Crouching Dirichlet, Hidden Markov Model: Unsupervised POS Tagging with Context Local Tag Generation

Taesun Moon   Katrin Erk   and Jason Baldridge

Department of Linguistics
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA

Empirical Methods in Natural Language Processing 2010

Introduction

Unsupervised HMM POS tagging: Problems

The standard HMM is a very poor approximation of natural language

- Markov independence assumption is too strong
- Parameter configuration is too restrictive

Introduction

A simple dichotomy in natural language

- For many languages, words can generally be grouped into function words and content words
- There are few function words by type
- Individually, these function words appear relatively frequently
- There are many content word by type
- Individually, these content words appear relatively infrequently

Introduction

Some numbers from the Penn Treebank WSJ

Number of word types (Total tokens) per tag

- NN = 9321 (164K)
- JJ = 8591 (75K)
- DT = 24 (101K)
- CC = 22 (29K)

Conditional probability of most frequent word given tag

- $p(\text{company}|\text{NN}) = 0.02$
- $p(\text{other}|\text{JJ}) = 0.02$
- $p(\text{the}|\text{DT}) = 0.59$
- $p(\text{and}|\text{CC}) = 0.69$

Transition probabilities

- $p(\text{NN}|\text{JJ}) = 0.45$
- $p(\text{NNP}|\text{NNP}) = 0.38$
- $p(\text{DT}|\text{PDT}) = 0.91$
- $p(\text{VB}|\text{MD}) = 0.80$

Introduction

Some assumptions in IR

- Function words are almost always stopwords
- tf-idf is predicated on the difference in variance of words across documents
- LSI, LDA, etc. capture the variance of content words across documents

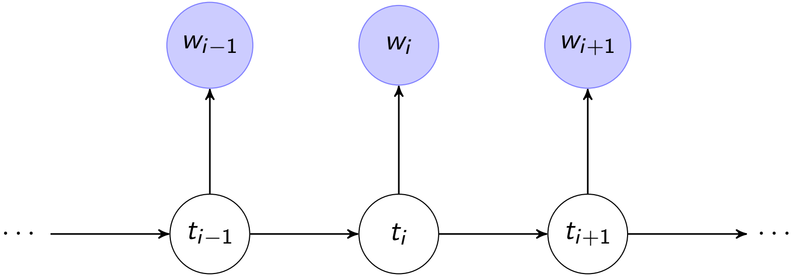
Introduction

Solutions

- Directly model statistical dichotomy between content and function words
- Capture possible variance of content words and tags across contexts
- Retain HMM framework and extend from there

Models

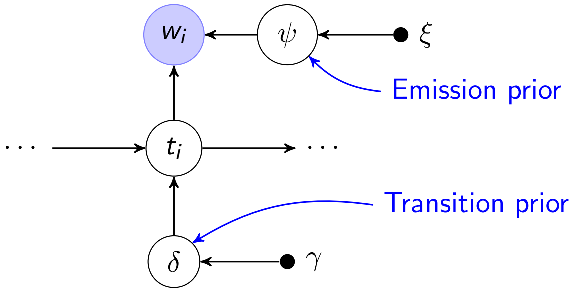
Standard Hidden Markov Model



- Difficult to model sparsity of words given tag
- Is still competitive with Bayesian HMM

Models

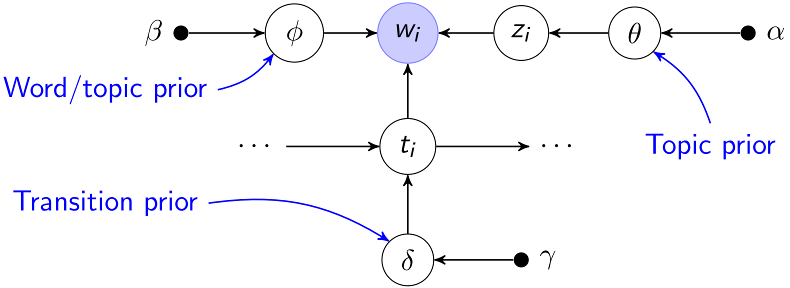
Bayesian Hidden Markov Model



- Can model sparsity through hyperparameters
- Difficult to capture content/function dichotomy

