Rooting through Direction – New and Old Approaches

Hoenen, Armin
hoenen@em.uni-frankfurt.de
Goethe Universität Frankfurt, Deutschland

Introduction

Computational stemmatology is the discipline dealing with the reconstruction of the copy history of historical works, primarily transmitted by handwriting. The result is a directed acyclic graph or tree. Trees generated by computational means are often unrooted. While in biology, all lifeforms are connected in a huge tree of life (compare tolweb.org), for texts such a tree seems an unsuitable metaphor. For this reason, rooting a tree in biology is relatively easy while it is relatively hard in stemmatology. Biologists in the majority of cases use a so-called outgroup, that is a species remotely - but not too remotely - related to the group under investigation. Then they identify root that node which is closest to the outgroup since the outgroup is considered to indicate where the group connects with the tree of life. There are some alternative rooting procedures none of which is applicable to stemmatology generally. Haigh (1970, 1971) proposes to assign root according to probabilities assuming a certain type of birth-process as a generating function for stemmata but immediately relativises his proposal saying that the process used is not historically realistic. Yet other approaches are pursued by Marmerola et al. (2016). We implement the approach of Haigh (1970), one of the approaches of Marmerola et al. (2016) (called minimum-cost heuristic) and present our own two approaches which we then test on 3 artificial stemmatological datasets.

Approach

Correctly classifying all directions of the edges in a tree would provide us with root. Such a classification would be one method to obtain a root, but would have a caveat. If only a few edges would be classified wrongly, the resulting contradiction could be resolvable in more than one equally probable way. Instead of looking at single edges or the complete tree, we focus on paths, namely all paths from one leaf to another leaf. The situation we start from is thus an unrooted tree (UT) with exactly one covertly underlying rooted tree (RT), the gold standard. What we can distinguish in the UT are only leafs and internodes. For any leaf, we suppose that there is at least one fellow leaf which is so remote that their latest common ancestor is root. Furthermore we do know that any shortest path from leaf to leaf in the UT must pass through the latest common ancestor of both leafs. We want to denote it by # l_i, l_j but we initially don't know which of the nodes on the path in the UT it is. # is pivotal for directionality in the RT: all directions of edges towards the two leafs point away from it. Thus, traversing the path by edges from one leaf to the other, direction changes in the RT when passing #. If, as we supposed, for each leaf there is at least one other leaf for which # is root, then, comparing all leaf-leaf paths starting in a leaf l_1, the # most distant from l_1 must coincide with root since there can be no later common ancestor in the RT. Moreover, all leafs should converge on the same node as the most distant #. Thus, if one could detect that node on any leaf-leaf path which is pivotal for direction, one would know for each such path the direction change point # and in consequence root. Detecting the most probable point of direction change on a path may be easier and lead to less contradictions than classifying each edge for direction since the path itself provides context. However, there is a case we have so far not taken into account: the case that RT is a planted tree. In that case, the highest direction change convergence point (HDCCP) would be the first node in the RT that has more than one child. Also, there would be one leaf, on the path to which all other leaves do not detect any direction change, which our algorithm tracks. For a graphical explanation in case root is internal, see Figure 1.

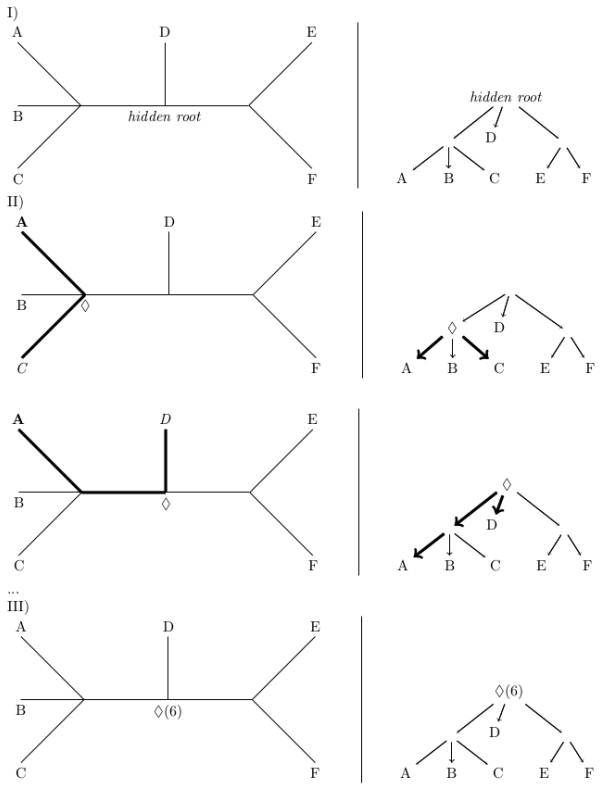

*Figure 1. Detection of root from an unrooted tree (I, left) with an underlying rooted one (right). For each leaf on each path to another leaf, the node of direction change is detected and marked,here with # (II, **A**-C and **A**-D exemplarily). For each leaf separately, the #*

furthest away from it is chosen and if direction change detection works 100 %, all nodes converge on true root (III).

