Segmentation evaluation metrics, a comparison grounded on prosodic and discourse units

Klim Peshkov, Laurent Prévot
Aix Marseille Université & CNRS, Laboratoire Parole et Langage, Aix-En-Provence, France
klim.peshkov@lpl-aix.fr, laurent.prevot@lpl-aix.fr

Abstract

Knowledge on evaluation metrics and best practices of using them have improved fast in the recent years Fort et al. (2012). However, the advances concern mostly evaluation of classification related tasks. Segmentation tasks have received less attention. Nevertheless, there are crucial in a large number of linguistic studies. A range of metrics is available (F-score on boundaries, F-score on units, WindowDiff ((WD), Boundary Similarity (BS) but it is still relatively difficult to interpret these metrics on various linguistic segmentation tasks, such as prosodic and discourse segmentation. In this paper, we consider real segmented datasets (introduced in Peshkov et al. (2012)) as references which we deteriorate in different ways (random addition of boundaries, random removal boundaries, near-miss errors introduction). This provide us with various measures on controlled datasets and with an interesting benchmark for various linguistic segmentation tasks.

Keywords: evaluation; segmentation; discourse; prosody

1. Introduction

Knowledge on evaluation metrics and best practices of using them have improved fast in the recent years (Fort et al., 2012). However, the advances concern mostly evaluation of classification related tasks. Segmentation tasks have received less attention. Nevertheless, there are crucial in a large number of linguistic studies. A range of metrics is available (F-score on boundaries, F-score on units, WindowDiff ((WD), Boundary Similarity (BS) but it is still relatively difficult to interpret these metrics on various linguistic segmentation tasks, such as prosodic and discourse segmentation. In this paper, we consider real segmented datasets (introduced in (Peshkov et al., 2012)) as references which we deteriorate in different ways (random addition of boundaries, random removal boundaries, near-miss errors introduction). This provide us with various measures on controlled datasets and with an interesting benchmark for various linguistic segmentation tasks.

The analyses presented in (Mathet et al., 2012) concern segmentation and categorization with a longer discussion on categorization. They also consider more perturbations in the datasets than we do. Finally, they consider evaluation of multiple segmentations while we worked only with a reference and one damaged segmentation. On the other hand, they do not pay much attention to the nature and structure of the data. Our approach is closer to our needs, because, as it will be shown below, the measures behave differently on different data. Therefore, we provide a more precise insight on the these metrics for segmentation of spoken data.

2. Survey of the metrics

2.1. Precision / Recall metrics

Precision and recall are conventional evaluation metrics from information retrieval. When applied to segmentation task, separate measures for left boundaries, right boundaries and the entire units can be used. This method was used, for example, for the shared task of CoNLL-2001 (Conference on Computational Natural Language Learning) (Tjong et al., 2001).

2.2. WindowDiff

When used for segmentation evaluation, information retrieval metrics have a serious drawback. They do not take in consideration the distance between the borders of the segmentations being compared. Near-miss errors are penalized as heavily as insertion or deletion of borders and using a threshold value for accommodating these cases can result in a bias. WindowDiff metrics was introduced to address this problem (Pevzner and Hearst, 2002). The algorithm operates as follows. It consists in moving a fixed-length window along the two segmentations, one unit at a time. For each position, the algorithm compares the numbers of borders in both segmentations. If the number of borders is not equal, the difference of the numbers is added to the evaluated algorithm's penalty. The sum of penalties is then divided by the number of measures, yielding a score between 0 and 1. The score 0 means that the segmentations are identical.

Initially, WindowDiff was created for text segmentation tasks. When applying it to the evaluation of units in time-aligned transcripts, we had to adapt it by introducing a time-based (instead of unit-based) step for moving the window. Results shown below were obtained with a step of 50 milliseconds.

2.3. Boundary Similarity

As explained in (Fournier and Inkpen, 2012; Fournier, 2013), Window-based methods also suffer from a variety of problems. We retain the following from their lists of issues: 'unequal penalization of error types', 'an arbitrarily defined window size parameter (whose choice greatly affects outcomes)', the 'lack of clear intuition'. (Fournier and Inkpen, 2012) proposes a new method for comparing two segmentations that answer these issues. They add that a "symmetric" measure that do not use the notion of a reference but more similarly to intercoder agreement, simply evaluate the distance between two segmentation. The key idea consists in thinking about the size of the units and then compute an edit distance based on the sequences of the units size.

3. Cohen's κ-score

Finally, since it is a well-known intercoder agreement metric we also looked at Cohen's kappa (Cohen, 1960). This measure is of a different nature and therefore is not strictly comparable but we argue that it is still useful in this context for two reasons: (i) it helps us interpreting and understanding the results of the various metrics; (ii) it also helps interpreting the κ-score in our intercoder agreements evaluations.

