0.69	0.54	0.69	

Table 2: Video game corpus: restricted scale, marks from 0 to 2.

Table 1 and 2 show that agreement between judges varies widely when marking scales are modified. Table 1 shows that there is an insufficient agreement among judges on the wide scale, with κ coefficients lower than 0.20, while the agreement between these same judges can be considered as good on the restricted scale, with κ coefficients between 0.54 and 0.79 (Table 2), the median being at 0.74.

In order to confirm the validity of the change in scales, we used the κ to test how each judge agreed with himself, between his two sets of results (Table 3). Therefore, we compared judgments made by each judge using the initial value scale and converted towards the restricted scale, with judgments made by the same judge directly using the restricted value scale. This measurement shows the degree of correspondence between both scales for each judge. Among the three judges who took part in the test, the first and third one agree well with themselves, while for the second one, the agreement is only moderate.

Judge	**1**	**2**	**3**
1	0.74		
2		0.46	
3			0.70

Table 3: Video game corpus: agreement of each judge with himself when scales change.

We did the same for a second corpus, of film reviews. The test involved five judges, and the scale

[7] `http://www.assemblee-nationale.fr/12/debats/`

[8] `http://www.avoir-alire.com`

[9] `http://www.jeuxvideo.com/etajvbis.htm`

change was smaller, since it was from five values to three, and not from twenty to three. For this scale change, we merged the two lowest values (0 and 1) into one (0), and the two highest ones (3 and 4) into one (2), and the middle value in the wide scale (2) remained the intermediate one in the restricted scale (1). This scale change was the most relevant one, since, with 29.7% of the documents, the class of the middle mark (2) accounted for almost one third of the corpus. However, the two other groups of documents are less well balanced. Indeed, the lowest mark concerns less documents than the highest one: 4.6% and 10.3% respectively for the initial marks 0 and 1, while one finds 39.8% and 15.6% of documents for the marks 3 and 4. Grouping the documents in only two classes, by joining the middle class with the two lowest ones, would have yielded a better balance between classes, with 44.6% of documents for the lower mark and 55.4% for the higher one, but that would have been less meaningful.

Results from human judges are shown in the Tables 4 and 5 for both scales.

Judge	**Ref.**	**1**	**2**	**3**	**4**	**5**
Ref.		0.10	0.29	0.39	0.46	0.47
1	0.10		0.37	0.49	0.48	0.35
2	0.29	0.37		0.36	0.30	0.43
3	0.39	0.49	0.36		0.49	0.54
4	0.46	0.48	0.30	0.49		0.60
5	0.47	0.35	0.43	0.54	0.60	

Table 4: Film review corpus: wide scale, marks from 0 to 4

Judge	**Ref.**	**1**	**2**	**3**	**4**	**5**
Ref.		0.27	0.62	0.53	0.56	0.67
1	0.27		0.45	0.43	0.57	0.37
2	0.62	0.45		0.73	0.48	0.54
3	0.53	0.43	0.73		0.62	0.62
4	0.56	0.57	0.48	0.62		0.76
5	0.67	0.37	0.54	0.62	0.76	

Table 5: Film review corpus: restricted scale, marks from 0 to 2.

Agreements between human judges ranked from bad to moderate for the wide scale (the five original values in this corpus), while these agreements rank from insufficient to good in the case of the restricted scale with three values. We can see that differences induced by the scale change are much less important than with the video game corpus. This agrees well with the scales being much closer to each other.

By first performing a hand-made evaluation, and secondly, matching between themselves the results from the judges, we found a way to assess with greater precision the difficulty of the evaluation task we were about to launch. Concerning the first two review corpora (films and books, video games), we attached values good, average and bad to the three selected classes. The scale for scientific paper reviews was also restricted to three classes for which following values were selected: paper accepted as it stands or with minor edits, paper accepted after major edits, paper rejected. Finally, since its original scale had only two values, the corpus of parliamentary debates underwent no change of scale.

3.2 Choice of a topical category set

In order to determine which topical categories should be recognized in the 2008 task of classifying documents by genre and topic, we performed a manual evaluation of a sample of the corpus with 4 human judges. The sample included 30 Le Monde papers for the journalistic genre, and 30 Wikipedia entries for the encyclopedic genre. Only the title and body of each article was kept in the sample, and the tables were deleted. All marks of inclusion in either corpus were also deleted (references to Le Monde and Wikipedia tags).

The test ran this way: each article was put in a separate file, and the evaluators had to identify the genre and the topical category under which it was published. All articles were included in one set, which means evaluators had to choose, between all categories and genres, which ones to match with each document. This test was made with a first selection of 8 categories, shared by both genres, listed in Table 6.

Table 7 shows that results from human judges in terms of precision and recall were excellent on the identification of genre (F-scores between 0.94 and 1.00) and quite good on the identification of categories (F-scores between 0.66 and 0.82).

