A bioinformatics solution to inter-rater agreement for forced time-alignment of data from underresourced languages

Matthew Stave,[1] François Delafontaine,[1] Frank Seifart,[2] Ludger Paschen,[2]
(1) DDL, 14 avenue Berthelot, 69363 Lyon, France
(2) ZAS, Schützenstrasse, 10117 Berlin, Deutschland
matthew.stave@cnrs.fr, francois_delafontaine@outlook.com,
frank.seifart@berlin.de, paschen@leibniz-zas.de

Résumé

Une solution de la bioinformatique à l'accord inter-annotateurs.
Un alignement précis des séquences d'annotations est un prérequis à l'analyse phonologique, et un aspect important de la linguistique de l'oral. Le projet DoReCo exploite l'alignement phonémique produit par le logiciel MAUS (Kisler et al. 2017) et, pour en évaluer la préci sion, a mesuré l'accord inter-annotateurs entre cet alignement automatique et des alignements manuels, ce qui pouvait impliquer l'ajout, le retrait ou l'altération d'unités d'annotation. Cette situation se révèle très problématique pour l'accord inter-annotateurs. L'algorithme de Needleman-Wunsch, du champ de la bioinformatique, offre une solution pratique et puissante. Son implémentation, comparée à une correction manuelle, connecte plus de 95% des unités correctement. L'algorithme offre une précision nouvelle pour la mesure de l'accord inter-annotateurs, et s'applique à d'autres tâches où cette mise en relation est requise.

Abstract

Precise time-alignment for sequences of annotations is a prerequisite for phonological analysis, and an important aspect of oral linguistics. The DoReCo project relies on phonemic time-alignment by MAUS software (Kisler et al. 2017) and, to evaluate its precision, must measure inter-rater agreement between MAUS-aligned and manually-aligned segments, which involve adding, removing, and changing annotation units. This situation proves highly problematic for inter-rater agreement. The Needleman-Wunsch algorithm, from the bioinformatics field, offers a practical and powerful solution to that problem. Its implementation, when compared with a manual correction, matched over 95% of all units correctly. The algorithm offers a newfound precision for inter-rater agreement measurement, and has further applications where precise matching is required.

Mots-clés : Linguistique de corpus ; linguistique de l'oral ; accord inter-annotateurs ; Needleman-Wunsch ; bioinformatique.

Keywords: Corpus linguistics ; oral linguistics ; inter-rater agreement ; Needleman-Wusch ; bioinformatics.

1 Introduction

We propose an innovative solution to a well-known problem in the area of inter-rater agreement, regarding the problem of measuring segmentation agreement on annotations which differ in content and/or number of units (Mathet et al. 2015). This solution from bioinformatics, called the Needleman-Wunsch algorithm (Needleman & Wunsch 1970), allows for the pairing of annotation units and, from there, the precise measurement of agreement.

First we will present the context of this work, that is, the DoReCo project and its specific inter-rater agreement task (point 2). We will then review some methods to address that task (point 3) before presenting the Needleman-Wunsch implementation and its results (point 4).

2 An alignment problem in the DoReCo project

DoReCo[1] is a new French-German collaborative project designed to bring together spoken language corpora from 50+ languages, taken from documentations of small and often endangered languages. This is done both to bring awareness to lesser-studied languages and language communities, and to enable access to a more diverse sample of languages for researchers to test linguistic hypotheses. The DoReCo project itself will use the corpora to explore a number of questions related to universal claims made about language production in articulatory phonetics and information-processing.

One important contribution of the project is the time-alignment of all the transcriptions, using the MAUS time-alignment software (Kisler et al. 2017). This is necessary for answering questions of phonetic (in)compressibility and final lengthening. Using MAUS, transcriptions are time-aligned at the word and phoneme level, using a global phonemic-alignment model, and these aligned corpora are then made available to the public.

Due to this reliance on the MAUS alignment, it was necessary to test the accuracy of the obtained phonemic tiers. This was to be done by comparing MAUS's segment boundaries with segment boundaries from manually-corrected tiers, to determine the agreement between word and phoneme boundaries. However, measuring agreement between sequences with different time boundaries (segmentation), as opposed to different content (categorization), is known to be problematic (Mathet & Widlöcher 2016). As it turns out, the manual corrections of the MAUS alignments can involve adding, removing, or altering the content of segments, thus rendering the comparison uncertain in multiple ways. Without knowing which two segments to compare, assessing the accuracy of the MAUS alignment is not possible.

3 Existing tools & methods

The standard method for evaluating categorical agreement of items between annotators is Cohen's kappa (Cohen 1960, 1968), which calculates the proportion of observed agreement and normalizes it by the predicted chance agreement. Other methods, such as boundary distance, can be used to evaluate continuous agreement of temporal boundaries, but these methods require that items be properly aligned with corresponding items on the other tier(s). When tiers are not aligned (e.g. after insertions or deletions), it becomes necessary to first determine this alignment between the tiers to be compared. Without this alignment, at best the mean distribution can be obtained, as for example with Krippendorff's alpha (Krippendorff 1970).

