Multilingual Part-of-Speech Tagging: Two Unsupervised Approaches

Naseem, T.; Snyder, B.; Eisenstein, J. & Barzilay, R. (2009)

Presenter: Chris Cervantes

Outline

- Conceptual Background
- Formal Descriptions
- Experiments

Outline

- Conceptual Background
  - Monolingual POS Tagging
  - The Role of Additional Languages
  - Overview of Approaches
    - Merged Node Model
    - Latent Variable Model
- Formal Descriptions
- Experiments

Monolingual POS Tagging

- Unsupervised monolingual part-of-speech (POS) tagging assigns tags to words, where tags are learned from unlabeled text
  - Tags are treated as a linear sequence of hidden variables and words as emitted observations
  - Often represented as a Hidden Markov Model (HMM)
- Necessary components for HMM POS tagger
  - Initial and final states
  - Transition probabilities
  - Emission probabilities
  - Initial state distributions
    - These probabilities can also be expressed as transition probabilities from a start-of-sentence tag to all the other tags

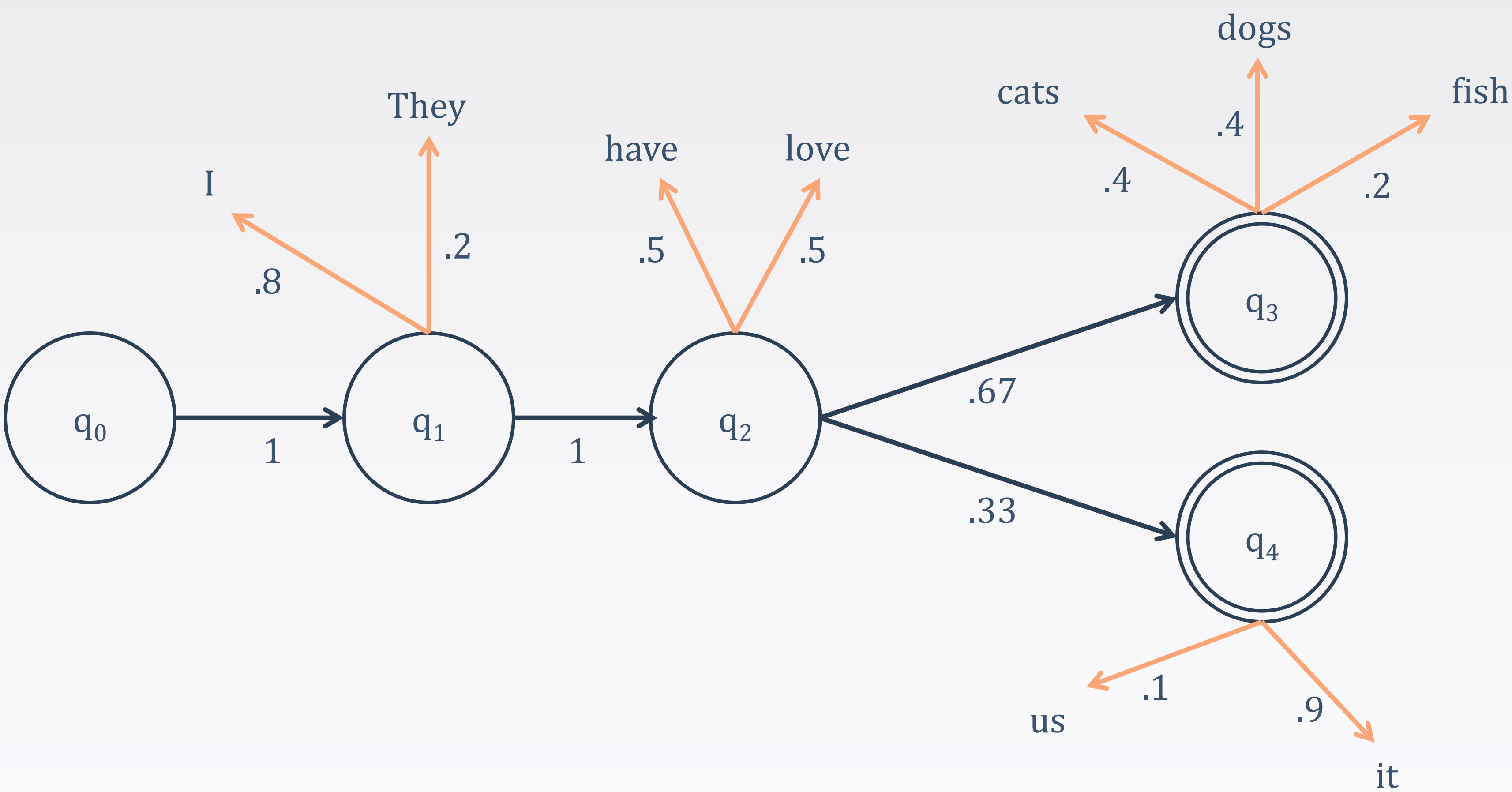
Monolingual POS Tagging

- Unsupervised monolingual part-of-speech (POS) tagging assigns tags to words, where tags are learned from unlabeled text
  - Tags are treated as a linear sequence of hidden variables and words as emitted observations
  - Often represented as a Hidden Markov Model (HMM)
- Necessary components for HMM POS tagger
  - Initial and final states
  - Transition probabilities
  - Emission probabilities
  - Initial state distributions
    - These probabilities can also be expressed as transition probabilities from a start-of-sentence tag to all the other tags

