

# Evaluation and NLP<sup>1</sup>

Didier Nakache<sup>1,2</sup>, Elisabeth Metais<sup>1</sup>, and Jean François Timsit<sup>3</sup>

<sup>1</sup> CEDRIC /CNAM: 292 rue Saint Martin - 75003 Paris, France

<sup>2</sup> CRAMIF: 17 / 19 rue de Flandre - 75019 Paris, France

<sup>3</sup> Réanimation médicale CHU Grenoble INSERM U578 - 38043 Grenoble, France

datamining@wanadoo.fr, metais@cnam.fr,  
jff.timsit@outcomerea.org

**Abstract.** F-measure is an indicator which has been commonly used for 25 years to evaluate classification algorithms in textmining, based on precision and recall. For classification and information retrieval, some prefer to use the break even point. Nevertheless, these measures have some inconvenient: they use a binary logic and don't allow to apply a user (judge) assessment. This paper proposes a new approach for evaluation. First, we distinguish classification and categorization from a semantic point of view. Then, we introduce a new measure: the K-measure, which is an overall of F-measure, and allows to apply user requirements. Finally, we propose a methodology for evaluation.

**Keywords:** evaluation, measure, classification, categorization, NLP.

## 1 Introduction

Natural language processing produces many algorithms for classification, clusterisation and information retrieval. The performance of these algorithms is computed from several measures, like precision and recall. To make the reading of performance easier, [Van Rijsbergen 79] created a synthetic measure: the F-measure, which is a combination of these two indicators. Today, needs are diversified, problems are more complex, but we have kept the same indicator for 25 years [Sparck Jones 2001]. Is this use still justified? Without renouncing to existing scales, how to integer new needs? In several domains, like in medicine, some users may consider that a medium result is a bad, or inappropriate, result. So, we had to find an indicator able to answer this problem, without losing qualities of existing measures.

To do this, we introduce our paper by precisising the main concepts: evaluation, classification, categorization, and information retrieval. We will propose a definition for each one (section 2). Section 3 presents the state of the art for evaluation and main indicators. Finally, after having analyzed the F-measure properties (section 4), we will propose a new approach for evaluation, adapted for each case, and allowing to integrate user's requirements (section 5).

---

<sup>1</sup> This work is partially financed by MENRT for the RNTS Rhea project.

## 2 Etymology and Definitions

Terms ‘classification’ and ‘clusterisation’ have different histories and origins. No scientific definition could be found, except in Webster dictionary which gives two meanings for the word classification: ‘taxonomy’ and .... category.

According to the history of these two terms and their current meaning, **we propose to define classification** as being action of arranging a whole set into hierarchical or ordered structure, in existing classes or not, or the result of this action, and **clusterisation** as being action (or its result) of grouping elements with common characteristics. Nevertheless, in a classification, it will be possible to quantify or valorize the difference between proposition and requirement. We can consider an answer as being partially true, and associate a metric to the difference. Finally, **information retrieval** is different from classification and categorization by a great set of enabled answers (potentially infinite), by missing of referential, and often obligation of human evaluation. Classical application could be a web crawler, or AI answers to a request. With so many different tasks, evaluation methods and indicators can be different.

## 3 State of the Art

### 3.1 What Is Evaluation?

Evaluation consists in measuring the difference between an expected result and the final result. No metric is associated, but we use to generate a number between 0 and 1 without unity. Some elements are very subjective and can’t be automated. Tefko Saracevic [Saracevic 70] insists on the main role of judge.

### 3.2 Indicators and Measures: Toward the F Measure

A system can answer to a request according to the following model:

	Pertinent	Not pertinent	Total
Found (or proposed)	a	B	a+b
Not found (or not proposed)	c	D	c+d
	a+c	b+d	a+b+c+d=N

From this contingency table, NLP community computes several distances:

$$\text{precision} = a/(a+b), \text{ recall} = \frac{a}{a+c}, \text{ pertinence} = \frac{a+d}{a+b+c+d}, \text{ error} = \frac{b+c}{a+b+c+d},$$

$$\text{fallout} = \frac{b}{b+d}, \text{ silence} = \frac{c}{a+c}, \text{ specificity} = \frac{d}{b+d}, \text{ noise} = \frac{b}{a+b}, \text{ overlap} = \frac{a}{a+b+c}$$

and generality = a/N

Finally, 4 single measures (a, b, c, d) generate 10 basic indicators, themselves combined to generate other measures. In most of the cases, we only use precision

and recall. From these different measures, several synthetic indicators have been created, but the most famous is the F-Measure from [Van Rijsbergen 79]:

F-Measure =  $((1+\beta^2)*Precision*Recall) / ((\beta^2*Precision)+Recall)$ , with usually  $\beta^2 = 1$   
 It can be noticed that this measure doesn't take pertinence into consideration and is binary: an answer is "good" or "not good".

### 4 Analysis of F-Measure

First, we have demonstrated that the F-measure is a harmonic average of precision and recall. Then, we have observed its properties. When precision has the same value as recall, we get: Precision = Recall = F1-measure. Therefore, the result is comprehensive and we try to maximize it by maximization of both precision and recall (like for 'Break Even Point' approach). Indeed, it would be difficult to evaluate an algorithm which would have a good precision and a bad recall (or reverse).

Let's compute harmonic mean M of precision P and recall R:  $\frac{2}{M} = \frac{1}{P} + \frac{1}{R}$  so

$$\frac{2}{M} = \frac{P + R}{P * R}, \text{ and } \frac{M}{2} = \frac{P * R}{P + R}. \text{ Finally, we get: } M = \frac{2 * (P * R)}{P + R} = F1$$

We notice that F1-measure is a harmonic mean of precision and recall. Nothing can justify this choice from a mathematical point of view. Nevertheless, harmonic mean has an interesting property which is: the result strongly decreases when only one of its components decreases. At the opposite, it grows strongly when the parameters are both high. This property is interesting because it would give a low result for algorithms which would improve precision or recall exclusively in prejudice of the other one.

We can demonstrate that property for the F1-measure: we have  $F1 = \frac{2 * P * R}{P + R}$ , with precision=P and recall=R. Let's have  $S = P + R$  and  $D = P - R$ . Our problem becomes: how to improve F-measure when S increases (so precision AND recall are high) and D is minimized (keeping precision and recall closed). We have:

$$S^2 - D^2 = (S + D)(S - D) = (P + R + P - R) * ((P + R) - (P - R)) = 2P * 2R = 4PR$$

And finally:  $F1 = \frac{2 * P * R}{P + R} = \frac{S^2 - D^2}{2S} = \frac{S}{2} - \frac{D^2}{2S}$ ; that's the reason why F-measure is improved when S increases, and decreases when D increases. If one of the components is low, the resulting mean is low too. The Fn measure has another interesting property: it allows to modify importance of precision or recall.

### 5 Proposition of New Indicator: Toward K-Measure

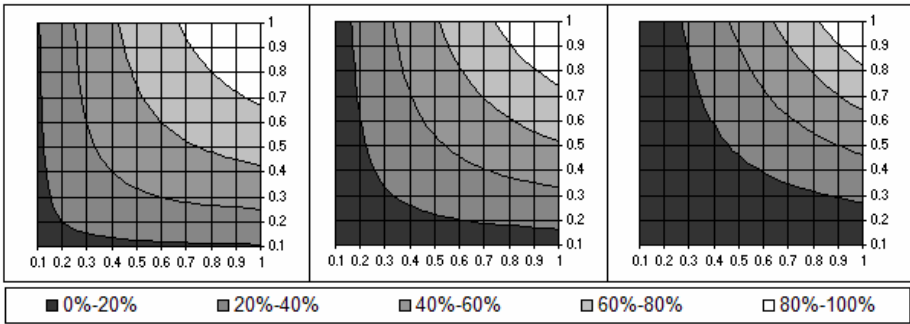
In section 1, we tried to define classification and distinguish it from categorization and information retrieval. Now, we are going to find a new measure, with more possibilities for evaluation.