Models

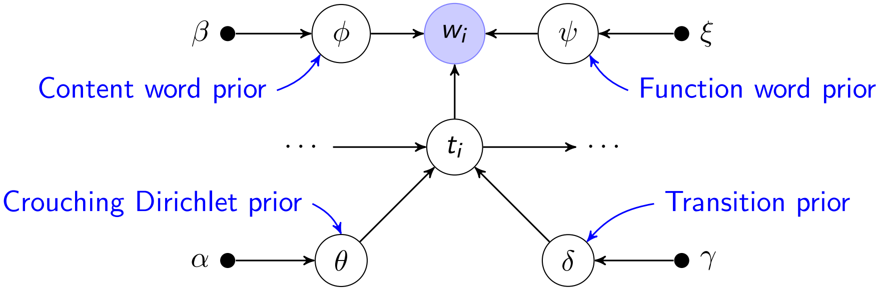
Latent Dirichlet Allocation/Hidden Markov Model [Griffiths et al. 2005]



- An LDA that jointly handles stopword removal
- Captures topic/non-topic dichotomy
- Conflates all topical content words into a single state

Model I

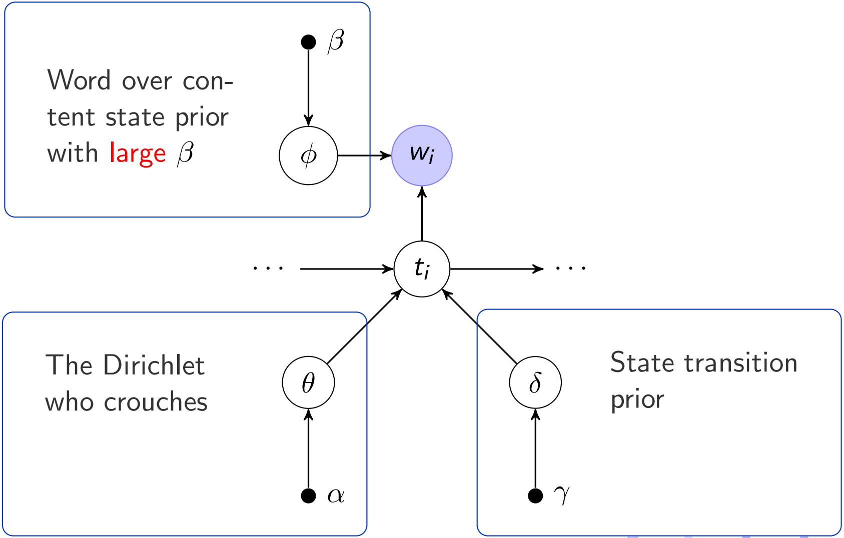
Crouching Dirichlet, Hidden Markov Model



- Define a composite distribution that models content states and function states separately
- Content words given tag are less sparse than function words given tag
- Assume that content words and tags have greater variance across contexts

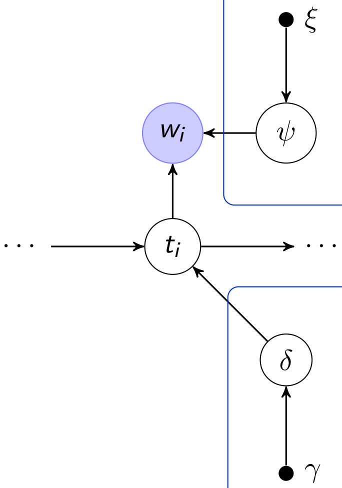
Model I

Crouching Dirichlet, Hidden Markov Model (Content States)



Model I

Crouching Dirichlet, Hidden Markov Model (Function States)