In the Bayesian model, these distributions are drawn from priors

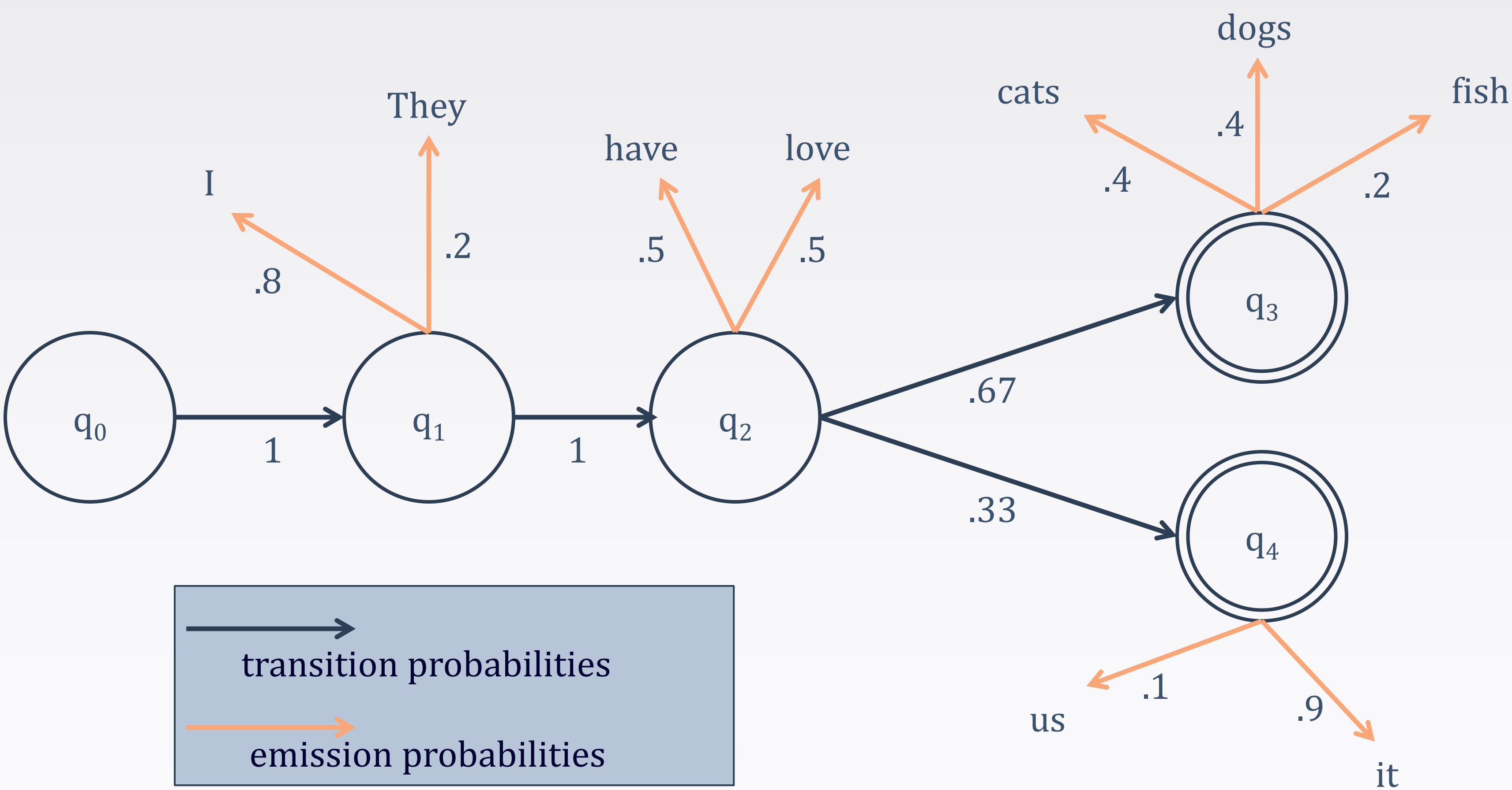
Monolingual POS Tagging

- HMM POS tagger example



Monolingual POS Tagging

- HMM POS tagger example



Role of Additional Languages

- Languages have different patterns of ambiguity
  - Words with POS ambiguity
    - "can" in English might be a standard verb, auxiliary verb, or noun
  - Structural ambiguity
    - articles in English reduce next-POS possibilities
- Different ambiguity patterns are very likely to occur in different places / for different reasons across languages
  - Unannotated multilingual data serves as a learning signal in an unsupervised system
  - **Key Idea: combining information from multiple languages creates a clearer picture of each**