Direction

In practice, assigning direction is utterly complicated by the fact that only very few features allow an unambiguous detection of direction. In biology, Gogarten et al. (1989), Iwabe et al. (1989) have used gene duplications to root groups of bacteria at the root to the tree of life itself for which outgroups are not applicable. The logic behind this is that gene copies are relatively specific and that once present they are unlikely to be lost. The best parallel in textual criticism are probably certain types of skips such as line skips. The problem for a copyist having a model with a lineskip is that they cannot restitute the text unless knowing it by heart or having another exemplar. Now, lineskips appear to be no mass phenomenon and other indications for direction are rather rare and relatively unstable. For instance, one could try to classify variants for (linguistic or relative) age as done in the so-called CBGM method (Mink 2004) and then assign direction change to the node on a leaf-leaf path which has relatively most oldest variants.[1] But, the classification could be tricky and require masses of input data. Or one could take a heuristic explained by West (1973:32) who elaborates on groups of manuscripts sharing variants but this criterion may be too weak in our case and in addition lead to some overweighting of vulgate variants. We choose psycholinguistically obtained asymmetric letter confusion probabilities instead. Those are largely constrained to one-to-one aligned letter pairs and in a nutshell give a probability of whether the letter sequence <tap> is more likely to be copied from or to <top>. Hoenen (2018) used these to determine distances (MMD) to compute stemmata and found that they are only interpretable with roughly a third of the cases of variation in the three artificial datasets Parzival (Spencer et al. 2004), Notre Besoin (Baret et al. 2004) and Heinrichi (Roos & Heikkilä 2009). Furthermore due to orthographic depth (Katz & Frost 1992) Finish (Heinrichi) was more processable than English (Parzival) or French (Notre Besoin). We detect that node as most probable DC point on a leaf-leaf path which has a minimal sum of MMD on the way to each leaf. Marmerola et al. (2016) use a heuristic where they take each node in UT as tentative root, sum the weights on all paths to all leafs and finally, comparing all possible roots take the overall minimum cost root. Weights come from one of their similarity functions, but in our implementation, we simply take the Hamming distance (Hamming 1950) and the MMD. Additionally, we implement a similar approach using the MMD distance only. We take each UT node as tentative root, then sum weights not on all paths root-leaf *in* but simply on all edges *of* the actual tree, since our distance matrix is asymmetric and will thus lead to a different sum for each tentative root. We choose the minimum cost root (LCM mincost). Finally, we implement the above outlined approach of the HDCCP root where we use again the letter confusion matrices, this time to determine the most likely DC nodes (#) on the leaf-leaf paths (HDCCP LCM). We used the matrix of Geyer (1977)[2] for lowercase andPaap et al. (1982) for uppercase.

Results

The results can be seen in Table 1.

Tradition	True root	Haigh MML	Marmerola et al. (2016), Hamming/ MMD	LCM mincost, MMD	HDCCP LCM, MMD
Parzival	LM	LM	LM	JW(2)	SD(2)
Notre Besoin	n10	n9 (2)	n9(2)	n9(2);n11(1)	n9(2); planted
Heinrichi	h1	h1	h1	h16(4)	h1

Table 1. Results for different approaches to rooting. In brackets length of path to true root.

Discussion

While in 1970/1971 the artificial traditions were not yet present, Marmerola et al. (2016) likewise do not report their results on the complete artificial datasets but on their estimated stemmata. Both approaches yielded very good results and identified root twice. LCM mincost performed clearly worse, surprisingly also worse than Marmerola et al. (2016) with MMD, probably since the magnitude of differences when summing the paths is larger. Moreover, HDCCP LCM on the same distance (MMD) performs better than LCM mincost and identified root in one case. For the second tradition for which the Gold Standard is a planted tree, all methods had problems. HDCCP LCM detected that root could be a leaf and gave as candidates for that case n8, n10 (true root), n4 and n6. Hoenen (2018) had reported that for Finish due to orthographic depth, the psycholinguistic matrices would perform best which is consistent with our results. The artificial traditions are hardly representative of historical traditions in size or depth and thus allow only a rough approximation of the implications of these preliminary results. Also, MMD is only one way to approach the detection of directionality. The artificial traditions have been produced by recent scribes, their time-depth is not comparable to historical traditions and their sizes and source languages are not representative of historical data, which is why a more in depth investigation and a closer look onto the results is a priority for future elaboration of the method.