**Case of Categorization**

After analysis of F-measure, we found a formula which could integrate those needs and introduce **K-Measure**, based on F-measure:

$$\mathbf{K-Measure} = (1+\beta^2) * (\mathbf{Precision} * \mathbf{Recall})^\alpha / ((\beta^2 * \mathbf{Precision}) + \mathbf{Recall})$$

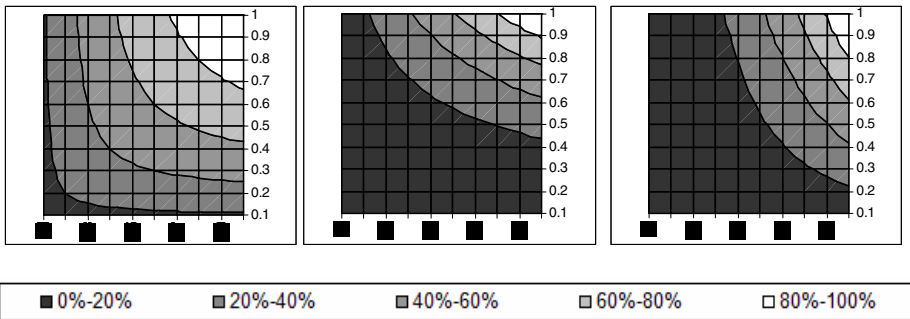
First, we can see that if  $\alpha=1$ , then K-measure is equal to F-measure. If  $\alpha=1$  and  $\beta^2=1$ , we get the usual F1-measure. So, the K-measure is a generalization of the F-measure. This is particularly useful because we can keep the history. Now, let's see properties when  $\beta^2=1$ , and  $\alpha$  parameter is varying with values 1, 1.2, and 1.6:



**Fig. 1.** Evolution of K measure when only  $\alpha$  parameter is varying

We notice that when  $\alpha$  parameter increases, the requirement level increases too. For example, if precision=recall=0.4, F-measure = 0.4, and k-measure = 0.13 with  $\alpha=1.6$  (three times less). This result will be considered as bad, while F-measure considers it as medium. So, we can formalize a requirement level, just increasing  $\alpha$  parameter.

We can observe that favor precision or recall is preserved, by increasing  $\beta$  parameter. And finally, we can use both parameters  $\alpha$  and  $\beta$ :



**Fig. 2.** Values of K-measure for  $(\alpha=1; \beta=1)$ ,  $(\alpha=2; \beta=0.2)$ ,  $(\alpha=2; \beta=5)$

In conclusion, K-measure has very interesting properties for evaluation:

- It is an overall of F measure which keeps its properties,
- It allows to express a judge requirement level,
- It can as well represent a Break Even Point approach when  $\alpha = 0.5$ .

It is a formula of convergence, and an overall of different approaches used nowadays.

**Case of Classifications**

As proposed in section 1, a classification distinguishes from categorization because we can use a distance measure between classes. [Budanitsky 2001] demonstrated that best measure of semantic distance was Jiang and Conrath measure:

$$d = \text{Dist}_{JC}(c1 : c2) = 2 \log(p(\text{lsc}(c1 : c2))) - (\log(p(c1)) + \log(p(c2)))$$

with  $\text{lsc}(c1 : c2)$  = largest common group.

If we call ‘d’ that distance (with  $d=1$  when classes are very far), then precision and recall can be defined like this:

Precision =  $a / B$  et Recall =  $a / c$ ,

$a$  = Count of pertinent and proposed classes (= correct classification),

$B$  = proposed class but not pertinent: we consider the distance ‘d’ with the nearest correct class. Then compute  $(1-d)$  to have  $B$  near 1 when distance is weak,

$c$  = Count of not proposed and pertinent classes.

**It is then possible to use K-measure.**

**Case of Information Retrieval (IR)**

Information retrieval is different from classification and categorization because of the large number of possible answers. Example of classical application would be a web crawler.