Overview of Approaches

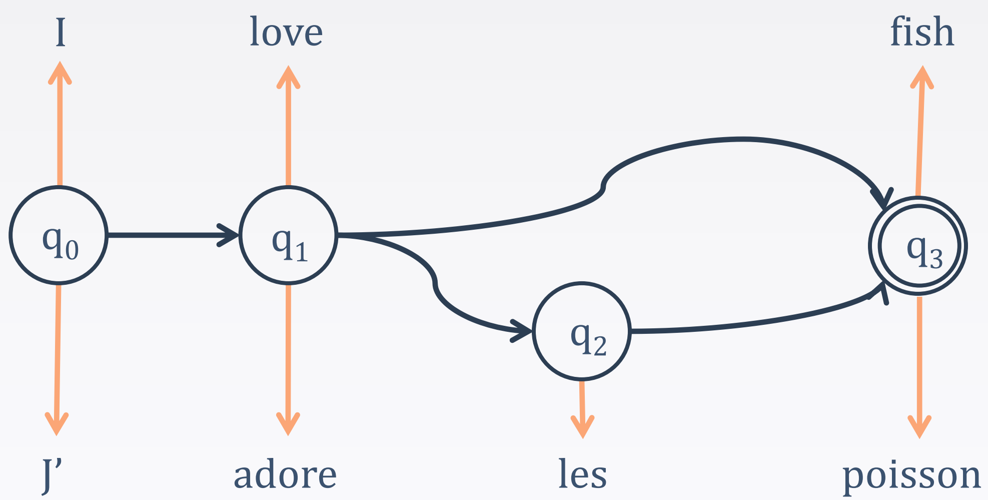
- Observed data
  - Corpus of parallel sentences in multiple languages
  - Word alignments between parallel sentence pairs are given via a black box mechanism and so are treated as observed
- Tags are drawn from tag dictionaries
  - Not completely unsupervised
- Two approaches
  - Merged Node Model
  - Latent Variable Model

Merged Node Model

- Model relies on language pairs
- HMM nodes are created by merging tag nodes from different languages
    - Nodes represent a pair of tags, one per language
- Each node emits two words, one per language

Merged Node Model

- Model relies on language pairs
- HMM nodes are created by merging tag nodes from different languages
  - Nodes represent a pair of tags, one per language
- Each node emits two words, one per language

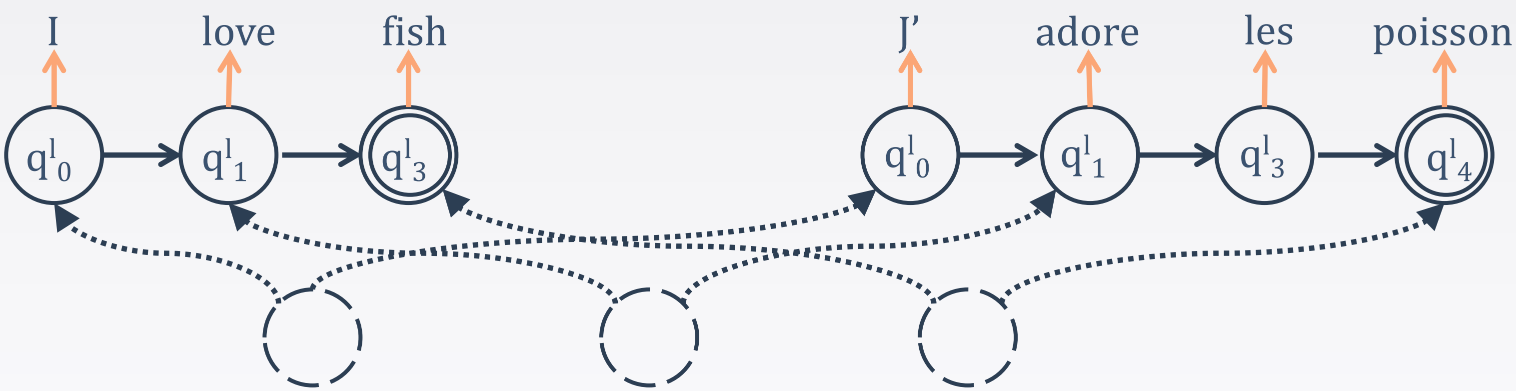


Latent Variable Model

- Operates over any number of languages with parallel text
- Like in the monolingual model, HMM nodes represent single tags and emit single words
- Assumes an additional layer of superlingual tags that inform which node to transition to

