Manual Corpus Annotation: Giving Meaning to the Evaluation Metrics

Yann MATHET[1,2,3] *Antoine WIDLÖCHER*[1,2,3] *Karën FORT*[4,5]
Claire FRANÇOIS[6] *Olivier GALIBERT*[7] *Cyril GROUIN*[8,9]
Juliette KAHN[7] *Sophie ROSSET*[8] *Pierre ZWEIGENBAUM*[8]

(1) Université de Caen Basse-Normandie, UMR 6072 GREYC, Caen, France
(2) ENSICAEN, UMR 6072 GREYC, Caen, France (3) CNRS, UMR 6072 GREYC, Caen, France
(4) LIPN, Villetaneuse, France (5) LORIA, Vandœuvre, France (6) INIST–CNRS, Vandœuvre, France
(7) LNE, Trappes, France (8) LIMSI–CNRS, Orsay, France
(9) INSERM, UMR_S 872, Eq 20 & UPMC, Paris, France

`{yann.mathet, antoine.widlocher}@unicaen.fr,`
`karen.fort@loria.fr, claire.francois@inist.fr,`
`{olivier.galibert, juliette.kahn}@lne.fr,`
`{cyril.grouin, sophie.rosset, pierre.zweigenbaum}@limsi.fr`

Abstract

Computing inter-annotator agreement measures on a manually annotated corpus is necessary to evaluate the reliability of its annotation. However, the interpretation of the obtained results is recognized as highly arbitrary. We describe in this article a method and a tool that we developed which "shuffles" a reference annotation according to different error paradigms, thereby creating artificial annotations with controlled errors. Agreement measures are computed on these corpora, and the obtained results are used to model the behavior of these measures and understand their actual meaning.

Keywords: inter-annotator agreement, manual corpus annotation, evaluation.

Proceedings of COLING 2012: Posters, pages 809–818,
COLING 2012, Mumbai, December 2012.

1 Introduction

The quality of manual annotations has a direct impact on the applications using them. For example, it was demonstrated that machine learning tools learn to make the same mistakes as the human annotators, if these mistakes follow a certain regular pattern and do not correspond to simple annotation noise (Reidsma and Carletta, 2008; Schluter, 2011). Furthermore, errors in a manually annotated reference corpus (a "gold-standard") can obviously bias an evaluation performed using this corpus as a reference. Finally, a bad quality annotation would lead to misleading clues in a linguistic analysis used to create rule-based systems.

However, it is not possible to directly evaluate the validity of manual annotations. Instead, inter-annotator agreement measures are used: at least two annotators are asked to annotate the same sample of text in parallel, their annotations are compared and a coefficient is computed. The latter can be of many types and the well-known Kappa-family is described in details in (Artstein and Poesio, 2008). However, as pointed out by the authors of this article, the obtained results are difficult to interpret. Kappa coefficients, for example, are difficult to compare, even within the same annotation task, as they imply a definition of the markables that can vary from one campaign to the other (Grouin et al., 2011). More generally, we lack clues to know if a Kappa of 0.75 is a "good" result, or if a Kappa of 0.8 is twice as good as one of 0.4 or if a result of 0.6 obtained using one coefficient is better than 0.5 with another one, and for which annotation task.

We first briefly present the state of the art (Section 2), then detail the principles of our method to benchmark measures (Section 3) and show on some examples how different coefficients can be compared (Section 4). We finally discuss current limitations and point out future developments.

2 State of the art

A quite detailed analysis of the most commonly used inter-annotator agreement coefficients is provided by Artstein and Poesio (2008). They present the pros and cons of these methods, from the statistical and mathematical points of view, with some hints about specific issues raised in some annotation campaigns, like the prevalence of one category. A section of their article is dedicated to various attempts at providing an interpretation scale for the Kappa family coefficients and how they failed to converge. Works such as (Gwet, 2012) are also to be mentioned. They present various inter-rater reliability coefficients and insist on benchmarking issues related to their interpretation.