Conclusion

We have demonstrated the first implementation of a new rooting algorithm which is applicable not only for stemmatological but also biological (molecular and character data, substitution matrices) and historical linguistic data (lexico-statistics, sound shifts). The approach yielded encouraging results.

Fußnoten

1. Here, one could also use variants classified as older through the principles of lectio difficilior or lectio brevior. Some first Machine Learning based classifications suggest that this principle, at least for the artificial datasets mentioned below is a too low frequency phenomenon for directionality classification.
2. Boles & Clifford (1989) data yielded the same results on HDCCP LCM.

Bibliographie

Baret, Philippe / Macé, Caroline / Robinson, Peter (2004): *"Testing methods on an artificially created textual tradition"* in: **Baret, Philippe / Macé, Caroline / Bozzi, Andrea / Cignoni, Laura**: *The evolution of texts: confronting stemmatological and genetical methods* Linguistica Computationale XXIV-XXV. Pisa/ Rom: Instituti Editoriali e Poligrafici Internationali, 255–281.

Boles, David B. / Clifford, John E. (1989): *"An upper- and lower case alphabetic similarity matrix, with derived generation similarity values"*. Behav. Res. Methods Instrum. Comput. 21, 597–586.

Geyer, L. H. (1977): *"Recognition and confusion of the lowercase alphabet"*. Percept. Psychophys. 22, 487–490.

Gogarten, Johann P. / Kibak, Henrik / Dittrich, Peter / Taiz, Lincoln / Bowman, Emma J. / Bowman, Barry J. / Manolson, Morris F. / Poole, Ronald J. / Date, Takayasu / Oshima, Tairo (1989): *"Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes."* Proc. Natl. Acad. Sci. 86, 6661–6665.

Haigh, John (1970): *"The recovery of the root of a tree"* J. Appl. Probab. 7, 79–88.

Haigh, John (1971): *"Mathematics in the Archaeological and Historical Sciences"* Edinburgh University Press, Scotland, UK, 396–400.

Hamming, Richard W. (1950): *"Error detecting and error correcting codes."* Bell System technical journal, 29(2), 147-160.

Hoenen, Armin (2018): *"Multi Modal Distance - An Approach to Stemma Generation with Weighting"* in: Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2018. Miyazaki (Japan).

Iwabe, Naoyuki / Kuma, Kei-ichi / Hasegawa, Masami / Osawa, Syozo / Miyata, Takashi (1989): *"Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes"* Proc. Natl. Acad. Sci. 86, 9355–9359.

Katz, Leonard / Frost, Ram (1992): *"The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis"* Haskins Lab. Status Rep. Speech Res. SR-111, 147–160.

Marmerola, Guilherme D. / Oikawa, Marina A. / Dias, Zanoni / Goldenstein, Siome / Rocha, Anderson (2016): *"On the Reconstruction of Text Phylogeny Trees: Evaluation and Analysis of Textual Relationships"* PloS One 11, e0167822.

Mink, Gerd (2004): *"Problems of a highly contaminated tradition: the New Testament: Stemmata of variants as a source of a genealogy for witnesses"* in: **van Reenen, P., den Hollander, A., van Mulken, M. (Eds.)**: *Studies in Stemmatology II. John Benjamins*, pp. 13–86.

Paap, Kenneth R. / Newsome, Sandra L. / McDonald, James E. / Schvaneveldt, Roger W. (1982): *"An activation-verification model for letter and word recognition: the word-superiority effect"* Psychol. Rev. 89, 573–594.

Roos, Teemu / Heikkilä, Tuomas (2009): *"Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets"* Lit. Linguist. Comput. 24, 417–433.

Spencer, Matthew / Davidson, Elizabeth A. / Barbrook, Adrian C. / Howe, Christopher J. (2004): *"Phylogenetics of artificial manuscripts"* J. Theor. Biol. 227, 503–511.

West, Martin L. (1973): *"Textual Criticism and Editorial Technique: Applicable to Greek and Latin texts"* Teubner, Stuttgart.