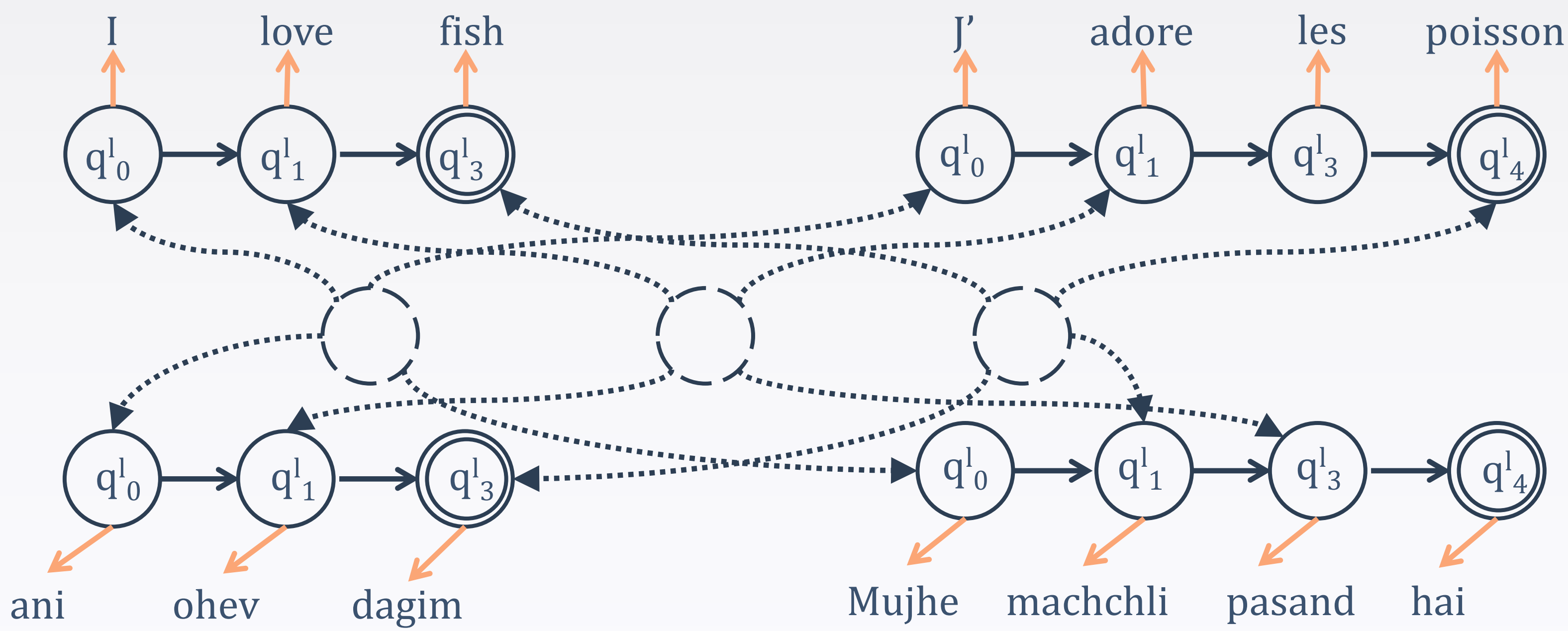
Latent Variable Model

- Operates over any number of languages with parallel text
- Like in the monolingual model, HMM nodes represent single tags and emit single words
- Assumes an additional layer of superlingual tags that inform which node to transition to



Latent Variable Model

- Operates over any number of languages with parallel text
- Like in the monolingual model, HMM nodes represent single tags and emit single words
- Assumes an additional layer of superlingual tags that inform which node to transition to



Outline

- Conceptual Background
- Formal Descriptions
    - Merged Node Model
    - Latent Variable Model
- Experiments

Merged Node Model

- Terms
  - **T / T'** : Tag set for respective languages
  - **t / t'** : Individual tag for respective languages
  - **< t, t' >** : A tag pair (one tag from each language). Tag pairs are the nodes in this HMM
    - <t, t'> ϵ T x T'
  - **ω** : Coupling distribution, which informs how the tags are merged into pairs
  - **< y_i, y_j' >** : Aligned tag pair. Where <t,t'> is a tag pair from the set of any two tags (one per language), <y_i, y_j'> is aligned between the two languages
    - <y_i, y_j'> is conditioned on y_{i-1}, y'_{j-1}, and the coupling parameter $\omega(y_i, y_j')$
  - **W / W'** : Vocabulary for respective languages

Merged Node Model

- Generative story
  - Transition / Emission Parameters
  - Coupling Parameter
  - Data

Merged Node Model

- Generative story
  - Transition / Emission Parameters
    - For each $t \in T$
      - Draw a transition distribution ϕ_t over tags T
      - Draw an emission distribution Θ_t over words W
    - For each $t' \in T'$
      - Draw a transition distribution ϕ_t' over tags T'
      - Draw an emission distribution Θ_t' over words W'
  - Coupling Parameter
  - Data

Merged Node Approach

multinomials, each drawn from a symmetric Dirichlet prior

- Generative story
  - Transition / Emission Parameters
    - For each $t \in T$
      - Draw a transition distribution ϕ_t over tags T
      - Draw an emission distribution θ_t over words W
    - For each $t' \in T'$
      - Draw a transition distribution ϕ_t' over tags T'
      - Draw an emission distribution θ_t' over words W'
  - Coupling Parameter
  - Data

Merged Node Model

- Generative story
  - Transition / Emission Parameters
  - Coupling Parameter
    - Draw a bilingual coupling distribution, ω, over tag pairs TxT'
  - Data

