Climbing the Tower of Babel: Unsupervised Multilingual Learning

Presented By David Erdos
January 23, 2011

Machine Learning for Natural Language Processing (048716)
Faculty of Electrical Engineering
Technion

Reference

Snyder B, Barzilay R. Climbing the Tower of Babel: Unsupervised Multilingual Learning. In: Joachims JFAT, ed. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Haifa, Israel: Omnipress; 2010:29-36. Available at: http://www.icml2010.org/papers/905.pdf.

Overview

Overview

- Electronic text is being produced at a vast and unprecedented scale all over the world
- Most languages are currently beyond the reach of NLP due to several factors
- Languages exhibit significant variation in the underlying linguistic structures

Overview

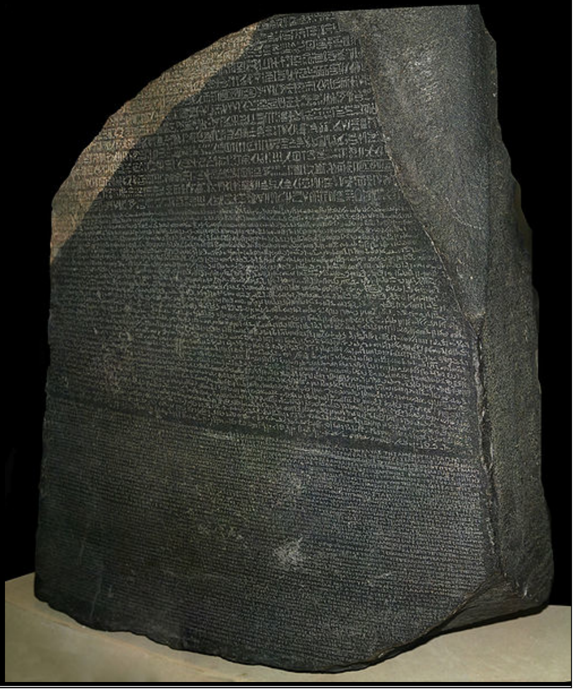
- This diversity in structure of languages can be harnessed to our advantage
- The authors utilize what is referred to as a *multilingual learning* framework
- This framework is based on the hypothesis that cross-lingual variations in linguistic structure correspond to *variations in ambiguity*

Variations in Ambiguity

- “I ate pasta with cheese”
  - Was pasta eaten with a cheese based utensil?
  - Or, was pasta eaten that had cheese on it?
- “What can he do?” - “מה הוא יכול לעשות”

Overview

- One of the goals was scalability in languages
- Unsupervised multilingual learning applied to the following tasks:
  - Morphological segmentation
  - Part-of-speech tagging
  - Parsing



Part-of-Speech Tagging

Part-of-speech Tagging

- Automatically determine the part-of-speech (noun, verb, adjective, etc.) of each word in the given context of a sentence
- A word with ambiguity in one language may correspond to an unambiguous word in another language

The Model

- A separate HMM is used for each language
- An additional layer of cross lingual variables (*superlingual tag*) is added
- Standard HMM joint-probability:

$$P(\mathbf{w}, \mathbf{y}) = \prod_i P(y_i | y_{i-1}) P(w_i | y_i)$$

The Model

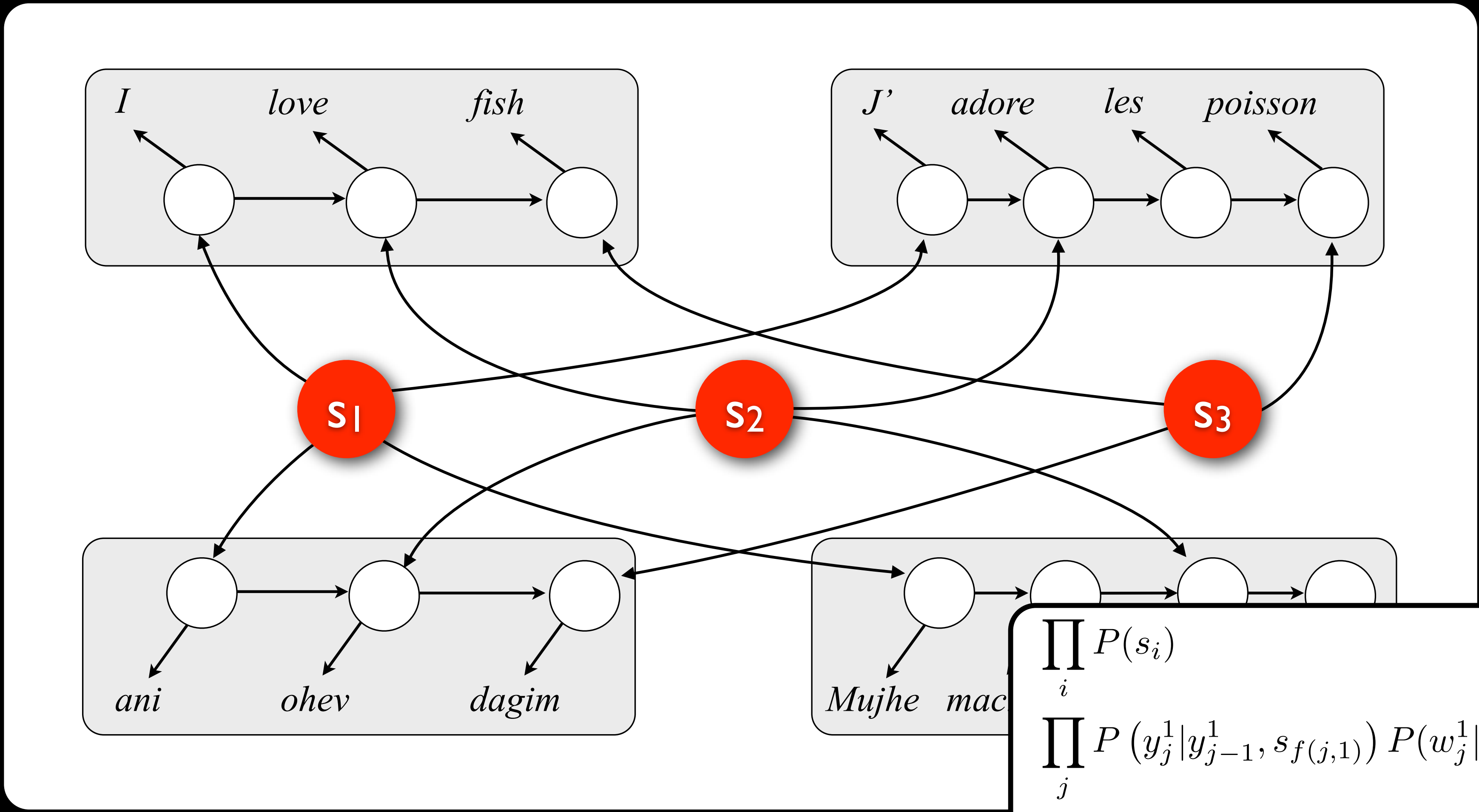
- Latent (hidden) variable model: the probability of bilingual parallel sentences ***(w^1, w^2)***, bilingual part-of-speech sequences ***(y^1, y^2)*** and superlingual tags **s** is given by:

$$\prod_i P(s_i)$$

$$\prod_j P\left(y_j^1|y_{j-1}^1, s_{f(j,1)}\right) P(w_j^1|y_j^1)$$

$$\prod_k P\left(y_k^2|y_{k-1}^2, s_{f(k,2)}\right) P(w_k^2|y_k^2)$$

The Model



$$\prod_i P(s_i)$$

$$\prod_j P\left(y^1_j | y^1_{j-1}, s_{f(j,1)}\right) P(w^1_j | y^1_j)$$

$$\prod_k P\left(y^2_k | y^2_{k-1}, s_{f(k,2)}\right) P(w^2_k | y^2_k)$$

Superlingual Tags

- Formally, each superlingual value provides a set of multinomial probability distributions, one for each language's part-of-speech inventory

Superlingual value "2"

	Noun	Verb	Determiner
English	0.9	0.1	0.0
French	0.8	0.1	0.1
Hindi	1.0	0.0	0.0

Superlingual value "5"

	Noun	Verb	Determiner
English	0.5	0.4	0.1
French	0.4	0.6	0.0
Hindi	0.5	0.5	0.0

Superlingual Tags

- The number of superlingual values is left unbounded
- To encourage sparse cross-lingual regularities a Dirichlet process prior is used
- The actual number of superlingual values is dictated by the data (11 for a pair of languages, 17 for eight languages)