To find a good indicator, we started from the score used by [Voohrees 2003]

$$\frac{1}{Q} \sum_{i=1}^Q \frac{n}{i}$$

where  $n$  represents the number of good answers in range  $i$ ,  $Q$  is the number

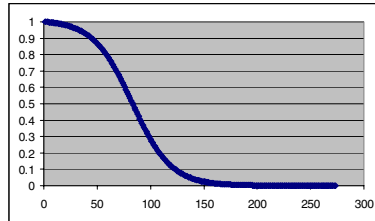
of questions. To represent a requirement level (for example: “I want that good answers are in the first 30, because it is the length of a web page”, we need to modulate the initial Boolean and linear approach by integrating a sigmoid function. After empirical researches, we could find a coefficient  $W_i$  which solves our problem:

$$w_i = \frac{1 + e^{(-k)}}{1 + e^{(-k \times \left(\frac{N-i+1}{N}\right) - l)}}$$

With  $k$  and  $l$ , two parameters (default values are  $k=15$ , and  $l=0.7$ ),  $N$  represents the number of answer, and  $i$  the range of the answer.

Let's see the properties of that equation when k and l are varying (in our example, we have N=273)

K=15, l=0.7



We can see that if the required answer doesn't appear in top 50, the score is strongly down, and quite null if higher than 150. The l parameter moves inflexion point (right and left), and k changes the slope.

The two parameters allow generating any requirement level. This score favor fact of giving good answers in first. To compute final indicator, we just multiply weight by pertinence. For automatic computing, we can use a Boolean approach: 1 for a good answer, otherwise 0. But for human evaluation, each judge can give a percentage. Global evaluation indicator becomes:

$$D \text{ Measure} = \frac{\sum_{i=1}^N \text{Pertinence}_i * w_i}{\sum_{i=1}^N w_i} = \frac{\sum_{i=1}^N \text{Pertinence}_i * \frac{1 + e^{(l-k)}}{1 + e^{(-k * (\frac{N-i+1}{N}) - l)}}}{\sum_{i=1}^N \frac{1 + e^{(l-k)}}{1 + e^{(-k * (\frac{N-i+1}{N}) - l)}}}$$

For automatic computing, we can use a Boolean approach: 1 for a good answer, otherwise 0. But for human evaluation, each judge can give a percentage.

This evaluation indicator is interesting because it allows:

- To represent requirement level,
- To be able to evaluate otherwise than with 0 and 1,
- To control requirements.

## 6 Conclusion

In this paper, we first defined classification and categorization. In the first case, we were able to measure the distance between classes, but not in second case as it is binary. The F-measure, which was created 25 years ago, has been established as a standard for evaluation. Since then, the needs evolves but not evaluation. Analysis of F-measure helps us to create a new measure: K-measure, an overall of F-measure able to integrate requirements. We demonstrate how to use k-measure for classification as well as to integrate the distance between the results. Finally, we propose a new measure for information retrieval which enhances finding good answers first and allow the expression of needs.

K-measure provides the following advantages: a meta measure of convergence between Van Rijsbergen's F-measure and Joachims's break even point. It has mathematical properties which allow to create a synthetic indicator from any other measure. Finally, it allows to integrate the judge approach of Saracevic and to formalize the required levels. Therefore, we consider say that it is a measure which converges the three approaches without modifying any of their properties.

In our future works, we will experiment these measures, particularly their impacts on classical measures.

## Acknowledgements

We would like to thank:

- Professor Jacky AKOKA and the Cedric laboratory - CNAM (Conservatoire National des Arts et Métiers) and
- Pierre KEBAILI and Jacques TONNER (CRAMIF : Caisse Régionale d'Assurance Maladie d'Ile de France)

for supporting us.

## Bibliography

- [Budanitsky 2001] A. Budanitsky and G. Hirst : "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures" Department of Computer Science Univ. of Toronto
- [Saracevic 70] Tefko Saracevic, "Introduction to Information Science", 111-151. New York: R.R. Bowker, 1970. Chap. 3 : The concept of "relevance" in information science: A historical review.
- [Sparck Jones. 2001] K. Sparck Jones. "Automatic language and information processing: Rethinking evaluation". *Natural Language Engineering*, 7(1):29-46. 2001
- [Van Rijsbergen 79] K. Van Rijsbergen, "Information Retrieval", (2nd Ed.) Butterworths, London. [www.dcs.gla.ac.uk/Keith/Preface.html](http://www.dcs.gla.ac.uk/Keith/Preface.html)
- [Voorhees 2003] E. M. Voorhees : "Evaluating the Evaluation: Edmonton", May-June 2003. Main Papers , pp. 181-188. Proceedings of HLT-NAACL 2003