Many authors, among whom (Grouin et al., 2011; Fort et al., 2012), tried to obtain a more precise assessment of the quality of the annotation in their campaigns by computing different coefficients and analyzing the obtained results. However, their analyses lack robustness, as they only apply to similar campaigns. Other studies concerning the evaluation of the quality of manual annotation identified some factors that influence inter- and intra-annotator agreements, thereby giving clues on their behavior. Gut and Bayerl (2004) thus demonstrated that the inter-annotator agreement and the complexity of the annotation task are correlated: the larger the number of categories, the lower the inter-annotator agreement. However, categories prone to confusion are in limited number. The meta-analysis presented by Bayerl and Paul (2011) extends this research on the factors influencing agreement results, identifying 8 such factors and proposing useful recommendations to improve manual annotation reliability. However, neither of these studies provides a clear picture of the behavior of the agreement coefficients

[0] This work has been partially financed by OSEO, the French State Agency for Innovation, under the Quaero program.

or of their meanings. The experiments detailed in (Reidsma and Carletta, 2008) constitute an interesting step in this direction, focusing on the effect of annotation errors on machine learning systems and showing the impact of the form of disagreements on the obtained quality (random noise disagreement being tolerable, but not patterns in disagreements). This work puts Kappa-like coefficients results into perspective but presents a tool-oriented view, limited to these coefficients. In summary, the domain lacks a tool providing a clear and generic picture of the agreement coefficients behavior, allowing to better qualify the obtained agreement results.

3 Generating benchmarking corpora: the Corpus Shuffling Tool

The method presented in this section is currently restricted to annotation campaigns consisting in delimiting a span of text and characterizing it. It will be extended in the future to relations and more complex structures.

3.1 Objectives and principles

Manual annotation, as already mentioned, is subject to human errors. Except for very simple annotation tasks, these errors may involve several paradigms. Indeed, each manually annotated element may diverge from what it should be (which is called the reference, see below), in one or multiple ways, including: (i) the location is not correct (the frontiers of an element do not exactly match those of the reference); (ii) the characterization is not correct (wrong category, or wrong feature value); (iii) the annotation does not belong to the reference (false positive); or (iv), on the contrary, a reference element is missing (false negative). All of these error paradigms tend to damage the annotations, so each of them should be taken into account by agreement measures. We propose here to apply each measure to a set of corpora, each of which embeds errors from one or more paradigms, and with a certain magnitude (the higher the magnitude, the higher the number of errors). This experiment should allow us to observe how the measures behave w.r.t. the different paradigms, and with a full range of magnitudes. The idea of creating artificial damaged corpora is inspired by Pevzner and Hearst (2002), then Bestgen (2009) in thematic segmentation, but our goal (giving meaning to measures) and our method (e.g. applying progressive magnitudes) are very different.

3.2 Protocol

Reference. A reference annotation set (called *reference*) is provided to the system: a true Gold Standard or an automatically generated set based on a statistical model. It is assumed to correspond exactly to what annotations should be, with respect to the annotation guidelines.

Shuffling. A shuffling process is an algorithm that automatically generates a multi-annotated corpus given three parameters: a reference annotation, a number n of annotators to simulate, and a coefficient $0 \leq m \leq 1$ called magnitude (in reference to earthquake measures). Each time it is run, it creates a set of n parallel annotations (simulating n different annotators) on the corpus, but with a quality damaged according to magnitude m.

Process. The system iteratively runs a given shuffling process on the full range of possible magnitudes (from 0 to 1) with a parametrizable step (default is 0.05).

Agreement measure graph. For each of these annotation sets, i.e., for each magnitude, we submit each agreement measure we want to evaluate, and record its score. At the end of the process, we obtain a graph showing how a given measure reacts to a progressive shuffling, where the x-axis represents the magnitude from 0 to 1, and the y-axis represents the agreement.