Evaluation

- The model is evaluated on a parallel corpus of eight languages
- Inference performed using Markov Chain Monte Carlo sampling
- Test is performed on held out monolingual data for each language

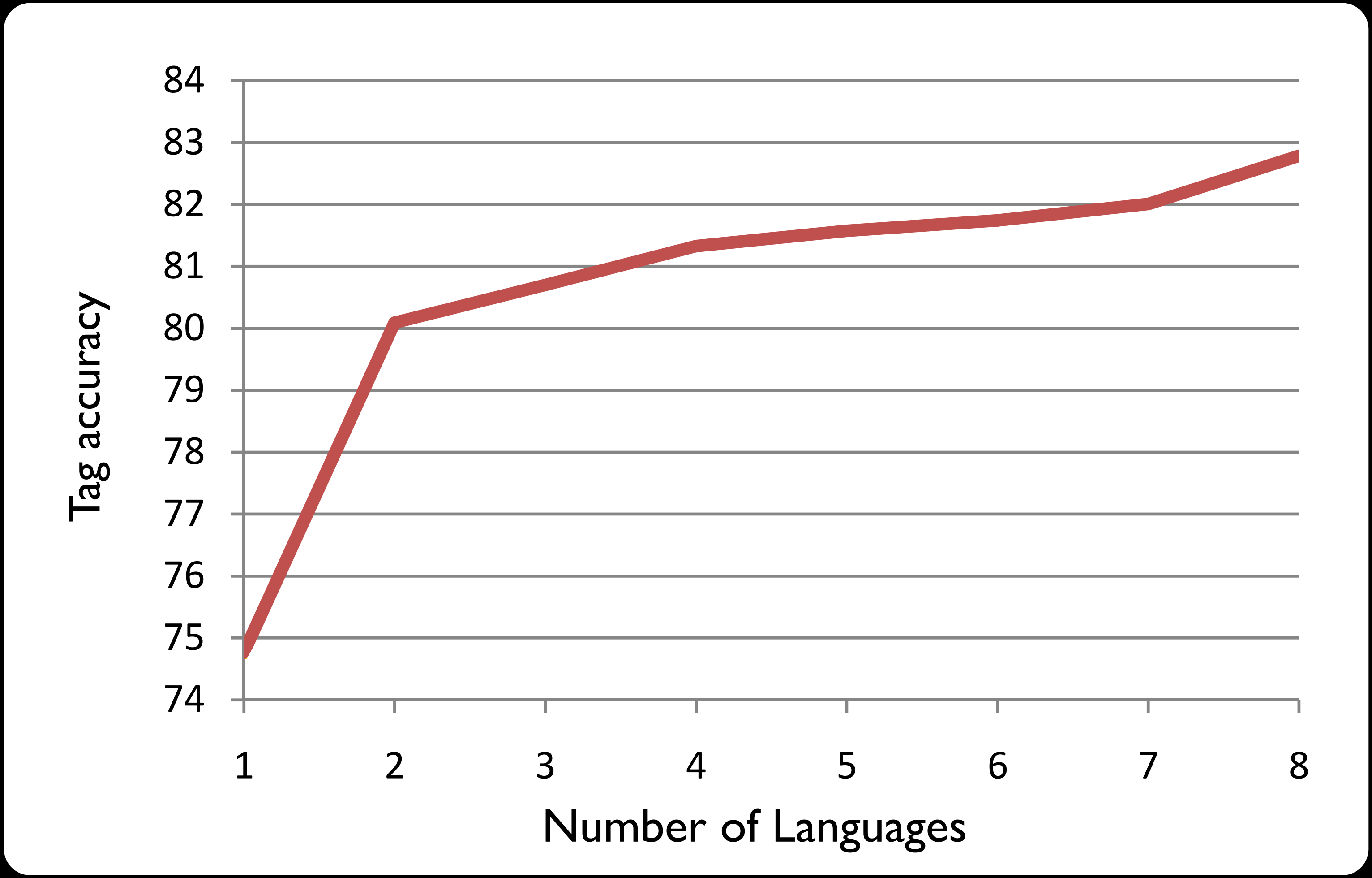
Evaluation

- The algorithm was run over all of the 255 subsets of the eight languages in the corpus
- The average change in performance as the number of languages increases was examined
- In the monolingual scenario, the model reduces to a Bayesian HMM (Goldwater & Griffiths, 2007)

