How Good are Typological Distances for Determining Genealogical Relationships among Languages?

Taraka Rama[1] and Prasanth Kolachina[2]
[1] Språkbanken, University of Gothenburg & [2] Language Technologies Research Centre, IIIT-Hyderabad

Introduction

- **Language family**: There are more than 7000 languages in this world (Lewis 2009), which fall into more than 140 genetic families.
- **Typological relatedness**: Languages can also share structurally common features such as *word order*, *similar phoneme inventory size* and *morphology*. But typological relatedness does not imply genetic relation between languages.
- In computational linguistics, genealogical distances between two language families have been shown to be useful for predicting the difficulty of machine translation (Birch et al. 2008).

Contributions

- Do we really need a clustering algorithm to measure the internal classification accuracy of a language family?
- How well do the typological distances within a family correlate with the lexical distances derived from ASJP lists?
- Given that there are more than dozen vector similarity measures, which vector similarity measure is the best for the above mentioned tasks?

Resources

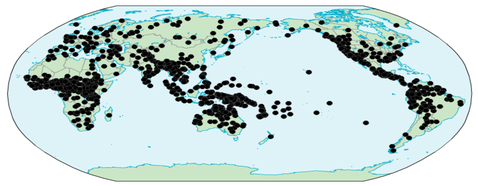
- **WALS**: The WALS database [a] has typological features for 2676 languages of 144 types. We removed entries for all languages with less than 25 attested features and features with less than 10% attestations (Georgi et al. 2010). Each of the WALS features (binary or multi-valued) are also binarized by recording presence or absence of a particular feature value.
- **ASJP**: A database of Swadesh word lists (Swadesh 1952) (short concept meaning list) for more than 58% of the world's languages (Brown et al. 2008).

[a] accessed on 2011-09-22

Previous work

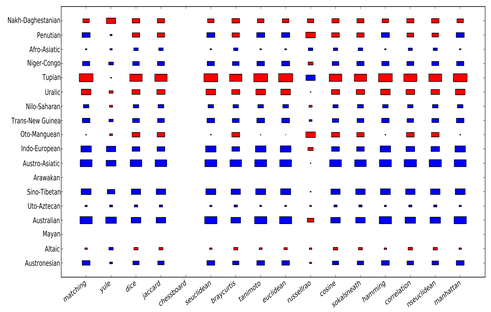
- Daume III (2009) and Georgi et al. (2010) use typological features from WALS to investigate relation between phylogenetic groups and feature stability.
- Georgi et al. (2010) motivate the use of clusters derived from typological distances to project linguistic resources from "resource-rich" to "low-resource" languages.
  - Do not take into account *geographical bias* in the dataset.

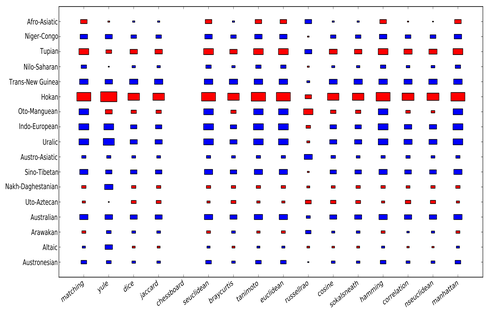
Experiments



- Distances between feature vectors from the WALS typological database are computed using 15 different similarity measures.
- For each similarity measure
  - The typological distance matrix is compared to a 2D WALS classification matrix
  - Pair-wise correlation between the typological distance matrix and the lexical distance matrix used in ASJP classification.
  - ASJP distance between two languages is computed as average pair-wise length-normalized Levenshtein distance (LDND).

Results





- **Internal classification accuracy** using point-biserial correlation shown above.
- Australian, Austro-Asiatic, Indo-European and, Sino-Tibetan language families, except for 'russellrao'.
- The worst performing language family is Tupian. Tupian has 5 genera with one language in each and a single genus comprising the rest of family.
- None of the vector similarity measures seem to perform well for Austronesian and Niger-Congo families.
- Correlation between typological distance matrices and distance matrix obtained using lexicostatistical lists from ASJP.
- Pairwise correlation ρ is high across Australian, Sino-Tibetan, Uralic, Indo-European and Niger-Congo families.
- The Hokan family shows the lowest amount of correlations across all distance measures.

Conclusions and Future Work

- Choosing the right vector similarity measure when calculating typological distances makes a difference in the internal classification accuracy.
- Choice of similarity measure does not influence the correlation between WALS distances and LDND distances within a family.
- Combination of a smaller set of typological features (from the ranking of Wichmann & Holman (2009)) and right similarity measure might achieve higher accuracies.

References

Birch, A., Osborne, M. & Koehn, P. (2008), Predicting Success in Machine Translation, *in* 'Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Honolulu, Hawaii, pp. 745–754.

Brown, C., Holman, E., Wichmann, S. & Velupillai, V. (2008), 'Automated classification of the world's languages: a description of the method and preliminary results', *Sprachtypologie und Universalienforschung* **61**(4), 285–308.

Daume III, H. (2009), Non-parametric bayesian areal linguistics, *in* 'Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, pp. 593–601.

Georgi, R., Xia, F. & Lewis, W. (2010), Comparing language similarity across genetic and typologically-based groupings, *in* 'Proceedings of the 23rd International Conference on Computational Linguistics', Association for Computational Linguistics, pp. 385–393.

Lewis, P. M., ed. (2009), *Ethnologue: Languages of the World*, Sixteenth edn, SIL International, Dallas, TX, USA.

Swadesh, M. (1952), 'Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos', *Proceedings of the American philosophical society* **96**(4), 452–463.

Wichmann, S. & Holman, E. W. (2009), 'Assessing temporal stability for linguistic typological features', *München: LINCOM Europa* .

Acknowledgements

The research presented here was supported by the Swedish Research Council (the project Digital areal linguistics, VR dnr 2009-1448) and by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken. We would like to thank Harald Hammarström, Lars Borin and Søren Wichmann for the discussions and their insights into this work. We would also like to thank the anonymous reviewers for comments on the paper.