3.3 Overview of the implemented shuffling processes

All processes described here come from real observations of various corpora: *false positives* and *false negatives* are usual in many campaigns, *fragmentation* and *shift* are observed in thematic segmentation, *category mistake* is so usual that most agreement measures (e.g. Kappa) address it, and *combination* of them appear for instance in discourse annotation.

3.3.1 False negative

A *false negative* is the fact for an annotator not to annotate an element belonging to the reference. It is simulated as follows: (i) magnitude $m = 0$: the annotator did not miss any annotation; (ii) magnitude $m = 1$: the annotator missed all the annotations and therefore did not produce any; and (iii) $0 < m < 1$: each element to be annotated has a probability m of being missed.

3.3.2 False positive

Reversely, a *false positive* is the fact for an annotator to annotate an element not belonging to the reference. We made the following decisions: (i) the maximum shuffling corresponds to adding x times the number of elements of the reference, x default value being 1; (ii) for $0 \leq m \leq 1$, $m \cdot x$ elements are added; (iii) the way annotations are added is done with respect to the characteristics of the reference (statistical distribution of categories, etc.)

3.3.3 Fragmentation

Sometimes, annotators have the choice between using several contiguous elements (of the same category), or just one (covering the same text spans). Fragmentation simulates this by splitting up reference elements. The protocol is as follows, n being the number of reference elements: (i) the number of fragmentations to apply is $n_{frag} = n \cdot x \cdot m$, where x is settable (its default value is 1); (ii) for each fragmentation to apply, one element is chosen at random, and is split (not necessarily in its center); (iii) the fragment of a split may be re-split next time, so that we finally get several levels of fragmentation.

3.3.4 Shift

In some annotation campaigns, annotators manage to properly identify the phenomenon to annotate, but have trouble locating it perfectly. We try to reproduce this error paradigm with the *shift* shuffling, that moves the frontiers of the reference annotations. It is defined as follows: (i) for magnitude m, we take into account the average length of the elements (w.r.t. their category), using for instance a statistical model as described in section 3.2, called *maxlength*, to compute the possible shifting latitude of each frontier, called *maxlat*, as follows: $maxlat = maxlength \cdot m \cdot x$, where m is the magnitude and x is a parameter whose default value is 2; (ii) a number is chosen at random for each frontier, in the range from *-maxlat* to *+maxlat*, and the frontier is shifted by this algebraic value. This shuffling process is conceptually more difficult to design than the previous ones because the shuffling space being finite (the text length), it is difficult to know to what extent it is actually possible to shuffle the annotations.

3.3.5 Category mistake

A frequent annotation mistake is to assign a wrong category to an annotated element. Two important phenomena are to be mentioned, that are quite frequent and lead to some important differences among current measurement methods: **Prevalence** is the fact that some categories are more frequent than others. Some measures take this phenomenon into account in their definition of so-called (and controversial) chance correction in order not to overrate the observed agreement. **Overlapping** is the fact for two categories to cover, even slightly, a same phenomenon: in such cases, annotators happen to choose a wrong but not so different category, and some measures consider them as less important mistakes. The question now is to define how best to simulate, the more gradually possible, a progressive category assignation mistake. To define such a simulation, we rely, for a given magnitude, on a matrix that indicates, for each category of the reference annotation, what is the probabilistic distribution of the chosen categories for 100 annotations. We have made the following choices: (i) for $m = 0$, we use the **perfect matrix** A (as given in table 1); (ii) for $m = 1$, we use the worst matrix B or C depending on the choice of simulating prevalence (B) or not (C). The **prevalence matrix** B simulates a (semi) random behavior with respect of the prevalence observed in the reference, while the **noPrevalence matrix** C reflects a full random choice; (iii) Besides, overlapping is simulated by the **overlapping matrix** D, which describes the way an annotator, for a given category in the reference, makes mistakes more often in favor of friendly categories than in favor of others; (iv) then, for each $0 < m < 1$, we built a matrix by weighted averaging of perfect matrix (100% weighted at $m = 0$) and worst matrix (100% weighted at $m = 1$), as shown in Figure 1 (right). When the overlapping option is chosen, the overlapping matrix is integrated in the averaging, with a weight distribution being zero at $m = 0$ and $m = 1$, and a maximum in the intermediate magnitudes, as shown in Figure 1 (left). Indeed, we consider such errors as neither belonging to perfect annotation, nor to worst annotation.