Merged Node Model

- Generative story
  - Transition / Emission Parameters
  - Coupling Parameter
    - Draw a bilingual coupling distribution, ω, over tag pairs TxT'
  - Data

also multinomial, drawn from symmetric Dirichlet prior ω_0

Merged Node Model

- Generative story
  - Transition / Emission Parameters
  - Coupling Parameter
  - Data
    - For each parallel sentence
      - Draw alignment a, a set of integer pairs (i,j) indicating aligned indices in parallel sentences.
      - Draw a bilingual POS tag sequence, $(y_1, \ldots, y_m), (y_1', \ldots, y_n')$
      - For each POS tag y_i, emit a word $x_i \sim \theta_{y_i}$
      - For each POS tag y_j', emit a word $x_j' \sim \theta_{y'_j}$

Merged Node Model

- Generative story
  - Transition / Emission Parameters
  - Coupling Parameter
  - Data
    - For each parallel sentence
      - Draw alignment a, a set of integer pairs (i,j) indicating aligned indices in parallel sentences.
      - Draw a bilingual POS tag sequence, $(y_1, \ldots, y_m), (y_1', \ldots, y_n')$
      - For each POS tag y_i, emit a word $x_i \sim \Theta_{y_i}$
      - For each POS tag y_j', emit a word $x_j' \sim \Theta_{y'_j}$

$a \sim A_0$, a prior distribution over alignments provided by their black box mechanism

Merged Node Model

- Generative story
  - Transition / Emission Parameters
  - Coupling Parameter
  - Data
    - For each parallel sentence
      - Draw alignment a, a set of integer pairs (i,j) indicating aligned indices in parallel sentences.
      - Draw a bilingual POS tag sequence, $(y_1, \ldots, y_m), (y_1', \ldots, y_n')$
      - For each POS tag y_i, emit a word $x_i \sim \Theta_{y_i}$
      - For each POS tag y_j', emit a word $x_j' \sim \Theta_{y'_j}$

$a \sim A_0$, a prior distribution over alignments provided by their black box mechanism

POS tag sequences are drawn according to:

$$P((y_1, \ldots, y_m), (y_1', \ldots, y_n') \mid a, \phi_t, \phi_t', \omega) =$$

$$\prod_{\text{unaligned } i} \phi_{y_{i-1}}(y_i) \prod_{\text{unaligned } j} \phi'_{y'_{j-1}}(y'_j)$$

$$\prod_{(i,j) \in a} \left(\frac{\phi_{y_{i-1}}(y_i)\, \phi'_{y'_{j-1}}(y'_j)\, \omega(y_i, y'_j)}{\sum_{y,y'} (\phi_{y_{i-1}}(y)\, \phi'_{y'_{j-1}}(y')\, \omega(y, y')} \right)$$

Merged Node Model

- Generative story
  - Transition / Emission Parameters
  - Coupling Parameter
  - Data
    - For each parallel sentence
      - Draw alignment a, a set of integer pairs (i,j) indicating aligned indices in parallel sentences.
      - Draw a bilingual POS tag sequence, $(y_1, \ldots, y_m), (y_1', \ldots, y_n')$
      - For each POS tag y_i, emit a word $x_i \sim \Theta_{y_i}$
      - For each POS tag y_j', emit a word $x_j' \sim \Theta_{y'_j}$

$a \sim A_0$, a prior distribution over alignments provided by their black box mechanism

x_i and x_j' are words from W and W', respectively

POS tag sequences are drawn according to:

$P((y_1, \ldots, y_m), (y_1', \ldots, y_n') \mid a, \phi_t, \phi_t', \omega) =$

$$\prod_{\text{unaligned } i} \phi_{y_{i-1}}(y_i) \prod_{\text{unaligned } j} \phi'_{y'_{j-1}}(y'_j)$$

$$\prod_{(i,j) \in a} \left(\frac{\phi_{y_{i-1}}(y_i)\, \phi'_{y'_{j-1}}(y'_j)\, \omega(y_i, y'_j)}{\sum_{y,y'} (\phi_{y_{i-1}}(y)\, \phi'_{y'_{j-1}}(y')\, \omega(y, y')} \right)$$

Merged Node Model

- Inference
  - Process occurs in a monolingual setting (and thus must be performed for each language in the pair)
  - Ideal transition and emission parameters

$$\hat{\theta}, \hat{\phi} = \underset{\theta,\phi}{\operatorname{argmax}} \int P(\theta, \phi, y, \omega \mid x, a, \theta_0, \phi_0, \omega_0)\, dy\, d\omega$$

  - Actual parameters are found with Gibbs sampling
    - θ, ϕ, and ω are all marginalized out
    - Only POS tags and priors are sampled
  - After sampling, parameters θ and ϕ are the maximum a posteriori estimates

Latent Variable Model

- Assumes an additional layer of superlingual tags
- Operates over any number of languages with parallel text
- Offers both a conceptual and a computational benefit over using the merged node model with more languages
  - Multilingual information can reduce linguistic ambiguity *during training*; combining bilingually trained models (like the merged node model) doesn't take advantage of this
  - State space in the merged node model grows exponentially with the number of languages, L
    - Since nodes are tag pairs, the size of the state space is $|T|^L$
    - ω has the same dimension.