4. Datasets

In this work, we systematize an evaluation work initiated in (Peshkov et al., 2013) on discourse and prosodic units (respectively DUs and PUs). In this previous work, we evaluated existing concurrent annotations with ConLL and WinDowDiff only and to provide some intuition of the metrics we damaged a reference annotation. However, the evaluation part was not systematic enough. Here, we start from the same datasets and damage them systematically in different ways. It is important to start from real datasets since the scores of the metrics are rather sensitive to the exact nature of the data, the ratio (size of base units)/(size of units segmented). The units' length distribution can affect the overall value and dynamics of the metrics as we will see in section 5..

4.1. The reference dataset

Both reference datasets were produced using Praat. Tokens aligned with signal was the base unit for determining the segmentation. The overall features of these datasets are provided in Table 1 while more precise information on the length distribution is presented in Figures 1 and 2.

	discourse	prosody
total time (minutes)	59.5	19.7
n of segments in reference	1582	1777
n of segments in base	7583	5040
segment avg dur (s)	2.26	0.67
segment avg length (in base units)	4.79	2.84

Table 1: Overall figures for the datasets

In table 1, *reference* refers to the segmentation we are interested in. When segmenting written texts the units are generally tokens, but for spoken data other options are also reasonable (a fixed time interval, syllables, phones, etc.). In this work, tokens were used. Therefore, for both datasets *base* refers to the tokens.

The distribution of DU lengths (Figure 1) is peculiar. One-token units are dominant while the rest of the distribution is decreasing slowly with length (being almost flat until a length of 10). The reasons are: (i) a high number of backchannels and other feedback items in the DU dataset; (ii) the fact that pauses are also units of one token.[1] The PU length distribution (Figure 2) is more standard. PUs are generally shorter and their frequency decays with duration. This difference in the distribution has an impact on the evaluation metrics as we will see below.

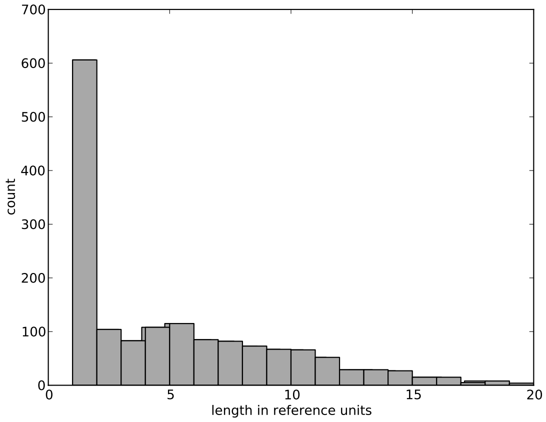

Figure 1: Distribution of discourse units' lengths (in tokens)

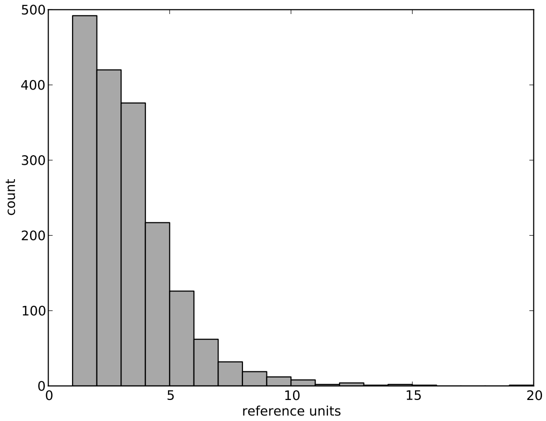

Figure 2: Distribution of prosodic units' lengths

4.2. Damaging the reference

Adding boundaries For each value of n from 1 to 49 with step 0.5, $n\%$ of randomly selected intervals are split into two to simulate false positives error. This way 96 variants of the original segmentation with gradually increasing amount of added boundaries are produced. Possible times for insertion are defined in the reference segmentation, R.

Removing boundaries For each value of n a variant with $n\%$ of removed boundaries is generated, simulating false negatives error. The removal is achieved by merging randomly selected intervals.

Moving boundaries For each value of n a variant with $n\%$ of shifted boundaries is generated. In this case, total number of boundaries does not change. This type of perturbation is introduced to simulate near-miss errors.

Depending on the data, several degrees of shifting are possible, with different shifting distance or amplitude. Amplitude of the shift, a, defines how far will a randomly selected

[1] Pauses are not technically DUs but the data being only composed of pauses and DUs, they must be integrated in the dataset to evaluate the segmentation.