Results

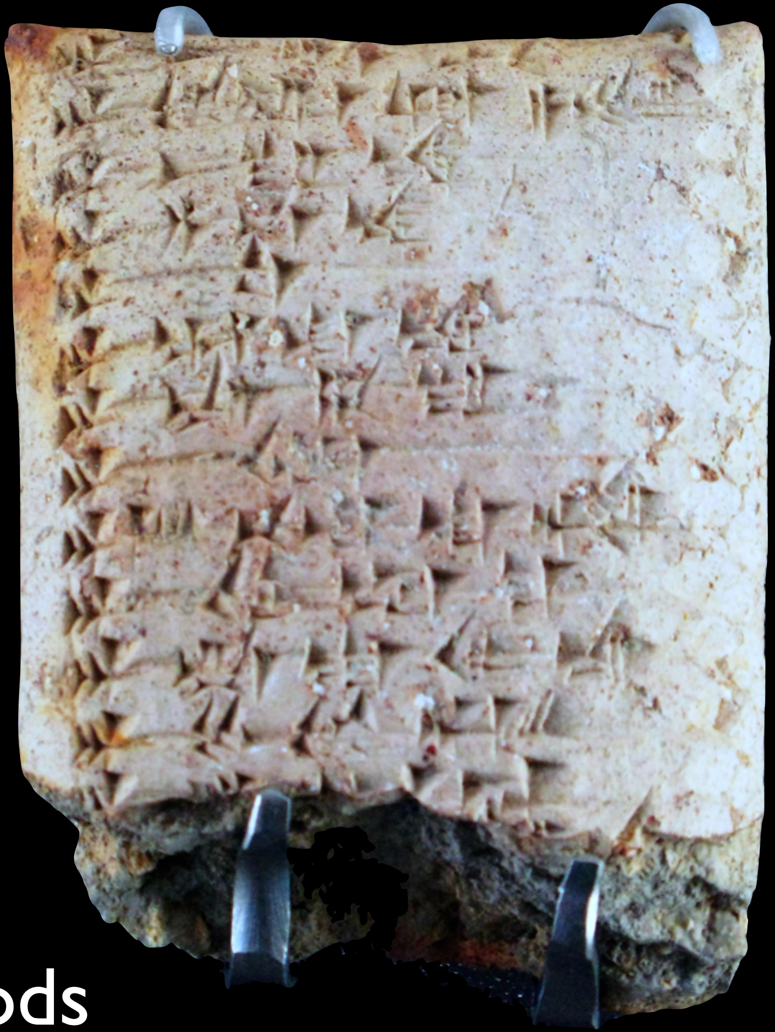
- With complete part-of-speech dictionary:
  - 91.1% average accuracy (monolingual)
  - 95% accuracy (multilingual)
- With partial part-of-speech dictionary:
  - 74.8% accuracy (monolingual)
  - 82.8% accuracy (multilingual)

Tag Accuracy



Lost Language Decipherment

Ugaritic



List of Ugaritic gods

13th Century BC

Lost Languages

- Previous work relies on the availability of parallel texts
- No parallel texts are available with lost languages
- Instead, this method relies on knowledge of similar languages