We also proceeded to the pairwise matching of results from human judges via the κ coefficient. Results show an excellent agreement of judges among themselves and with the reference for genre identification (Table 8). The agreement is mod-

Le Monde	**Wikipedia**
Notebook	*People*
Economy	*Economy*
France	*French Politics*
International	*International Politics*, minus category *French Politic*
Science	*Science*
Society	*Society*, minus subcategories *Politics, People,* *Sport, Media*
Sport	*Sport*
Television	*Television*

Table 6: Correspondence between categories from Le Monde and Wikipedia for the 8 categories in the test.

Judge	**1**	**2**	**3**	**4**
Genres	1.00	0.98	0.97	0.94
Categories	0.79	0.77	0.82	0.66

Table 7: F-scores obtained by human judges on the identification of genre and categories.

erate to good for categoy identification (Table 9). These good results led us to keep the corpora as they stood, since they appeared to constitute a good reference for the defined task. However, we made an exception for category *Notebook* (biographies of celebrities) which we discarded for two reasons. First, it is more of a genre, namely, ”biography”, rather than a topical category. Secondly, we found it rather difficult to assign a single category to articles which could belong in two different ones, as would be the case for the biography of a sportsman, which would fall under both categories *Notebook* et *Sport*.

Judge	**Réf.**	**1**	**2**	**3**	**4**
Réf.		1.00	0.97	0.93	0.87
1	1.00		0.97	0.93	0.87
2	0.97	0.97		0.90	0.83
3	0.93	0.93	0.90		0.87
4	0.87	0.87	0.83	0.87	

Table 8: κ coefficient between human judges and the reference: Identification of genre.

Judge	**Réf.**	**1**	**2**	**3**	**4**
Réf.		0.56	0.52	0.60	0.39
1	0.56		0.69	0.75	0.55
2	0.52	0.69		0.71	0.61
3	0.60	0.75	0.71		0.52
4	0.39	0.55	0.61	0.52	

Table 9: κ coefficient between human judges and the reference: Identification of categories.

Our task of genre and topic classification included two subtasks, one being genre and topic recognition for a first set of categories, the other one being only topic recognition for a second set of categories. Therefore, the corpus had to be divided in two parts. In order to find which categories had to go into which subcorpus, we decided to estimate, for each category, the difficulty of recognizing it. To do so, we calculated the precision and recall of each evaluator for each category. This measurement was obtained via a second evaluation of human judges, with a wider set of categories (by adding categories *Art* and *Literature*).

The ordering of categories by decreasing precision is following: *Sport* (1.00), *International* (0.80), *France* (0.76), *Literature* (0.76), *Art* (0.74), *Television* (0.71), *Economy* (0.58), *Science* (0.33), *Society* (0.26). This means no document in the *Sport* category was misclassified, and, contrariwise, categories *Science* and *Society* were the most problematic ones.

The ordering by decreasing recall is slightly different: *International* (0.87), *Economy* (0.80), *Sport* (0.75), *France* (0.70), *Art* (0.62), *Literature* (0.49), *Television* (0.46), *Society* (0.42), *Science* (0.33). Hence, articles in the *International* category were best identified. This ordering also confirms the difficulty felt by human judges concerning the categories *Society* and *Science*.

We decided to distribute the categories for each subtask according to a balance between easy and diffucult ones in terms of human evaluation:

- *Art, Economy, Sport, Television* for the subtask with both genre and category recognition;

- *France, International, Literature, Science, Society* for the subtask with only category recognition. For this second subset, we put together three categories which are topically close (*France, International* and *Society*).

4 Human judgments and software

4.1 Confirming the difficulty of a task

The 2007 edition of DEFT highlighted two main phenomena concerning the corpora involved in the task.

First, each corpus yielded a different level of difficulty, and this gradation of difficulty among corpora appeared both for human evaluators and competitors in the challenge (Paroubek et al., 2007).

	Judges	**Competitors**
Debates	0.77/1.00	0.54/0.72
Game reviews	0.73/0.90	0.46/0.78
Film reviews	0.52/0.79	0.38/0.60
Paper reviews	0.41/0.58	0.40/0.57

Table 10: Minimal and maximal strict F-scores between human evaluators and competitors in the challenge, 2007 edition.

During human tests, judges mentioned the great facility of finding about opinions expressed in the corpus of parliamentary debate. Next came corpora of video game reviews, and then of film and book reviews, whose difficulty was considered average, and last, the corpus of scientific paper reviews, which the judges perceived as particularly difficult. This gradation of difficulty among corpora was also found among competitors, following the same ordering of three levels of difficulty.

Secondly, the difficulties met by human evaluators are also found in the case of competitors. Upon finishing human tests, judges felt difficulties in evaluating the corpus of scientific paper reviews, yielding poor results. Now, the results of competitors on the same corpus are quite as poor, occupying exactly the same value interval as for human judges. Most competitors, by the way, obtained their worst results on this corpus.