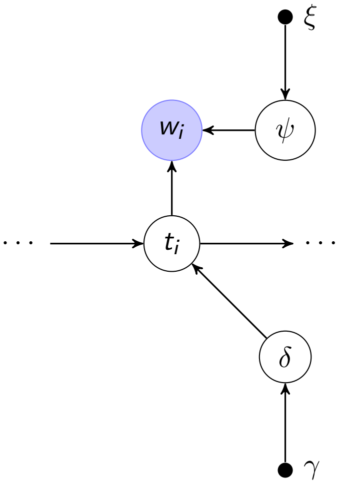


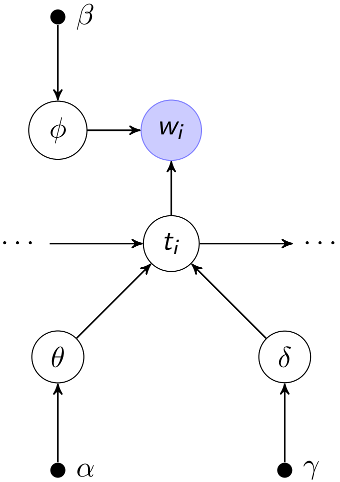
Word over function state prior with small ξ

State transition prior

Models

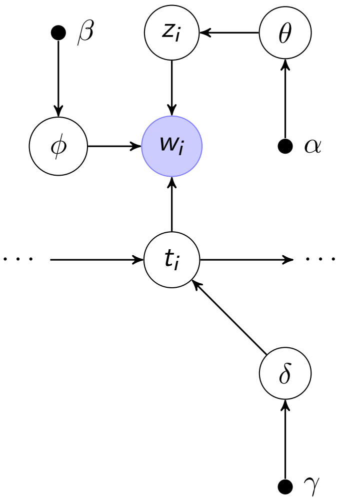
Bayesian Hidden Markov Model vs CDHMM





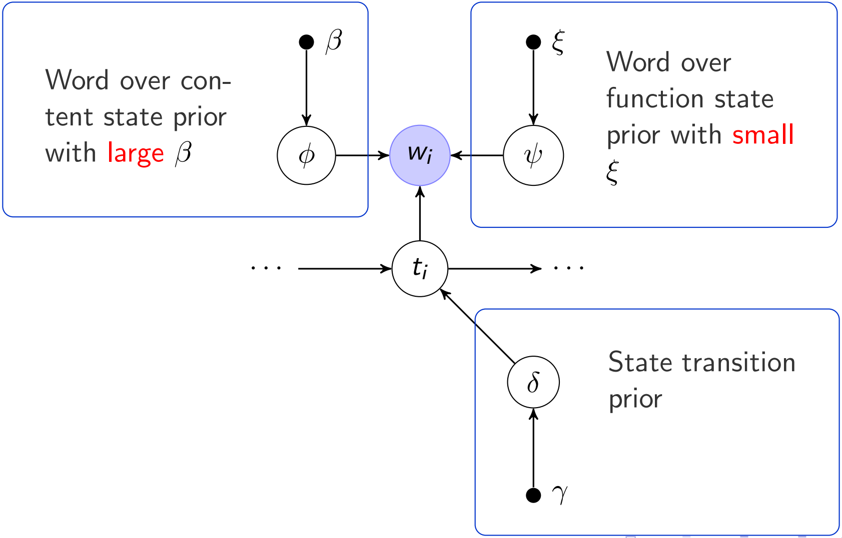
Models

LDAHMM vs CDHMM



Model II

HMM+



Word over content state prior with large β
β
ϕ
w_i
ψ
ξ
Word over function state prior with small ξ
$\cdots$
t_i
$\cdots$
δ
State transition prior
γ

Parameter inference: Collapsed Gibbs Sampling

Conditional distribution of interest

$$p(t_i|\mathbf{t}_{-i}, \mathbf{w})$$

Bayesian HMM conditional distribution

$$\frac{N_{w_i|t_i} + \xi}{N_{t_i} + W\xi} \frac{\left(N_{t_i|t_{i-1}} + \gamma\right)\left(N_{t_{i+1}|t_i} + \mathrm{I}[t_{i-1} = t_i = t_{i+1}] + \gamma\right)}{N_{t_i} + T\gamma + \mathrm{I}[t_i = t_{i-1}]}$$