Latent Variable Model

- Parameter generation
  - Draw an infinite sequence of distribution sets
    - $\Psi_1, \Psi_2, \ldots \sim G_0$
    - Ψ_i : a set of distributions over tags,
      one distribution per language l
      $(\varphi_i^{l}, \varphi_i^{l'}, \ldots)$
  - Draw an infinite sequence of mixture weights
    - $\pi_1, \pi_2, \ldots \sim GEM(\alpha)$
    - These mixture weights weight the sets of distributions, above

Latent Variable Model

- Parameter generation
  - Draw an infinite sequence of distribution sets
    - $\Psi_1, \Psi_2, \ldots \sim G_0$
    - Ψ_i : a set of distributions over tags, one distribution per language l (φ_i^{l}, $\varphi_i^{l'}$, …)
  - Draw an infinite sequence of mixture weights
    - $\pi_1, \pi_2, \ldots \sim \mathrm{GEM}(\alpha)$
    - These mixture weights weight the sets of distributions, above

G_0 is some base distribution

Latent Variable Model

- Parameter generation
  - Draw an infinite sequence of distribution sets
    - $\Psi_1, \Psi_2, \ldots \sim G_0$
    - Ψ_i : a set of distributions over tags, one distribution per language l (φ_i^{l}, $\varphi_i^{l'}$, …)
  - Draw an infinite sequence of mixture weights
    - $\pi_1, \pi_2, \ldots \sim \mathrm{GEM}(\alpha)$
    - These mixture weights weight the sets of distributions, above

G_0 is some base distribution

GEM(α) is a stick-breaking process

Latent Variable Model

- Given these parameters…
  - Superlingual tag z is drawn such that
    - z is drawn with probability π_z
    - z is an index of the infinite sequence of sets of multinomials $(\varphi_z^{l}, \varphi_z^{l'}, \ldots)$
  - POS tag y_i is drawn according to
    - $y_i \sim \dfrac{\phi_{y_{i-1}}(y_i) \prod_{m=1}^{M} \varphi^{l}_{z_m}(y_i)}{Z}$
    - i : Tag position
    - l : Language
    - $\phi_{y_{i-1}}(y_i)$: Transition distribution from the previous tag to this tag
    - z_m : Value of the m^{th} connected superlingual tag
    - $\varphi^{l}_{z_m}(y_i)$: Tag distribution for language l given by Ψ_{z_m}
    - Z : Sum of the product in the numerator over all values for y_i
    - M : All superlingual tag indices with which position l is associated

Latent Variable Model

- Given these parameters…
  - Superlingual tag z is drawn such that
    - z is drawn with probability π_z
    - z is an index of the infinite sequence of sets of multinomials $(\varphi_z^{l}, \varphi_z^{l'}, \ldots)$
  - POS tag y_i is drawn according to
    - $y_i \sim \dfrac{\phi_{y_{i-1}}(y_i) \prod_{m=1}^{M} \varphi^{l}_{z_m}(y_i)}{Z}$
    - i : Tag position
    - l : Language
    - $\phi_{y_{i-1}}(y_i)$: Transition distribution from the previous tag to this tag
    - z_m : Value of the m^{th} connected superlingual tag
    - $\varphi^{l}_{z_m}(y_i)$: Tag distribution for language l given by Ψ_{z_m}
    - Z : Sum of the product in the numerator over all values for y_i
    - M : All superlingual tag indices with which position l is associated

A beneficial consequence of drawing tags in this way is that a high probability tag at a given position must be allowed for by each incoming distribution

Latent Variable Model

- Generative story
  - Transition / Emission Parameters
  - Superlingual Parameters
  - Data

Latent Variable Model

- Generative story
    - Transition / Emission Parameters
        - For each language $l = 1, \ldots, n$ and for each tag $t \in T^l$
            - Draw a transition distribution, ϕ^l_t , over tags T^l
            - Draw an emission distribution Θ^l_t , over words W^l
    - Superlingual Parameters
    - Data

Latent Variable Model

- Generative story
  - Transition / Emission Parameters
    - For each language $l = 1, \ldots, n$ and for each tag $t \in T^l$
      - Draw a transition distribution, ϕ^l_t, over tags T^l
      - Draw an emission distribution Θ^l_t, over words W^l
  - Superlingual Parameters
  - Data

multinomials, each drawn from a symmetric Dirichlet prior

Latent Variable Model

- Generative story
  - Transition / Emission Parameters
  - Superlingual Parameters
    - Draw an infinite sequence of distribution sets
      - $\Psi_1, \Psi_2, \ldots \sim G_0$
      - Ψ_i : a set of distributions over tags, one distribution per language l (φ_i^l)
    - Draw an infinite sequence of mixture weights
      - $\pi_1, \pi_2, \ldots \sim GEM(\alpha)$
      - These mixture weights weight the sets of distributions, above
  - Data