The alikeness of results between judges and competitors reflects the complexity of the corpus: when preparing the campaign, we observed that reviews were quite short. Therefore, assigning a value had to rely upon a small amount of data. From that, we can derive a minimal size for documents to be used in this kind of evaluation. Moreover, a paper review can be seen as an aid for the author, to be expressed as positively as possible, even if it is also addressed to the Program Committee which has to accept or reject the paper. Therefore, the mark could prove more negative than the text of the review.

The case of comments about videogames is a different one. Indeed, giving a global mark on a scale of 20 is a difficult task. Therefore, this mark comes most often from a sum of smaller marks which rate either the whole document according to various criteria, or parts of this document. In our corpus, each reviewer rates the game according to several criteria, namely, graphics, playability, life span, sound track and scenario, from which a rather long text is produced, making the judgment an easier task to perform. However, the global mark differs from the sum of the smaller ones from various criteria, hence the difficulty for human judges to reckon this global mark on a scale of 20.

4.2 Confirmation of the expected success of competitors

Contrary to the 2007 edition, in which competitors obtained results that confirmed those of human judges, the 2008 edition gave them the opportunity to reach a higher level than human evaluators.

While genre identification yielded no special problem, either for human evaluators or for competitors, and the results obtained by both groups are similar, competitors reached better results than human judges in topical categorization.

Concerning genre identification, strict F-scores are situated between 0.94 and 1.00 for human judges, and between 0.95 and 0.98 for the best runs of competitors (each competitor was allowed to submit up to three collections of results, only the best one being used for the final ranking). As for topical categorization, strict F-scores go from 0.66 to 0.82 for human evaluators, and from 0.84 to 0.89 for best runs from competitors.

The equivalence of results on genre identification between judges and competitors can be explained by the fact that it was a simple, binary choice (the newspaper Le Monde vs. Wikipedia).

Contrariwise, competitors obtained better results in topical categorization, since machines have a stronger abstraction capacity than humans in presence of the 9 topical categories we defined (*Art, Economy, France, International, Literature, Science, Society, Sport* and *Television*). However, conditions were not quite similar, since human judges had to pick a category among eight, and not, like the automatic systems, a category within two subsets of four and five categories. Indeed, we dispatched the categories into two sets, by balancing categories that are easy or difficult for human evaluators. For the second set of categories, we carefully put together three semantically close ones, (*France, International* and *Society*, all three of them being about political and societal contents), to make the task more difficult. Although the second set of categories seems more complicated for human judges, half of the competitors obtained better results in topical categorization of the second set than of the first one.

5 Conclusion

The relevance of human judgment in an evaluation campaign is present from the beginning to the end of a campaign.

In a first step, testing a topic for a campaign among a limited number of human evaluators allows us to check the feasibility of a task. This checking relies both on the results obtained by judges (recall, precision, F-scores) and on their personal impressions after passing the test.

In a second step, the study of both the results obtained by the judges, and their pairwise matching involving such a comparator as the κ coefficient allows us to adjust the task (choice of a marking scale for DEFT'07 and selection of topical categories for DEFT'08).

Finally, the mutual comparison of competitors' results, at the end of the evaluation campaign, allows us to validate the choices we made at its starting point, and even to reposition the task when we shall launch a future campaign based on the same topic.

References

Adda, Gilles, Joseph Mariani, Patrick Paroubek, Martin Rajman, and Josette Lecomte. 1999. L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Langues*, 2(2):119–129, juin.

Burstein, Jill and Magdalena Wolska. 2003. Toward evaluation of writing style: Finding overly repetitive word use in student essays. In *10th Conference of the European Chapter of the Association for Computational Linguistics, EACL'03*, pages 35–42, Budapest, Hungary, april.

Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 2(22):249–254.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, (20):37–46.

Eck, Matthias and Chiori Hori. 2005. Overview of the iwslt 2005 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 5–14, Pittsburg, PA.

Grouin, Cyril, Jean-Baptiste Berthelin, Sarra El Ayari, Thomas Heitz, Martine Hurault-Plantet, Michèle Jardino, Zohra Khalis, and Michel Lastes. 2007. Présentation de DEFT'07 (DÉfi Fouille de Textes). In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, pages 1–8, Grenoble. Association Française d'Intelligence Artificielle.

Hovy, Eduard, Margaret King, and Andrei Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation*.

Hurault-Plantet, Martine, Jean-Baptiste Berthelin, Sarra El Ayari, Cyril Grouin, Patrick Paroubek, and Sylvain Loiseau. 2008. Résultats de l'édition 2008 du DÉfi Fouille de Textes. In *Actes TALN'08*, Avignon. Association pour le Traitement Automatique des Langues.

Paek, Tim. 2001. Empirical Methods for Evaluating Dialog Systems. In *Proceedings of the ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems*, pages 3–10.

Paroubek, Patrick, Jean-Baptiste Berthelin, Sarra El Ayari, Cyril Grouin, Thomas Heitz, Martine Hurault-Plantet, Michèle Jardino, Zohra Khalis, and Michel Lastes. 2007. Résultats de l'édition 2007 du DÉfi Fouille de Textes. In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, pages 9–17, Grenoble. Association Française d'Intelligence Artificielle.