EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
Language Evolution:
The Empirical Turn

Information-Weighted Sequence Alignment

Workshop “Trees and what to do with them”

Tübingen, March 23, 2018

Johannes Dellert

This research has been supported by the ERC Advanced Grant 324246 EVOLAEMP.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Table of Contents

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Cognate Detection

- **cognate sets** in quantitative historical linguistics: sets of etymologically related words (which includes borrowings)
- **cognate detection** task: partitioning a set of words with the same meaning into cognate sets
- can be viewed as a binary classification problem for word pairs: are a from language L_a and b from language L_b cognates?
- most common approach: compute some pairwise form distance measure, use distances as input for clustering algorithm
- benchmark for all recent advances: **LexStat** by List (2012)
- improvements over LexStat in B-Cubed score have been small:
  - ▷ Jäger and Sofroniev (2016): 0.700 → 0.718
  - ▷ Rama et al. (2017): 0.819 → 0.841 (NED: 0.804)
  - ▷ List et al. (2017): 0.883 → 0.894 (NED: 0.814)

Table of Contents

Information Weighting: Idea

- recent advances mainly driven by better clustering methods:
  - ▷ List et al. (2017) show that LexStat distances are the best, but InfoMap clustering beats UPGMA clustering on them
  - ▷ improvement in Rama et al. (2017) is also partially due to InfoMap clustering (in addition to better PMI scores)
- what about the other component? any clustering method would profit from improvements to the form distances
- observation: not all segments in a word are equally important
- simple rules like focusing on the first syllable do not generalize, a specialized model would be needed for every language
- instead: use trigram models to learn from the data which parts are more relevant for comparison!

Information Weighting: Definition

Segment-wise **information content** of *c* in context *abcde*:

$$I_L(c, [ab\_de]) := -\log \left\{ \frac{c_{abc} + c_{bcd} + c_{cde}}{c_{abX} + c_{bXd} + c_{Xde}} \right\}$$

- c_{abc}, c_{abX}, c_{Xbc}, c_{aXc} are trigram and extended bigram counts extracted from all word forms of *L*
- expanded by # at word boundaries (creating a full context)
- the quotient defines a probability distribution $P(c, [ab\_de])$ over possible segments *c* in context $[ab\_de]$
- $I_L(c, [ab\_de])$ is a measure of surprisal or self-information!

Information Weighting: Examples

- example for [t͡ɕ] in Polish *dać* [dat͡ɕ] “to give”:
  I_{pol}(t͡ɕ,[da_##])
  $= (c_{\text{dat͡ɕ}} + c_{\text{at͡ɕ\#}} + c_{\text{t͡ɕ\#\#}})\ /(c_{daX} + c_{daX\#} + c_{X\#\#})$
  $= (13 + 132 + 350)/(30 + 339 + 1124)$
  $= 1.287$
- for comparison: I_{pol}(d,[##_at͡ɕ]) $= 3.306$

Table of Contents

Information-Weighted Sequence Alignment (IWSA)

- idea: modify Needleman-Wunsch algorithm
- multiply achievable score for each operation by a combined information score baed on information models of both languages
- when computing the costs for an alignment, give a discount for alignment of ill-fitting material that has low information content in both languages
- at the same time, avoid aligning high-information material to low-information material (e.g. stems to suffixes)

IWSA: Definition

- aligning two IPA strings $a \in L_a$ of length m and $b \in L_b$ of length n
- combined information content for two aligned segments:

$$I^2_{L_a,L_b}(a_i, b_j) := \sqrt{\frac{I_{L_a}(a_i, [a_{i-2} \ldots a_{i+2}])^2 + I_{L_b}(b_j, [b_{j-2} \ldots b_{j+2}])^2}{2}}$$

- modified dynamic programming procedure for computing $sc(a, b) := M(m, n)$:

$$\begin{aligned}
M(0,0) &:= 0 \\
M(i,0) &:= M(i-1,0) + w(a_i, \epsilon) \cdot I^2_{L_a,L_a}(a_i, a_i) \\
M(0,j) &:= M(0,j-1) + w(\epsilon, b_j) \cdot I^2_{L_b,L_b}(b_j, b_j) \\
M(i,j) &:= min \begin{pmatrix} M(i-1,j-1) + w(a_i, b_j) \cdot I^2_{L_a,L_b}(a_i, b_j), \\ M(i-1,j) + w(a_i, \epsilon) \cdot I^2_{L_a,L_a}(a_i, a_i), \\ M(i,j-1) + w(\epsilon, b_j) \cdot I^2_{L_b,L_b}(b_j, b_j), \end{pmatrix}
\end{aligned}$$