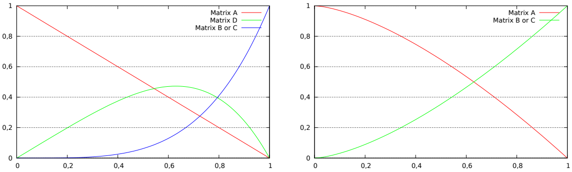

Figure 1: Weight distributions for averaging with overlapping (left) or without it (right)

	A:Perfect				B:Prevalence				C:NoPrevalence				D:Overlapping			
	Noun	Verb	Adj	Prep	Noun	Verb	Adj	Prep	Noun	Verb	Adj	Prep	Noun	Verb	Adj	Prep
Noun	100	0	0	0	27	9	18	45	25	25	25	25	0	80	15	5
Verb	0	100	0	0	27	9	18	45	25	25	25	25	80	0	0	20
Adj	0	0	100	0	27	9	18	45	25	25	25	25	15	10	0	75
Prep	0	0	0	100	27	9	18	45	25	25	25	25	5	20	75	0

Table 1: The four confusion matrices used for interpolation

Combining the two options, 4 different experiments can be built: with or without overlapping and with or without prevalence.

3.3.6 Combination

Each previously defined shuffling process involves a particular paradigm. However, in the real world, human annotation errors in a given campaign may involve several paradigms at the same time, sometimes on the same annotated element (e.g. a slight shift and a category mistake). To address this situation, we provide a shuffling process that combines as many shuffling processes as needed, defined as follows: (i) n sub-processes are chosen in a given order; (ii) for a magnitude m, the main process shuffles the reference annotation, successively applying each sub-process (in the given order) with magnitude m/n (hence, this multi-shuffling is not n times faster as classic ones).

4 Using shuffled corpora to compare measures: a brief overview

To demonstrate the consistency of the method we briefly show in this section how it can be used with two types of annotation paradigms.

4.1 Segmentation: Comparison of WindowDiff, G-Hamming and GM

Segmentation consists in determining frontiers between contiguous textual segments. We compare here two metrics already compared by (Bestgen, 2009): WindowDiff (WD) described in (Pevzner and Hearst, 2002) and Generalized Hamming Distance (GH) described in (Bookstein et al., 2002), as well as a new versatile measure, the Glozz Measure (GM), described in (Mathet and Widlöcher, 2011), which can be adapted to several paradigms, including segmentation. WD and GH cannot exactly be considered as agreement measures, as they are distances between a reference and a human annotation. These distances equal 0 when annotations are the same, and 1 in the worst case. We have adapted the results as follows: *agreement* $= 1 -$ *distance*. Moreover, since these metrics consider two annotators only, we have averaged the one-to-one results when working with 3 or more annotators.

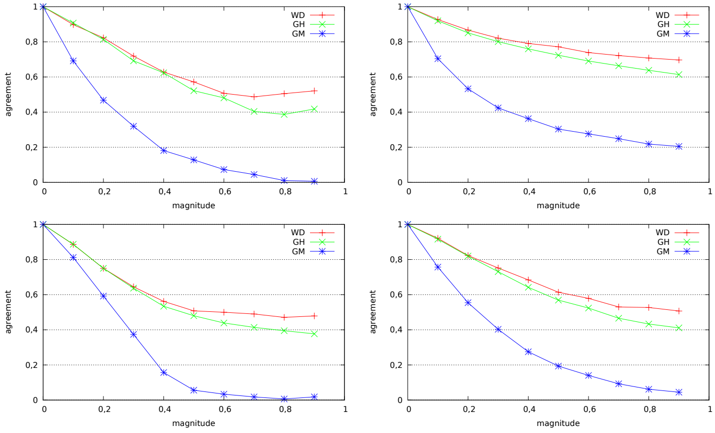

Figure 2: False negatives (upper left), false positives (upper right), shift (lower left) and combination (lower right)

