Multiple Word Alignment with Profile Hidden Markov Models

Aditya Bhargava and Grzegorz Kondrak
Department of Computing Science, University of Alberta
{abhargava, kondrak}@cs.ualberta.ca

MULTIPLE WORD ALIGNMENT

kwatro  kwa-tro
kwattro  kwattro
katr   k-a-tr-
den   d-e-n-
deny   d-e-ny
dzen   d-z-e-n-
dzienn  dz1-e-n-
giorno   g-i-orno
corteza  -c-o-r-tez-=-a-
cortex  -c-o-r-tex-=-ex
cortica  -c-o-r-tica-=-a-
cortecia  -c-o-r-teccia-=-a-
sorza   sc-o-r-za-=-a-

USES & APPROACHES

Alignment of two words useful for:
• String similarity (Mackay and Kondrak, 2005)
• Dialect distances (Nerbonne and Heeringa, 1997)
• Cognate identification (Mackay and Kondrak, 2005)
• Comparative reconstruction (Covington, 1996)

Multiple alignment gets us:
• String similarity vs. multiple words
• Better-informed cognate identification
• Better-informed comparative reconstruction (Covington, 1998)
• Sentence-level paraphrasing (Barsilay and Lee, 2003)

How to do it?
• One way: hand-crafted scales of similarity phoneme classes (Covington, 1998)
• Iterative pairwise
• Copy the computational biologists! (Durbin et al., 1998)

REFERENCES

Bhargava, Dyen, Data Corpus (distributions) and Lee, and Wilbert Profile HMM 93.20%
Minimum Levenshtein Distance 30%
Average Levenshtein Distance

RESULTS

Phoneme classes (Covington, 1998)

USES & APPROACHES

Cognate set matching:
• Build model from candidate sets
• Score word to sets using forward algorithm
• Choose set with highest score

Smoothing:
• Substitution matrix
• Added during Baum-Welch

EXPERIMENTS

Data:
• Comparative Indo-European Data Corpus (Dyen et al., 1992)
• Cognate data for words in 95 languages corresponding to 200 languages
• English orthography

Multiple alignment:
• Initialize a model (e.g. sample parameters from Dirichlet distributions)
• Train model to words using Baum-Welch
• Align words to model using Viterbi

CONCLUSIONS

• Profile HMMs can work for word-related tasks
• Multiple alignments are reasonable
• Cognate set matching performance exceeds that of average and minimum Levenshtein distance
• If multiple words need to be considered, Profile HMMs present a viable method

FUTURE WORK

• Model construction from aligned sequences: e.g. maximum a posteriori model construction
• Initial models for unaligned sequences: more informed, decrease guesswork
• Smoothing methods
• N-gram output symbols

REFERENCES