IWSA: Examples

Opacity represents $I^2_{L_a,L_b}(a_i, b_j)$, color represents $w(a_i, b_j)$:

German	f	ɛ	ɐ	z	ɪ	ŋ	k	ə	n	"to sink"
English	-	-	-	s	ɪ	ŋ	k	-	-	

Arabic	θ	a	l	-	d͡ʒ	"snow"
Hebrew	ʃ	ɛ	l	ɛ	g	

Information-Weighted Distance

For words *a* of length *m* and *b* of length *n*:

$$d(a, b) := 1 - \frac{2 \cdot \frac{sc(a,b)}{\max\{n,m\}}}{\frac{sc(a,a)}{m} + \frac{sc(b,b)}{n}}$$

- unusual normalization by length necessary
  due to very high self-similarity for pairwise similarity scores
- values concentrate in interval [0.6, 1.4],
  no centralisation or normalisation done in this study
- threshold for candidate cognate pairs: $d(a, b) < 1.2$

Table of Contents

Changes to PMI Score Inference

- staying within the PMI framework, building on resampling in the style of Kessler (2001) and List (2012):

$$w_{glo}(x, y) := \log \frac{p(x, y)}{\hat{p}(x, y)}$$

- in the information-weighted case, the $p(x, y)$ and $\hat{p}(x, y)$ are based on **weighted counts** as well:

$$c(x, y) := \sum_{L_1, L_2 \in \mathcal{L}} \sum_{\substack{(a,b) \in lex(L_a, L_b), \\ sc(a,b) < 1.2}} \sum_{\substack{1 \leq i \leq \max\{m,n\}, \\ al(a,b).a_i = x, \\ al(a,b).b_i = y}} I^2_{L_a, L_b}(a_i, b_i)$$

Local scores for sound correspondences

- global PMI scores based on 1.3M cognate candidate pairs from NorthEuraLex 0.9, and an equal number of random word pairs
- local PMI scores (inferred from the data for a single language pair) to represent some of the sound correspondences:

$$w_{L_1,L_2}(x,y) := \frac{w_{glo}(x,y) + \log \frac{p_{L_1,L_2}(x,y)}{\hat{p}_{L_1,L_2}(x,y)}}{2}$$

- $p_{L_1,L_2}(x,y)$ and $\hat{p}_{L_1,L_2}(x,y)$ are estimated like in the global case, five alternations of re-estimation and re-filtering of candidates

Table of Contents

Test Data: intersection of NorthEuraLex and IELex

The testset was generated from an intersection of NorthEuraLex with IELex cognacy judgments (from the webpage):

- 36 Indo-European languages
- 185 concepts
- 100156 binary cognacy judgments
- available as an appendix to my dissertation