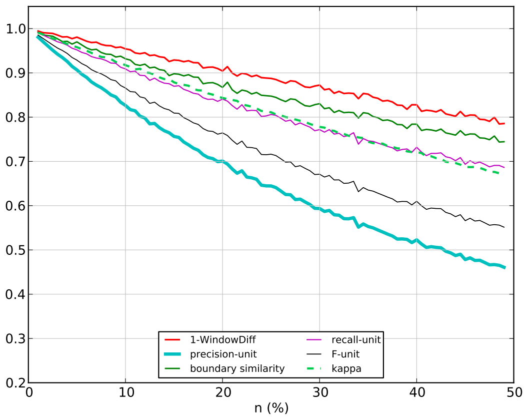

Figure 3: Adding boundaries to discourse dataset

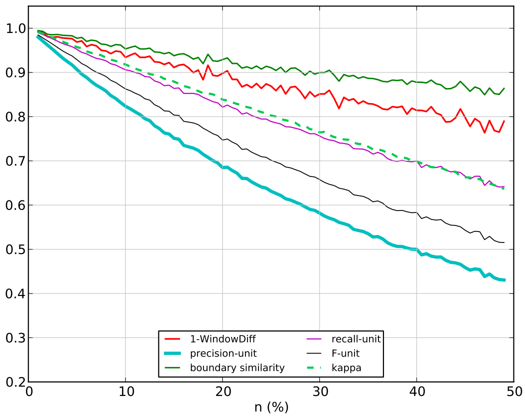

Figure 4: Adding boundaries to prosody dataset

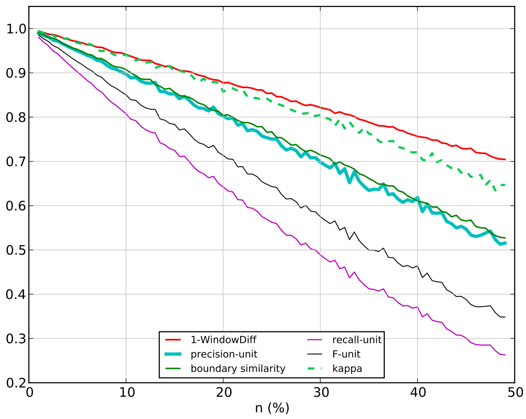

Figure 5: Removing boundaries from discourse dataset

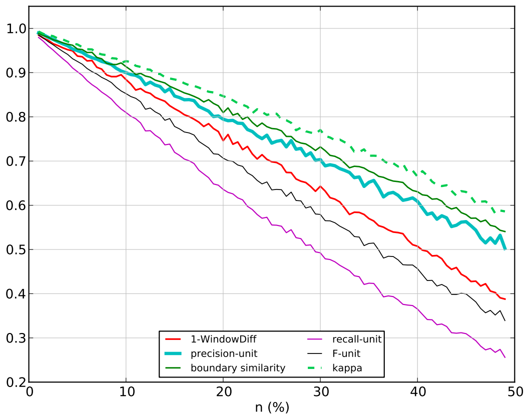

Figure 6: Removing boundaries from prosody dataset

boundary be moved in terms of units of the reference segmentation R.

The maximum value of a is equal to half the average unit length in reference segmentation. For example, for DU dataset, the average length is 4.79 reference units, which means $a_{max} = 2$. Consequently, for this segmentation, two kinds of shifting are used, with $a = 1$ and with $a = 2$.

5. Evaluation results

5.1. Adding boundaries

First of all, we should remind that actual score in our graphics does not mean that a given measure is more strict than another one. The only information the graphics provide are: (i) how to compare the scores; (ii) how the scores evolve according to the type of perturbation and (iii) how the scores evolve with regard to different structures of the datasets.

Overall, the figures 3 and 4 show that the measures are more tolerant to false positives in the case of discourse units. This is only due to the average length of units. As expected, precision decreases quickly while the decrease of recall is slower.[2] Interestingly, WindowDiff and Boundary Edit Distance are inverted between PU and DU datasets.

5.2. Removing boundaries

When removing boundaries, Figures 5 and 6 show a stronger slope than for the boundary addition and the difference between DU and PU is maintained. Again, WD and BS are inverted between PU and DU datasets.

5.3. Perturbating boundaries

Concerning the shifting of boundaries, see Figures 7 and 8 for near-misses and Figure 9 for bigger shifts. As in the previous cases, WD and BS are inverted.

Comparing near-miss and other errors on the DU, we note that structure of the data has more impact on WD and BS than the amplitude of the errors introduced. However, for given datasets, WD and BS are efficient in capturing the differences between near-misses and other errors, BS making this difference more salient.

5.4. Discussion

κ-score is less sensitive to boundary removal than to additions, although we could expect the opposite. There is a prevalence of no-boundary decisions in segmentation tasks, so removing instances from the dominant category rather

[2] There is still a decrease because a perfect match of both units' boundaries is evaluated.

