Proto-word reconstruction with NNs

Cognate sets & proto-words

- Cognate set: N-tuple of related/homologous words in n languages:

$$< father, Vater, vader, fa\eth ir > \tag{1}$$

- Proto-word: The common ancestor from which the words in the cognate set descend, in (1) from Proto-Germanic **fadēr*

1 Work to build on

- Bouchard-Côté et al. (2013), phylogenetic inference performed on a large Austronesian dataset (reversible-jump MCMC, so not strictly what we want). The goal is to reconstruct a phylogenetic tree, and Bouckaert et al. (2012), also inferring geographic diffusion of the IE family.

- Ciobanu and Dinu (2018), using conditional random fields & RNNs Ciobanu and Dinu (2018) to reconstruct Latin words from Romance cognate data.

- Meloni et al. (2019) also RNN, pipeline similar to neural machine translation. Inputs are character + language embedding vectors. (Encoder-Decoder)

- Cognate identification with siamese CNNs using either string similarity metrics as Levenshtein distance Soisalon-Soininen and Granroth-Wilding (2019) or phonetic feature arrays (multi-hot encodings) Rama (2016).

Features	p	b	f	v	m	8	4	t	d	s	z	c	n	S	Z	C	j	T	5	k	g	x	N	q	G	X	7	h	l	L	w	y	r	!	V
Voiced	0	1	0	1	1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1
Labial	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Dental	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alveolar	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Palatal/Post-alveolar	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Velar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0
Uvular	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
Glottal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
Stop	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0
Fricative	1	1	1	1	0	1	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0
Affricate	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nasal	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Click	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Approximant	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0
Lateral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Rhotic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Figure 1: Phone encodings in Rama (2016), p.1021

2 Data sources

- Wiktionary
  - Many datapoints of dubious quality
  - Would have to do most of the data extraction ourselves
  - Data dumps available here
  - Technically it should be possible possible to get RDF data
- Indo-European lexical cognacy database
  - Used by famous Bouckaert et al. (2012)
  - No longer maintained (since 2016)
  - Only Indo-European data, which may be a bit over-investigated
  - But: plain TSV
- Evolution of human language project (used in Hruschka et al. (2015)). provides cognate data for several Eurasian language families (Altaic, Tungusic, Mongolic, Japonic...)
  - I didn't know the format (dBase/.dbf), don't know exactly how to use
  - Somewhat outdated (2013)
  - Pro: Many languages from many families
  - Could try to reconstruct proto-Altaic (which is a deprecated clade) or proto-Transeurasian

3 Model architecture

- Code letters for phonological features
  - Word = $n_{letters} \times n_{features}$ array (following Rama (2016)).
    Example for PGmc **fadēr* with ASJP [1] encodings:

    $$\begin{Bmatrix} f & a & d & \bar{e} & r \\ f & a & 8 & e & r \end{Bmatrix}$$

  - The exact coding depends on the orthographical data available, or we have to do the encoding based on what we know about the exact phonetics of the languages.
- MT-like pipleine:
  - One cognate as input per time step → encoder
  - Decoder produces proto-word candidate
  - Encoding of the true proto-word as ground truth
  - Outputs not only probability distribution over a set of phonetic features, but phonetic encodings → visualization of errors (not present in the papers I found)
- CNN:
  - Difficult for me to get the intuition
  - In a cognate set like
  - Kernels should concentrate on phonetic features and/or languages most relevant for reconstruction
  - Architecture could be very flexible

[1] https://en.wikipedia.org/wiki/Automated_Similarity_Judgment_Program

4 Work split

1) Until next meeting (07.07.2020):

   - Decide on language family (should be the one we can get best data for?)
   - Get a data sample for development (can be small, produced by hand)
   - Decide on approach (RNN vs. CNN)

2) After project start:

   - Decide on default architecture (everybody)
   - Get training data (about 2-3000 cognate sets would be ideal, alternatively we can use Swadesh lists)

     → We shouldn't have to change the encoding later
   - Implement the model
   - Visualization

References

Bouchard-Côté, A., Hall, D., Griffiths, T. L., and Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960.

Ciobanu, A. M. and Dinu, L. P. (2018). Ab Initio: Automatic Latin Proto-word Reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1604–1614, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hruschka, D., Branford, S., Smith, E., Wilkins, J., Meade, A., Pagel, M., and Bhattacharya, T. (2015). Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution. *Current Biology*, 25(1):1–9.

Meloni, C., Ravfogel, S., and Goldberg, Y. (2019). Ab Antiquo: Proto-language Reconstruction with RNNs. *arXiv:1908.02477 [cs]*. arXiv: 1908.02477.

Rama, T. (2016). Siamese convolutional networks based on phonetic features for cognate identification. *arXiv:1605.05172 [cs]*. arXiv: 1605.05172.

Soisalon-Soininen, E. and Granroth-Wilding, M. (2019). Cross-Family Similarity Learning for Cognate Identification in Low-Resource Languages. In *Proceedings - Natural Language Processing in a Deep Learning World*, pages 1121–1130. Incoma Ltd., Shoumen, Bulgaria.