This task of aligning segments is often done manually, but when working with hundreds of thousands of annotation units across thousands of files, as in the DoReCo corpora, this kind of manual alignment becomes unfeasible. There has been some work on the development of automated

[1] <http://doreco.info>

methods of segment alignment. Holle & Rein (2015) have created the EasyDIAG tool, which uses categorical (same tier type and same label) and continuous (percent overlap) approaches to align segments. The STACCATO algorithm (Lücking et al. 2011) uses mutual overlap of multiple annotators to determine "nuclei" segments. Other approaches ignore alignment altogether, simply reporting the raw amount of overlap of any segment with another segment, as a proportion of the segment lengths, averaged across all segments (Strunk et al. 2014).

A method called Gamma (Mathet et al. 2015) has been specifically devised for this task. It uses a unified approach whereas pairing units and measuring their agreement is done parallel to each other, in one process. The result is not only a measurement, but also an automatically realigned annotation. It offers a good alternative to the most common method to our knowledge, which is atomization (*idem* : 440 ; Krippendorff 2004), or the reduction of the segmentation problem to a categorization one by segmenting the annotations further down into intervals of equivalent size. Atomization then allows for a simple Kappa-score, making for an easily sharable measurement.

These methods, however, while often quite useful for comparing aligned segments, proved inadequate for aligning MAUS sequences with manually-corrected sequences. This led us to explore approaches to sequence alignment in other fields, such as bioinformatics.

4 A bioinformatic solution

The Needleman-Wunsch algorithm (Needleman & Wunsch 1970) is a well-known algorithm in bioinformatics, where it is frequently used in aligning amino acid sequences in proteins, or nucleotide sequences in DNA strings. It has the advantage of returning the optimal alignment(s) of two sequences of items, based only on their labels and their positions in the sequences. With very little adjustment (here our labels are word or phoneme strings, rather than amino acids), this can be employed on linguistic annotations to find the optimal alignment of two sequences of words, phonemes, or other annotations.

The algorithm works by maximizing the similarity between the two sequences, allowing for three edit operations: insertions, deletions, and substitutions. Each of these edit operations can be weighted differently, depending on one's theoretical considerations (for our study we have weighted all three equally). A matrix (length(seqA) by length(seqB)) is created, populated by the similarity score for every pairwise combination in the two sequences: identical items receive +2, while insertions, deletions, and substitutions receive -1. This generates a set of optimal paths from the first pair of items to the final pair of items. The alignment paths with the highest score are returned. Within the DoReCo project, the implementation of these algorithm is done using a Python library called *biopython*, and its "pairwise2" function. A preliminary test was done on transcriptions extracted from ELAN tiers in three languages: Anal (India), Resigaro (Colombia), and Vera'a (Vanuatu), with a total of 2503 words.

To test the method, a Cohen's kappa was performed on the Needleman-Wunsch-aligned sequence and a manually-aligned sequence. The reliability of the algorithm compared to human matching of sequences resulted in a kappa-score of 0.97-1, with 95-99% of sequences being aligned the same way across the files: most of the divergence is due to misinterpreted pauses. At that level of reliability, and for the purpose of inter-rater agreement, this suggests the method can be entirely automated. As for the results it gave: we were notably able to establish mean differences in unit onsets and offsets, meaning how much each unit boundary, start and end, was moved, in (milli)seconds, as well as the proportion of moved boundaries and as such, the amount of work a manual correction required. We were also able to reliably track substitutions, additions and deletions due to corrections.

5 Perspectives

Preliminary results have been very promising. Further work will include expanding the sample of test languages, and of course making use of the aligned segments to assess the accuracy of the MAUS word and phoneme alignment. Which method to employ for this latter task is still under consideration.

We do however see applications even beyond inter-rater agreement. This pairing will, as examples, help with operations such as the merging of different transcription files, or the realignment of morphemic units under their word-corrected counterpart. While such tasks do require some manual correction of the automatic pairing to be fully effective, they are made possible by this very bioinformatics method.

References

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.

Holle, H. & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior research methods* 47(3), 837-847.

Lücking, A., Ptock, S. & Bergmann, K. (2011). Assessing agreement on segmentations by means of Staccato, the SegmenTation Agreement CalCulator According to Thomann. *Proceedings of the 9th international conference on Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*. <researchgate.net/publication/262171036_Assessing_Agreement_on_Segmentations_by_Means_of_Staccato_the_Segmentation_Agreement_Calculator_according_to_Thomann>.

Kisler, T., Uwe, R.D. & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer speech & language* 45, 326-347.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement* 30, 61-70.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks: Sage.

Mathet, Y., Widlöcher, A. & Métivier, J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational linguistics* 41(3), 437-479.

Mathet, Y. & Widlöcher, A. (2016). Évaluation des annotations : ses principes et ses pièges. *TAL* 57(2), 73-98.

Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48(3), 443-453.

Strunk, J., Schiel, F. & Seifart, F. (2014). Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. *LREC 2014*, Reykjavik, Iceland.