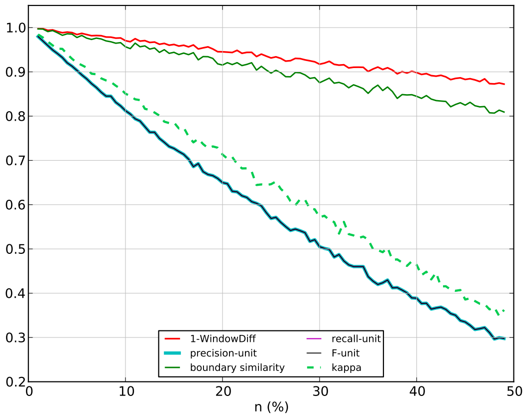

Figure 7: Introducing near misses in discourse dataset

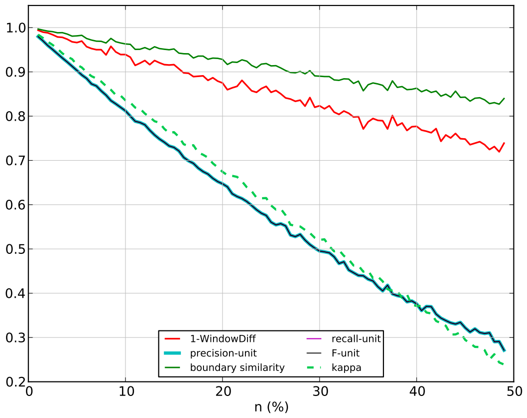

Figure 8: Introducing near misses in prosody dataset

than from the less represented one increases the agreement by chance which lowers κ. However, in our case we see that inserting completely erroneous boundaries is still worse than removing good ones for the κ score.
Concerning the interesting inversion of WD and BS on the two datasets, a deeper investigation is needed but it should be related to the difference in the length distributions. Indeed, WD and BS should not be sensitive to average unit length but they probably can be sensitive to drastically different length distributions.

6. Conclusions and Future Work

In this paper, we proposed a comparison of evaluation metrics for segmentation. Some interesting observations were made concerning the effect of the structure of the data. The results shown in the paper argue, once again, for the need to be careful when providing evaluation scores. Using more subtle scores is not enough, we have to be able to interpret them and our benchmark in a step in this direction.

As for future work, on the evaluation side itself, we would like to investigate hierarchical segmentations (Carroll, 2010) and to extend this work using multiple segmentations. Concerning the applications, we will take these

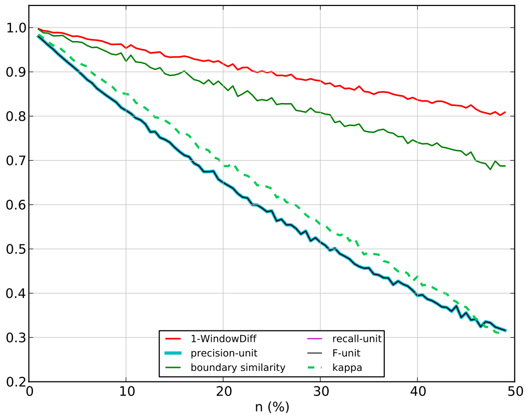

Figure 9: Introducing errors in discourse dataset

results into account when evaluating our annotation campaigns as well as automatic tools.

Acknowledgments

The authors would like to thank the region Provence-Alpes-Côte d'Azur and the company Semantia which contributed to financing this work.

7. References

Carroll, L. (2010). Evaluating hierarchical discourse segmentation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 993–1001. Association for Computational Linguistics.

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

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 Eight International Conference on Language Resources and Evaluation (LREC'12).*

Fournier, C. and Inkpen, D. (2012). Segmentation similarity and agreement. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 152–161.

Fournier, C. (2013). Evaluating text segmentation using boundary edit distance. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA*, volume 5.

Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012). Manual corpus annotation: Giving meaning to the evaluation metrics. In *Proceedings of the International Conference on Computational Linguistics (COLING 2012)*, page 809–818, Mumbaï, Inde, December. Quaero.

Peshkov, K., Prévot, L., Bertrand, R., Rauzy, S., and Blache, P. (2012). Quantitative experiments on prosodic and discourse units in the corpus of interactional data. In *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*, pages 181–182, Paris, September.

Peshkov, K., Prévot, L., and Bertrand, R. (2013). Evaluation of automatic prosodic segmentations. In *Proceedings of Prosody-Discourse Interface 2013*, Leuven, September.

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

Tjong, E., Sang, K., and Déjean, H. (2001). Introduction to the CoNLL-2001 shared task: clause identification. In *Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7*, page 8.