Latent Variable Model

- Generative story
  - Transition / Emission Parameters
  - Superlingual Parameters
  - Data
    - For each multilingual parallel sentence
      - Draw alignment *a* from A_m
        - *a* is a set of aligned indices across languages $(i_1, i_2, \ldots, i_n)$
      - For each set of indices in *a*
        - Draw superlingual tag z
      - For each language, l, and for each position i
        - Draw y_i such that
          - $y_i \sim \dfrac{\phi_{y_{i-1}}(y_i) \prod_{m=1}^{M} \varphi^{l}_{z_m}(y_i)}{Z}$
        - Draw word $w_i \in W^l$ according to Θ_{y_i}

Latent Variable Model

- Inference
  - Like in the merged node model, a sampling technique is used for inference
    - θ, ϕ, φ^{l}_{i}, and π are all marginalized out
    - Only POS tags and superlingual tags need to be sampled
  - In order to integrate over π during superlingual tag sampling, the Chinese Restaurant Process is used

Outline

- Conceptual Background
- Formal Descriptions
- Experiments
  - Full Lexicon Experiment
  - Reduced Lexicon Experiment
  - Analysis

Experiments

- George Orwell's *1984* is used as the experiment data
  - Parallel text in English, Bulgarian, Czech, Estonian, Hungarian, Slovene, Serbian, and Romanian
  - Provided as part of the Multext-East corpus, which is annoted with POS tags and provides a lexicon for each language
- Word alignments are provided with a black box mechanism (GIZA++)
- For the sake of comparison, two other systems are implemented
  - A monolingual Bayesian HMM
  - A supervised HMM (trained with annotated data)
- Merged node model results (which are constrained by pairings) are combined in three ways
  - Average across pairings
  - Best-pair using an oracle
  - Voting scheme

Full Lexicon Experiment

- Assume the full tag lexicon – set of possible POS tags – is known in advance
- Possible tags per word is 1.39
- Tagging Accuracy

	Avg	BG	CS	EN	ET	HU	RO	SL	SR
1. Random	83.3	82.5	86.9	80.7	84.0	85.7	78.2	84.5	83.5
2. Monolingual	91.2	88.7	93.9	95.8	92.7	95.3	91.1	87.4	84.5
3. MERGEDNODE: *average*	**93.2**	91.3	96.9	95.9	93.3	96.7	91.9	89.3	90.2
4. LATENTVARIABLE	**95.0**	92.6	98.2	95.0	94.6	96.7	95.1	95.8	92.3
5. Supervised	97.3	96.8	98.6	97.2	97.0	97.8	97.7	97.0	96.6
6. MERGEDNODE: *voting*	93.0	91.6	97.4	96.1	94.3	96.8	91.6	87.9	88.2
7. MERGEDNODE: *best pair*	95.4	94.7	97.8	96.1	94.2	96.9	94.1	94.8	94.5

Reduced Lexicon Experiment

- Three types of reduced lexicons are used.
  - All words with less than 5 instances are removed
  - All words with less than 10 instances are removed
  - Only the top 100 words are retained in the lexicon
- Possible tags per word is 7.54 in the "Top 100" model
- Tagging Accuracy

		Avg	BG	CS	EN	ET	HU	RO	SL	SR
Counts > 5	Random	63.6	62.9	62	71.8	61.6	61.3	62.8	64.8	61.8
	Monolingual	74.8	73.5	72.2	87.3	72.5	73.5	77.1	75.7	66.3
	MERGEDNODE: *average*	80.1	80.2	79	90.4	76.5	77.3	82.7	78.7	75.9
	LATENTVARIABLE	**82.8**	81.3	**83.0**	88.1	**80.6**	**80.8**	**86.1**	**83.6**	78.8
	MERGEDNODE: *voting*	80.4	80.4	78.5	90.7	76.4	76.8	84.0	79.7	76.4
	MERGEDNODE: *best pair*	81.7	**82.7**	79.7	**90.7**	77.5	78	84.4	80.9	**79.4**
Counts > 10	Random	57.9	57.5	54.7	68.3	56	55.1	57.2	59.2	55.5
	Monolingual	70.9	71.9	66.7	84.4	68.3	69.0	73.0	70.4	63.7
	MERGEDNODE: *average*	77.2	77.8	75.3	88.8	72.9	73.8	80.5	76.1	72.4
	LATENTVARIABLE	**79.7**	78.8†	**79.4**	86.1	**77.9**	**76.4**	**83.1**	**80.0**	75.9
	MERGEDNODE: *voting*	77.5	78.4†	75.3	89.2	73.1	73.3	81.7	76.1	73.1
	MERGEDNODE: *best pair*	79.0	**80.2**	76.7	**89.4**	74.9	75.2	82.1	77.6	**76.1**
Top 100	Random	37.3	36.7	32.1	48.9	36.6	36.4	33.7	39.8	33.8
	Monolingual	53.8	60.9‡	44.1	69.0	54.8*	56.8	51.4	49.4	44.0
	MERGEDNODE: *average*	**59.6**	60.1	52.5	73.5	59.5	59.4	61.4	56.6	53.4
	LATENTVARIABLE	57.9	**65.5**	49.3	71.6	54.3*	51.0	57.5	53.9	**60.4**
	MERGEDNODE: *voting*	62.4	61.5‡	55.4	74.8	62.2	60.9	64.3	62.3	57.5
	MERGEDNODE: *best pair*	63.6	64.7	**55.3**	**77.4**	**61.5**	**60.2**	**69.3**	**63.1**	56.9