CDHMM conditional distribution

$$\begin{cases} \frac{N_{w_i|t_i}+\beta}{N_{t_i}+W\beta} \frac{N_{t_i|d_i}+\alpha}{N_{d_i}+C\alpha} \frac{\left(N_{t_i|t_{i-1}}+\gamma\right)\left(N_{t_{i+1}|t_i}+\mathrm{I}[t_{i-1}=t_i=t_{i+1}]+\gamma\right)}{N_{t_i}+T\gamma+\mathrm{I}[t_i=t_{i-1}]} & t_i \in C \\ \frac{N_{w_i|t_i}+\xi}{N_{t_i}+W\xi} \frac{\left(N_{t_i|t_{i-1}}+\gamma\right)\left(N_{t_{i+1}|t_i}+\mathrm{I}[t_{i-1}=t_i=t_{i+1}]+\gamma\right)}{N_{t_i}+T\gamma+\mathrm{I}[t_i=t_{i-1}]} & t_i \in F \end{cases}$$

Data & Experiments

Corpora

- Penn Treebank Wall Street Journal: English, 1M words
- Brown: English, 800K words
- Tiger: German, 450K words
- Floresta: Portuguese, 200K words
- Uspanteko: 70K words, transcribed text, tagged over morphemes

Data & Experiments

Evaluation measures

- Greedily matched accuracy (one-to-one, many-to-one)
- Pairwise token level precision, recall, f-score
- Variation of information [Meila, 2007]

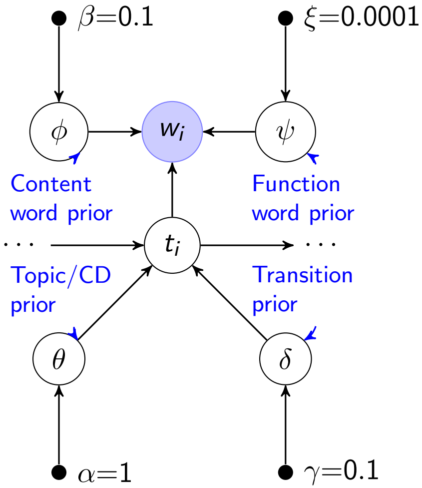
Data & Experiments

Models

- Bayesian HMM: Does not model content/function dichotomy
- LDAHMM: Models topic/non-topic dichotomy
- HMM+: Models content/function dichotomy w/ different blocked priors
- CDHMM: Models content/function dichotomy w/ different blocked priors and a crouching Dirichlet prior

Data & Experiments

Parameter settings



Hyperparameters

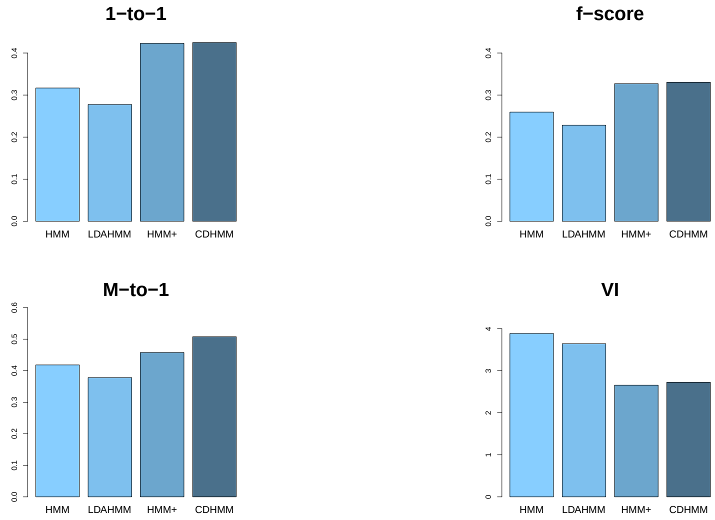
- Uninformative/symmetric priors
- No hyperparameter re-estimation

Other settings

- Total no. of states: 20/30/40/50
- No. of content states: 5
- Iterations: 1000
- 10 chains, single sample each