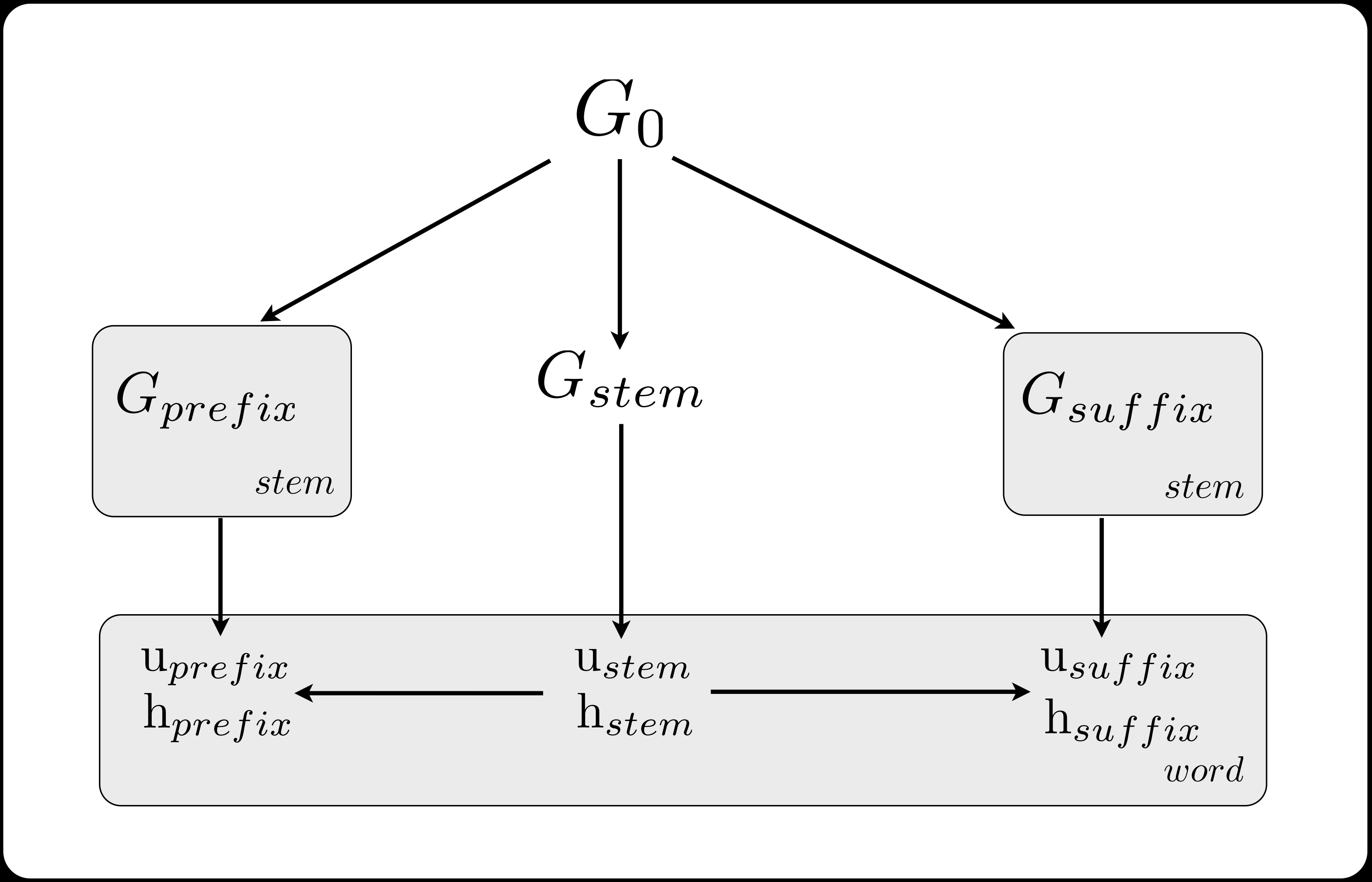
The Method

- The input consists of texts in a lost language, and corpus of non-parallel data in a known related language
- Common manual methods involve studying word and letter frequency
- Morphological analysis plays a key part in the process, frequent suffix/prefix occurances can be particularly helpful

The Method

- These intuitions are captured as a generative Bayesian model
- The model caries out implicit morphological analysis of the lost language utilizing the known morphological structure of the related language

Decipherment Model



Results

- Decipherment model applied to a corpus of Ugaritic text with 7,386 unique word forms
- A Hebrew lexicon is also used, which was extracted from the Hebrew Tanakh
- The model yields almost perfect decipherment of the alphabetic symbols
- Over half of the Ugaritic word forms with cognates in Hebrew were correctly identified

Hebrew - Ugaritic

	א	ב	ג	ד	ה	ו	ז	ח	ט	י	כ
𐎀	X										
𐎁		X									
𐎂			X								
𐎃								X			
𐎄				X			X				
𐎅					X						
𐎆						X					
𐎇							X				
𐎈								X			
𐎉									X		
𐎊										X	
𐎋											X

Conclusion

Conclusion

- Authors applied multilingual learning to traditional NLP tasks, with unannotated parallel texts
- Multilingual language models performed better than their monolingual counterparts
- This is a realistic scenario for many of the world’s languages

Questions?