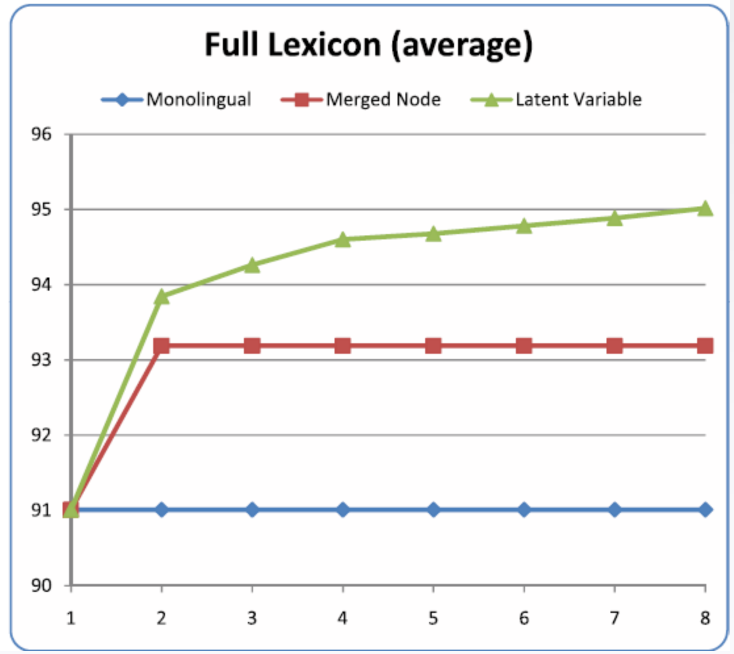
Analysis

- Performance would be helped if the optimal language partners could be predicted
  - Language relatedness isn't necessarily helpful
    - Slovene and Serbian are related and optimal partners
    - Bulgarian and English are optimal, but not closely related
  - Tag / word ambiguity is correlated negatively with a language's helpfulness as a partner

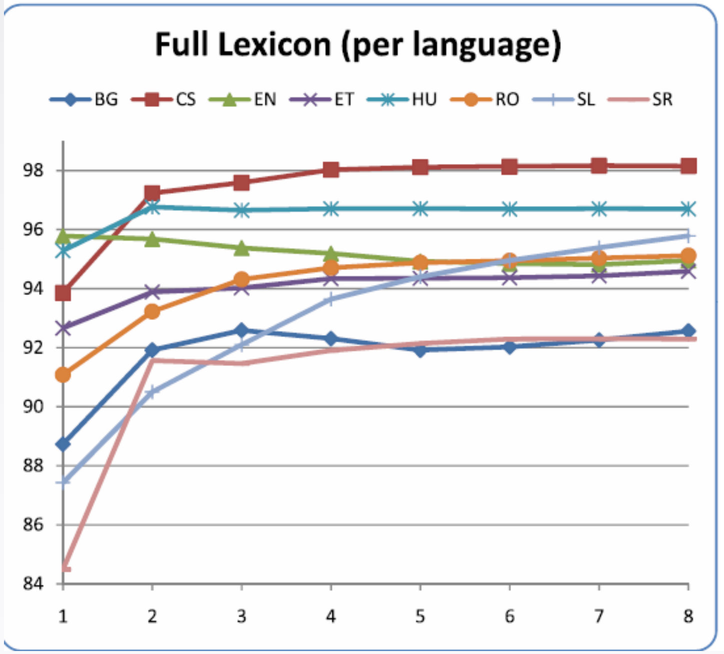
MergedNode Model									
		coupled with...							
accuracy for...	**Avg**	BG	CS	EN	ET	HU	RO	SL	SR
BG	91.3		90.2	**94.7**	92.3	90.6	91.2	91.1	88.7†
CS	96.9	95.3		97.5	**97.8**	96.3	96.4	97.4	97.4
EN	95.9	**96.1**	95.9†		95.8†	95.8†	95.8†	**96.1**	96.0
ET	93.3	93.0	94.0	92.9†		92.2†	93.0	**94.2**	93.9
HU	96.7	96.8	96.6	96.8	**96.9**		96.8	96.5	96.7
RO	91.9	**94.1**	90.6†	92.0	91.3	90.3†		91.3	93.9
SL	89.3	88.5	88.1	89.2	89.8	87.5†	87.5†		**94.8**
SR	90.2	88.5	88.2	**94.5**	94.2	89.5	85.0	91.4	

Analysis

Average performance as the number of languages increases



Average performance of the latent variable model of languages as the number of language increases



Analysis

- If the full lexicon is available, the two models proposed significantly improve on previous unsupervised methods
  - For most languages, performance is gained as more languages are added
- If only a reduced lexicon is available, the merged model is likely the better choice
- Performance varies greatly depending on which languages are chosen, but it’s difficult to determine what language is going to be helpful
  - This question is irrelevant in the latent variable model, since all languages are used