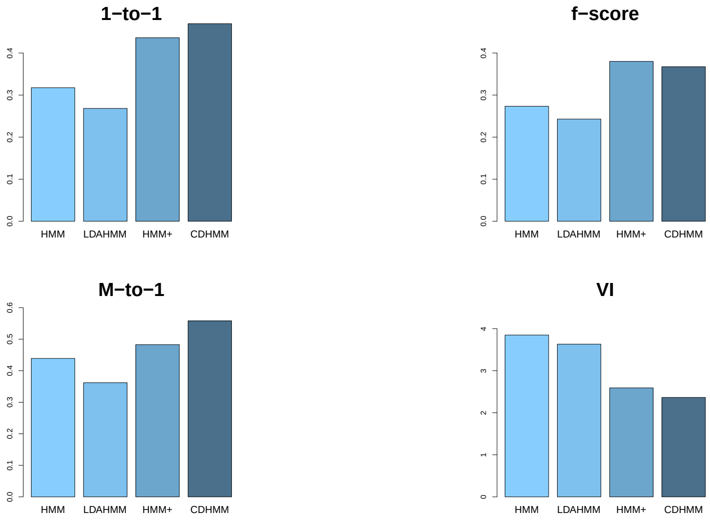
Full Results by Corpus

Wall Street Journal



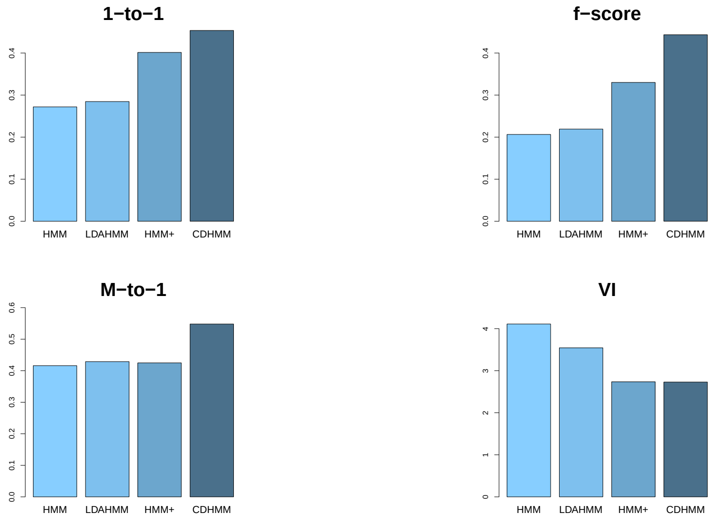
Full Results by Corpus

Brown



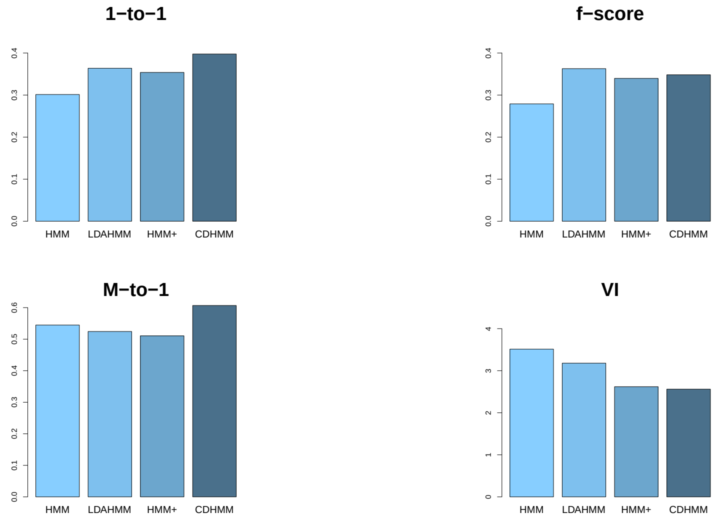
Full Results by Corpus

Tiger



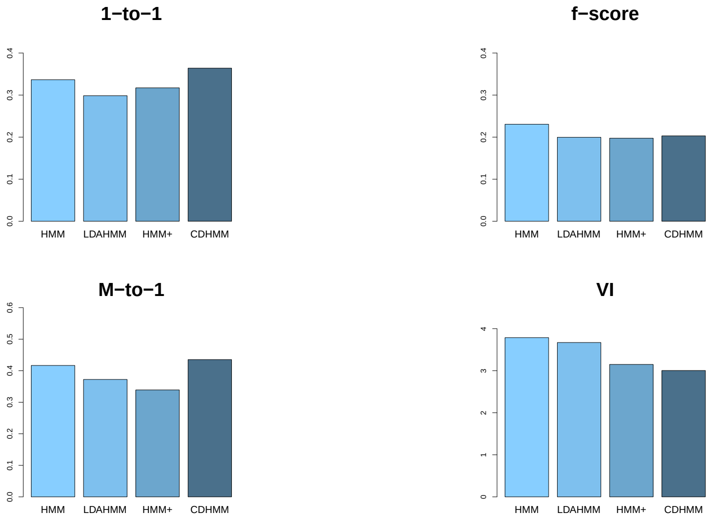
Full Results by Corpus

Floresta



Full Results by Corpus

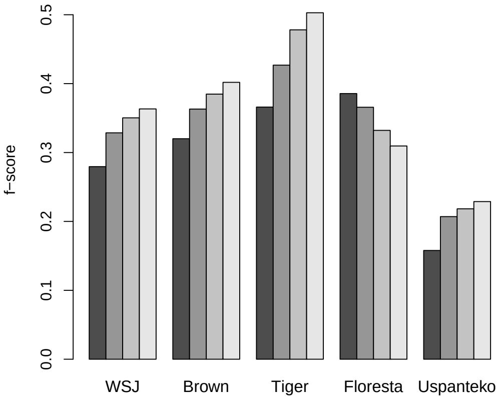
Uspanteko



Full Results by Corpus