Evaluation: Overview

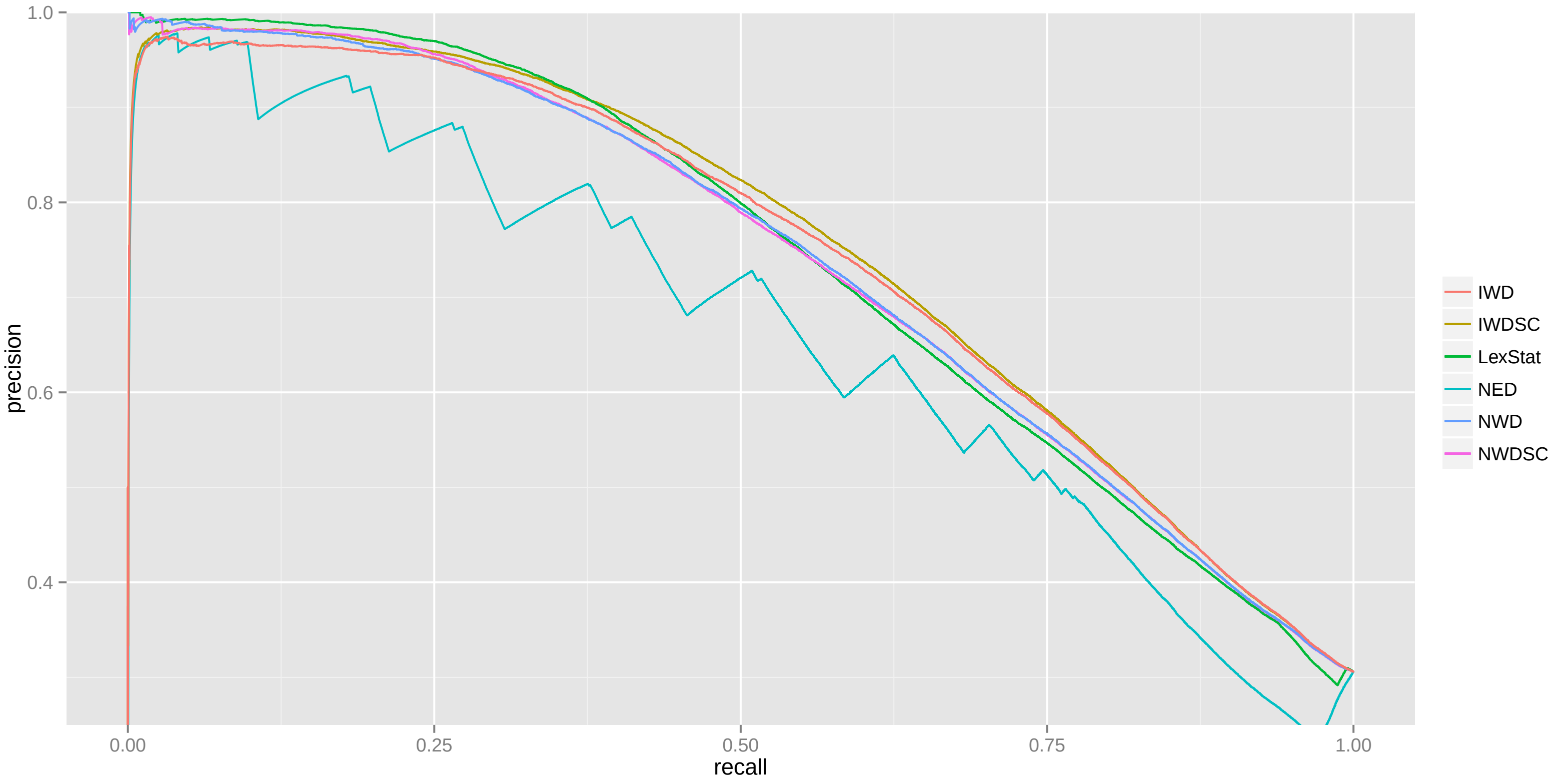
Methods being compared:

- **NED**: Normalized Edit Distance
- **LexStat**: LexStat Distance
- **NWD**: Needleman-Wunsch Distance
- **NWDSC**: NWD with Sound Correspondences
- **IWD**: Information-Weighted Distance
- **IWDSC**: IWD with Sound Correspondences

Evaluation measure: **average precision**

- precision averaged over all recall values
- equivalent to area under precision-recall graph
- threshold-independent criterion
- independent of clustering algorithm

Results: Precision-Recall Graphs



Results: Average Precision

Method	NED	LexStat	NWD	NWDSC	IWD	IWDSC
Avg. Prec.	0.604	0.728	0.741	0.747	0.764	**0.771**
Max. F-score	0.599	0.630	0.652	0.654	0.673	**0.679**
Precision	0.639	0.653	0.666	0.660	0.696	**0.706**
Recall	0.564	0.609	0.639	0.648	0.652	**0.654**

- NWD improves on LexStat by 1.3%, even without SC (advantage for full IPA model on many forms per language?)
- improvements through information weighting and sound correspondences are orthogonal:
  - ▷ information weighting leads to an increase of 2.3%
  - ▷ sound correspondences provide an additional 0.7%

Open Questions

- does information weighting work on smaller wordlists?
- does the advantage disappear on pre-stemmed data?
- how much difference does it make in clustering quality?
- performance of methods on cross-family datasets?
  (where similarity is less predictive of cognacy)

Acknowledgments

- Armin Buch (joint work on early version)
- Pavel Sofroniev (initial version of test set)
- all other members of the EVOLAEMP team
  (building NorthEuraLex, feedback at many stages)
- the ERC (Advanced Grant 324246)

References

Jäger, G. and Sofroniev, P. (2016). Automatic cognate classification with a Support Vector Machine. Proceedings of the 13th Conference on Natural Language Processing (KONVENS).

Kessler, B. (2001). *The significance of word lists. Statistical tests for investigating historical connections between languages.* CSLI Publications, Stanford.

List, J.-M. (2012). LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France. Association for Computational Linguistics.

List, J.-M., Greenhill, S. J., and Gray, R. D. (2017). The Potential of Automatic Word Comparison for Historical Linguistics. *PloS one*, 12(1):e0170046.

Rama, T., Wahle, J., Sofroniev, P., and Jäger, G. (2017). Fast and unsupervised methods for multilingual cognate clustering. *CoRR*, abs/1702.04938.