Figure 2 shows the behavior of these three measures for three paradigms and their combination: for **false negatives** WD and GH are quite close, with an almost linear response until magnitude 0.6. Their drawback is that their responses are limited by an asymptote, while GM shows a full range of agreements, but is not linear; again, for **false positives**, WD and GH are very similar, and their responses, if not asymptotic, show a lower limit at a quite high value (resp. 0.7 and 0.6). GM behaves in the same way as for false negatives, but with an asymptote at *agreement* $= 0.2$ much lower than WD and GH; once again, for **shifts**, WD and GH show an asymptote at about *agreement* $= 0.4$, when GM shows values from 1 to 0. Not surprisingly, when using **combination**, the overall responses look like an average of the other paradigms. This very first and brief comparison reveals that WD and GH are quite close, but GH scores are a little more severe, and with a wider range. For these reasons, according to this experiment, GH seems slightly better. GM is quite different, with almost a full range of agreements, probably because it takes chance into account.

4.2 Categorization: Comparison of Kappa, W-Kappa and GM

We focus here on categorization only, assuming a situation where the elements to annotate are pre-located. Four sets of corpora were created, with respect to the two available options described in section 3.3.5. The measures we compare here are Cohen's Kappa (Cohen, 1960), the weighted Kappa (Cohen, 1968), with two different weight matrices (W-Kappa 1 being much more forgiving than W-Kappa 2); and GM (Mathet and Widlöcher, 2011), with two different options, GM1 which has overlapping capabilities, and GM2 which has not. We also add a very simple percentage agreement value as a baseline (called BM, for Baseline Measure) for all the other measures. The results are shown in Figure 3. First of all, when neither overlapping nor prevalence is involved, all the measures behave almost in the same way (even though BM slightly overrates the agreement as magnitude increases, because it does not take chance

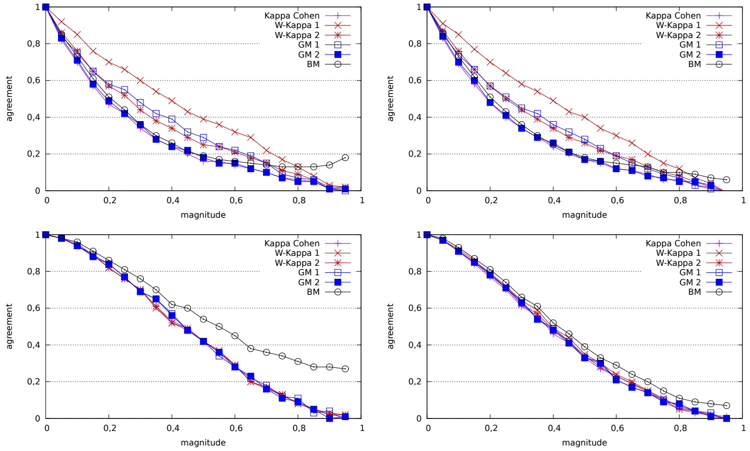

Figure 3: Results of different measures with prevalence (bottom-left), overlapping (top-right), overlapping+prevalence (top-left), and none (bottom-right)

into account). When a **prevalence** phenomenon occurs, all the measures (except BM) still perform equivalently, but BM increasingly overrates the agreement by up to about 0.25. Taking chance into account has more impact here. The **overlapping** phenomenon clearly opposes W-Kappa and GM to others. Whatever the prevalence option (top-left and top-right figures), the differences are important in the 0.1 to 0.6 magnitude range (where the overlapping matrix has more influence), with a difference of up to 0.15 for GM, and up to about 0.25 for W-Kappa-1. The latter reacts with more strength because we set it with a very forgiving weight matrix, while W-Kappa-2 is set with a less forgiving one, and is very close to GM whose weight matrix is data-driven. Besides, it is interesting to note that when applying these two measures to non-overlapping data (bottom figures), they behave almost exactly the same way as their basic versions not taking overlapping into account.