- CDHMM always wins or ties on accuracy
- CDHMM loses once to HMM+ on VI (Floresta)
- HMM wins once on f-score (Uspanteko)
- HMM+ ties with LDAHMM once on f-score (Floresta)
- CDHMM wins elsewhere on f-score

f-score dependent on number of states

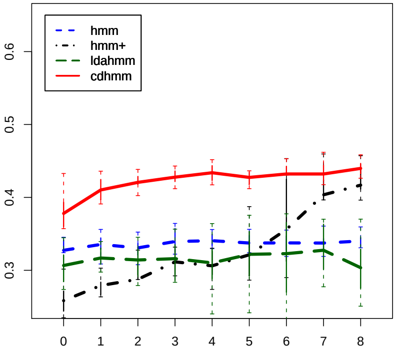


If number of model states are kept close to number of gold tags

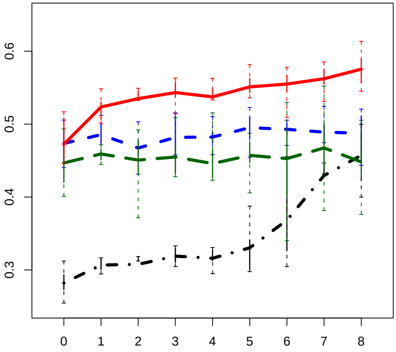
CDHMM wins or ties on every corpus on every measure

