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       →    dz-e--n-
dzien           dzie--n-
giorno          g--iorno

corteza         -c-o-rtez--a-
cortex          -c-o-rt----ex
cortica    →    -c-o-rtic--a-
corteccia       -c-o-rteccia-
scorza          sc-o-r--z--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 (Barzilay 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)

MMIIIM

```
AG...C
A-AG.C
AG.AA-
--AAAC
AG...C
```

EXPERIMENTS

Data:

- Comparative Indoeuropean Data Corpus (Dyen et al., 1992)
- cognation 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

Cognate set matching:

- Build model from candidate sets
- Score word to sets using forward algorithm
- Choose set with highest score

Smoothing:

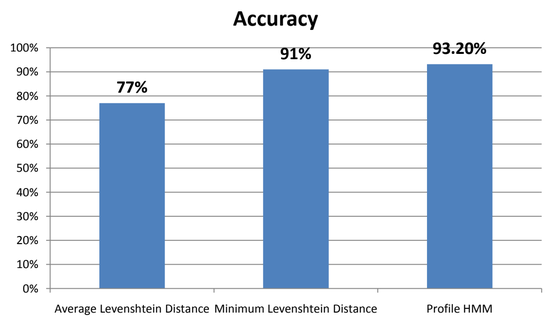
- Substitution matrix
- Added during Baum-Welch

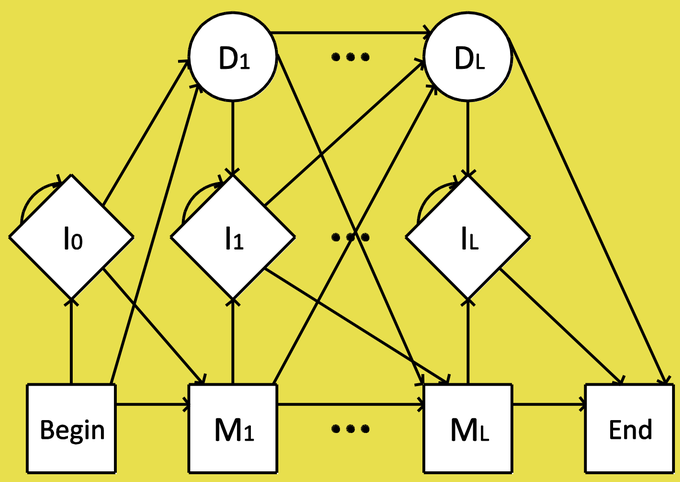
RESULTS

IIMMIMIIIMIIMI	**IMMIMIMIIMIIMI**
--MA-R---N--U-	-SCHO-R-----S-
U-M--R------IT	--C-O-RTEZ--A-
--MU-R------I-	--C-O-RT----EX
--ME-R------AV	--C-O-RTIC--A-
--MI-R----T-I-	--C-O-RTECCIA-
--MI-R----T-I-	-SC-O-R--Z--A-
--MI-R------ST	E-C-O-R--C--E-
--M--R------AM	-SC-OART----A-
--ME-R---C--I-	ISC-O-R--Z--A-
ZEM--R---IT-I-	ESC-O-R--X---Z
--M--R---IT-I-	--KRO----Z--U-
--M--R------ES	ISK-O-RT-H--A-
--M--R------EC	-SK-O-R-----A-
U-M--R------AM	--K-A-R-----A-
U-MIER------AC	--K-O-R-----A-
U-M--R---Z--EC	--K-U-R-----A-
U-MI-R------AT	
--ME-R------ET	
U-M--R---ET-I-	
--M--R---ET-I-	
U-MIER------AT	
--M--R------ET	
U-M--R------T-	
--M--R---ET-I-	
--ME-R----T-Y-	
--M---------O-	
--M---------O-	
--MORR---ER-E-	
--MOUR------IR	
--MO-R------I-	
--MOUR------I-	
--MO-R------IR	
--MO-R---IRSE-	
--MO-R---IR-E-	
--MORR------I-	
--MORR---ER-E-	
S-MU-R------IR	
--MU-R------I-	
--M--R------EL	
--MU-R---TH-A-	
--ME-R------I-	
--MAEL------YN	
--ME-R------EN	
--MA-R---N--U-	
--MA-R------UN	
--MA-RENAW--A-	
--ME-R---V--EL	
--MA-RU--E--IN	
--ME-R---V--EL	
--MA-R------W-	
--MA-R------W-	
--ME-R---W--U-	
U-M--R------E-	
--MU-R----D-AN	
--ME-R---N--A-	
--ME-R---N--A-	
--MO-R------A-	
--ME-R---N--E-	
--MO-R----D-AN	
--ME-R---N--EL	
U-M--R------E-	
U-MY-R---AT-Y-	
PAMI-R------AC	
--MORR------ER	
--MORR------ER	
--MA-R---N--IL	
--M--R------EL	

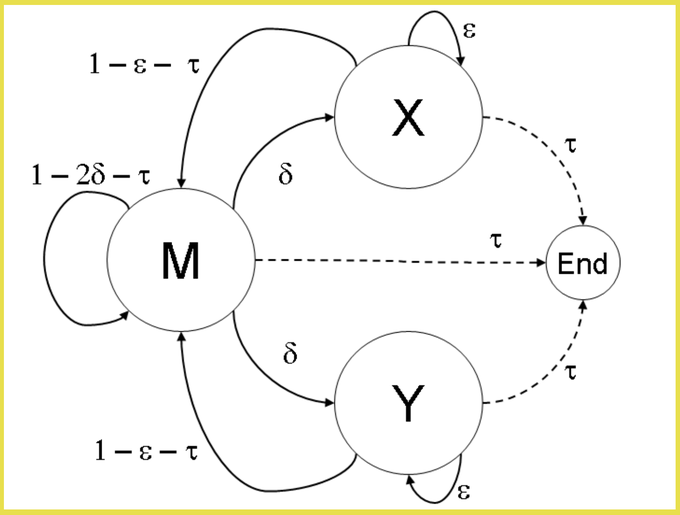
MIIMIIMI

```
D--E--N-
D--E--NY
D--E--N-
D--E--N-
Z--E--N-
DZ-E--N-
DZIE--N-
D--E--N-
D--A--N-
D--A--N-
DI-E--NA
D--E--IZ
D--E----
D--Y--DD
D--I--A-
D--I--E-
D-----I-
D-----I-
Z-----I-
Z--U--E-
Z-----U-
J--O--UR
DJ-O--U-
J--O--UR
G--IORNO
```





A prototypical Profile HMM of length *L*.



…contrast to Pair HMMs, which have been used for word similarity & cognate identification (Mackay and Kondrak, 2005).

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 consdered, 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

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple sequence alignment. In *Proc. of NAACL-HLT*, pages 16–23.

Michael A. Covington. 1996. An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481–496.

Michael A. Covington. 1998. Alignment of multiple languages for historical comparison. In *Proc. of COLING-ACL*, pages 275–279.

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5). Wesley Mackay and Grzegorz Kondrak. 2005. Computing word similarity and identifying cognates with pair hidden Markov models. In *Proc. of CoNLL*, pages 40–47.

John Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Proc. of the Third Meeting of ACL SIGPHON*.