5 Limitations and future work

Enhancing annotators' simulation. We shall try in the future to get closer to real annotation constraints. For instance, shifting is currently free, whereas in some campaigns annotating overlapping entities is prohibited. We will also address the question of differences of behavior between annotators.

Using real Gold Standard corpora. It is also possible to use a real Gold Standard corpus as a reference for the system, and then to shuffle it. We started this work with the TCOF-POS-tagged corpus (Benzitoun et al., 2012), for which annotators reached a 0.96 Kappa agreement, which corresponds to a magnitude of 0.1 in the Shuffling Tool, i.e., to a matrix averaged between the perfect one at 95% and the worst one at 5%.

Playing with more parameters. For each experiment this tool makes possible, it will be possible to generate sub-experiments, each of which taking into account a given parameter, including: (i) the number of annotators, (ii) the number of categories, (iii) the number of annotated elements, as already studied with statistical considerations by (Gwet, 2012).

Relations and more complex structures. Finally, we shall extend the current work, focused on entities as textual segments, to relations and sets of entities, in order to address other annotation types such as co-reference chains and discourse relations.

Conclusion

According to the results on various types of paradigms, and with quite different agreement measures, the proposed method and corpora happen to be consistent: as expected, it is confirmed that the different measures provide decreasing scores from 1 to 0. Some important differences as well as some similarities, appear between the studied methods. This seems promising for further comparisons, in particular for measures with multi error paradigms capabilities, e.g. Krippendorff's α_U (Krippendorff, 1995) and GM. To sum up, this tool will help to (i) objectively compare the behavior of different agreement measures, (ii) obtain a new and enhanced interpretation of their results: a given result of a given method corresponds to a certain magnitude, of which we have a clear and formal definition, (iii) set and enhance existing or future measures (checking improvements and regressions). The shuffling tool used in this work to generate the damaged corpora is written in Java and is freely available[1] under the GPL license and all the corpora we generated and used for this paper are also freely available.

[1] http://www.glozz.org/shufflingtool

References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Bayerl, P. S. and Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Benzitoun, C., Fort, K., and Sagot, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Proceedings of the Traitement Automatique des Langues Naturelles (TALN)*, pages 99–112, Grenoble, France.

Bestgen, Y. (2009). Quels indices pour mesurer l'efficacité en segmentation thématique? In *Actes de TALN'09*, page p. 10, Senlis (France).

Bookstein, A., Kulyukin, V. A., and Raita, T. (2002). Generalized Hamming distance. *Information Retrieval*, (5):353–375.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Fort, K., François, C., Galibert, O., and Ghribi, M. (2012). Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. 7 pages.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA. (poster).

Gut, U. and Bayerl, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Proceedings of the Speech Prosody*, pages 565–568, Nara, Japan.

Gwet, K. L. (2012). *Handbook of Inter-rater Reliability*. Advanced Analytics, LLC, third edition.

Krippendorff, K. (1995). On the reliability of unitizing contiguous data. *Sociological Methodology*, (25):47–76.

Mathet, Y. and Widlöcher, A. (2011). Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. In *Proceedings of the Traitement Automatique des Langues Naturelles 2011 (TALN 2011)*, Montpellier, France.

Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Reidsma, D. and Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Schluter, N. (2011). *Treebank-Based Deep Grammar Acquisition for French Probabilistic Parsing Resources*. PhD thesis, Dublin City University - Faculty of Engineering and Computing, School of Computing.