Results: Accuracy Learning curves

Wall Street Journal



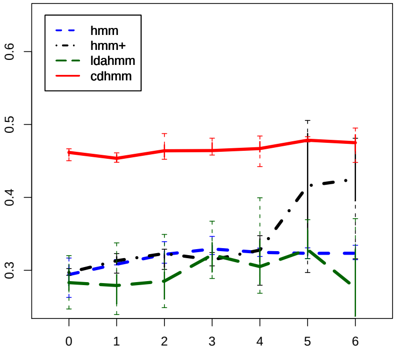
Wall Street Journal one-to-one



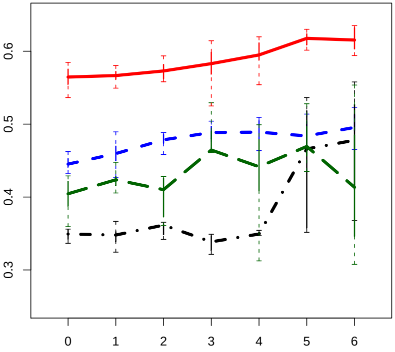
Wall Street Journal many-to-one

Results: Accuracy Learning curves

Brown



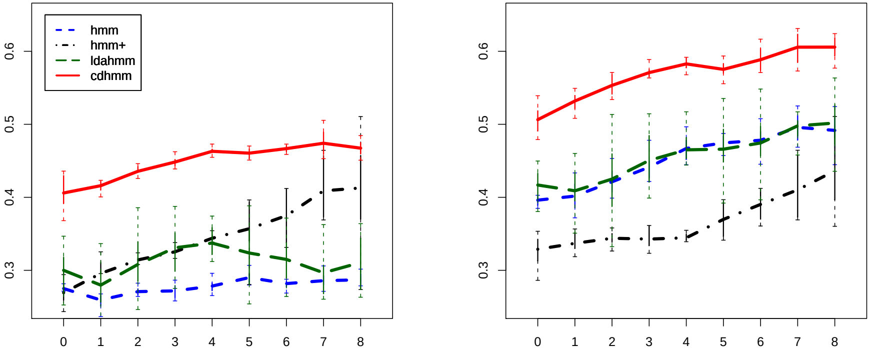
Brown one-to-one



Brown many-to-one

Results: Accuracy Learning curves

Tiger (German)

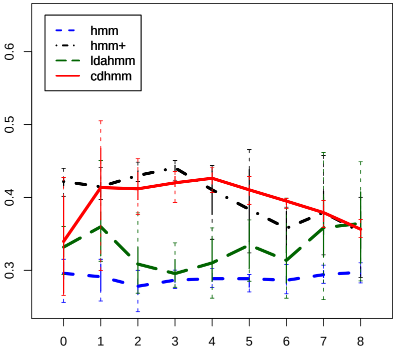


Tiger one-to-one

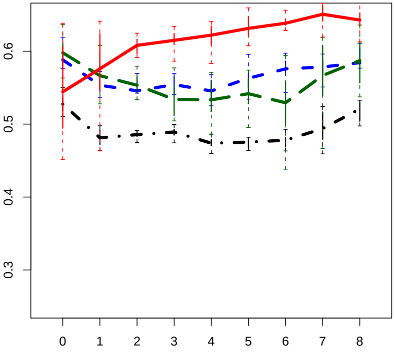
Tiger many-to-one

Results: Accuracy Learning curves

Floresta (Portuguese)



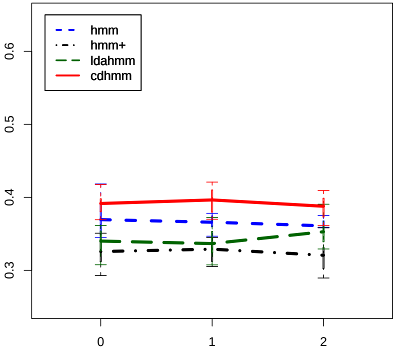
Floresta one-to-one



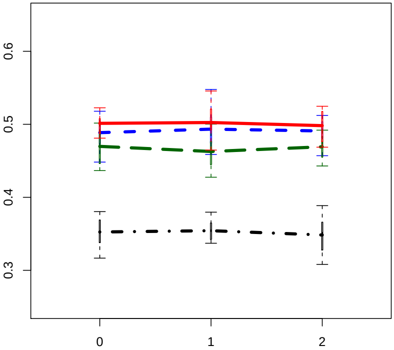
Floresta many-to-one

Results: Accuracy Learning Curves

Uspanteko



Uspanteko one-to-one



Uspanteko many-to-one

Conclusion

- Modeling content/function word dichotomy improves the HMM
- Capturing context variance of content words even further improves the HMM
- Merely modeling this dichotomy through blocked hyperparameters works as well

Thank you!

Bibliography

[Meila, 2007] Marina Meilă.
Comparing clusteringsan information based distance.
Journal of Multivariate Analysis, 2007.

[Griffiths et al., 2005] Thomas L. Griffiths, Mark Steyvers, David M. Blei and Joshua B. Tenenbaum.
Integrating topics and syntax.
Advances in neural information processing systems, 2005