A Bayesian Mixture Model for Part-of-Speech Induction Using Multiple Features

Christos Christodoulopoulos
Sharon Goldwater
Mark Steedman

School of Informatics,
University of Edinburgh

EMNLP, 2011

Unsupervised PoS Induction

- Why is it useful?
  - Low-density languages
  - Large amounts of unlabeled data
  - Unsupervised dependency parsing

Unsupervised PoS Induction

- Why is it useful?
  - Low-density languages
  - Large amounts of unlabeled data
  - Unsupervised dependency parsing
- Substantial amount of literature

Unsupervised PoS Induction

- Why is it useful?
  - Low-density languages
  - Large amounts of unlabeled data
  - Unsupervised dependency parsing
- Substantial amount of literature
- Performance is increasing
  - Complex machine learning methods

This work

- Simple generative model
- Easy to incorporate multiple features
  - Context, morphology, alignment
- Competitive results

This work

- Simple generative model
- Easy to incorporate multiple features
  - Context, morphology, alignment
- Competitive results

- Combine various ideas from literature

Insights from literature

Type-based

- Every token of a word gets the same cluster

Insights from literature

Type-based

- Every token of a word gets the same cluster
- Good approximation
  - 93% upper bound on WSJ

Insights from literature

Type-based

- Every token of a word gets the same cluster
- Good approximation
  - 93% upper bound on WSJ

- Used by many systems:
  - Brown et al. (1992), Clark (2003), Lee et al. (2010)

Insights from literature

Clustering model

- Alternative to HMMs
  - Schütze (1995), Toutanova & Johnson (2007), Lamar et al. (2010)

Insights from literature

Clustering model

- Alternative to HMMs
  - Schütze (1995), Toutanova & Johnson (2007), Lamar et al. (2010)

- Allows for additional features to be used

Insights from literature

Clustering model

- Alternative to HMMs
  - Schütze (1995), Toutanova & Johnson (2007), Lamar et al. (2010)

- Allows for additional features to be used
  - most importantly...

Insights from literature

Morphology modelling

Alignment features

Insights from literature

Morphology modelling

- Proven highly successful predictor
  - Clark (2003), Berg-Kirkpatrick et al. (2010)
- Type-level feature (more later)

Alignment features

Insights from literature

Morphology modelling

- Proven highly successful predictor
  - Clark (2003), Berg-Kirkpatrick et al. (2010)
- Type-level feature (more later)

Alignment features

- Parallel corpora
- Preliminary evidence
  - Naseem et al. (2009), Das & Petrov (2011)

Bayesian Mixture Model

	f1	f2	f3	f4	f5	f6	...
ball	98	24	0	0	85	0	
throwing	0	0	1	0	0	25	
big	278	0	0	0	62	0	
quickly	0	0	0	1	0	0	
loves	0	0	0	0	0	78	
...							

Bayesian Mixture Model

	the	green	f3	f4	f5	f6	...
ball	98	24	0	0	85	0	
throwing	0	0	1	0	0	25	
big	278	0	0	0	62	0	
quickly	0	0	0	1	0	0	
loves	0	0	0	0	0	78	
...							

Left context words

Bayesian Mixture Model

	the	green	ing	ly	f5	f6	...
ball	98	24	0	0	85	0	
throwing	0	0	1	0	0	25	
big	278	0	0	0	62	0	
quickly	0	0	0	1	0	0	
loves	0	0	0	0	0	78	
...							

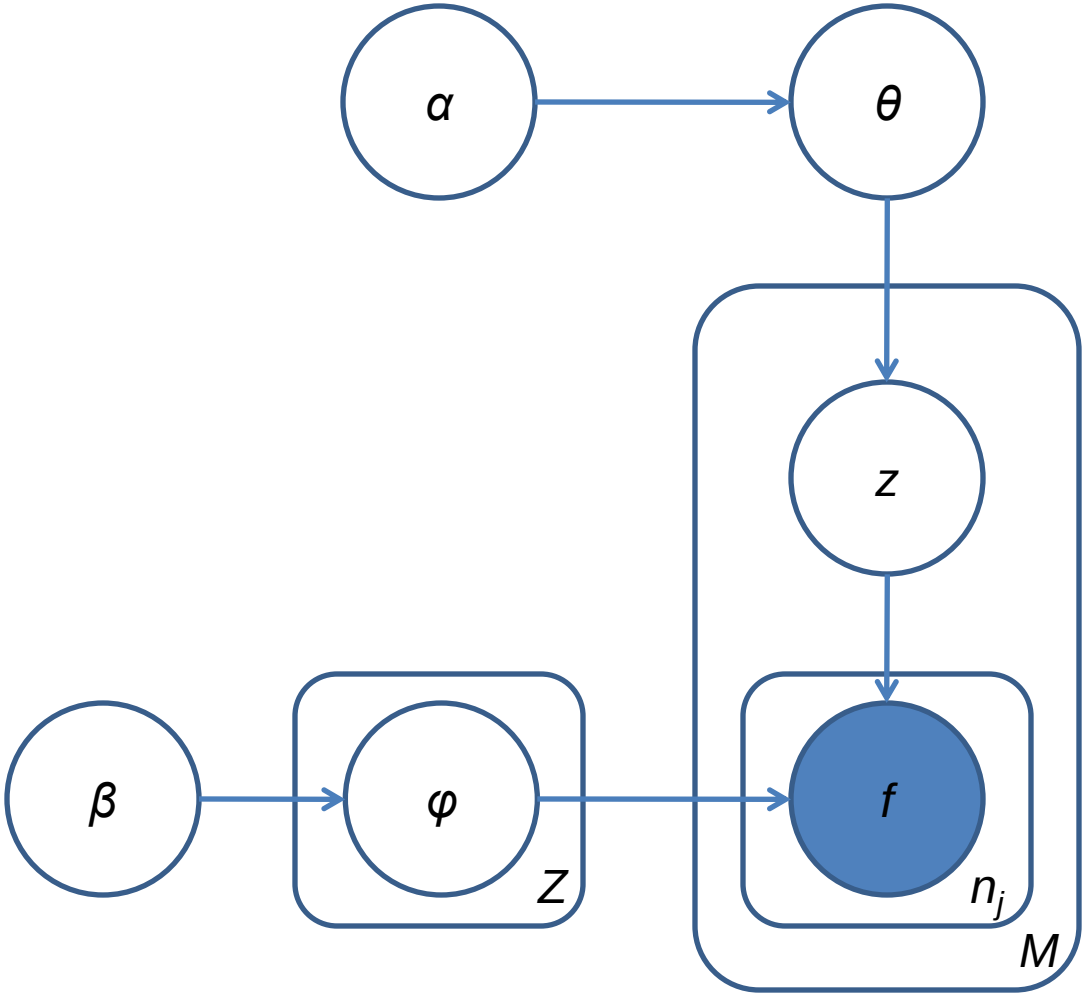
Morphology suffixes

Bayesian Mixture Model

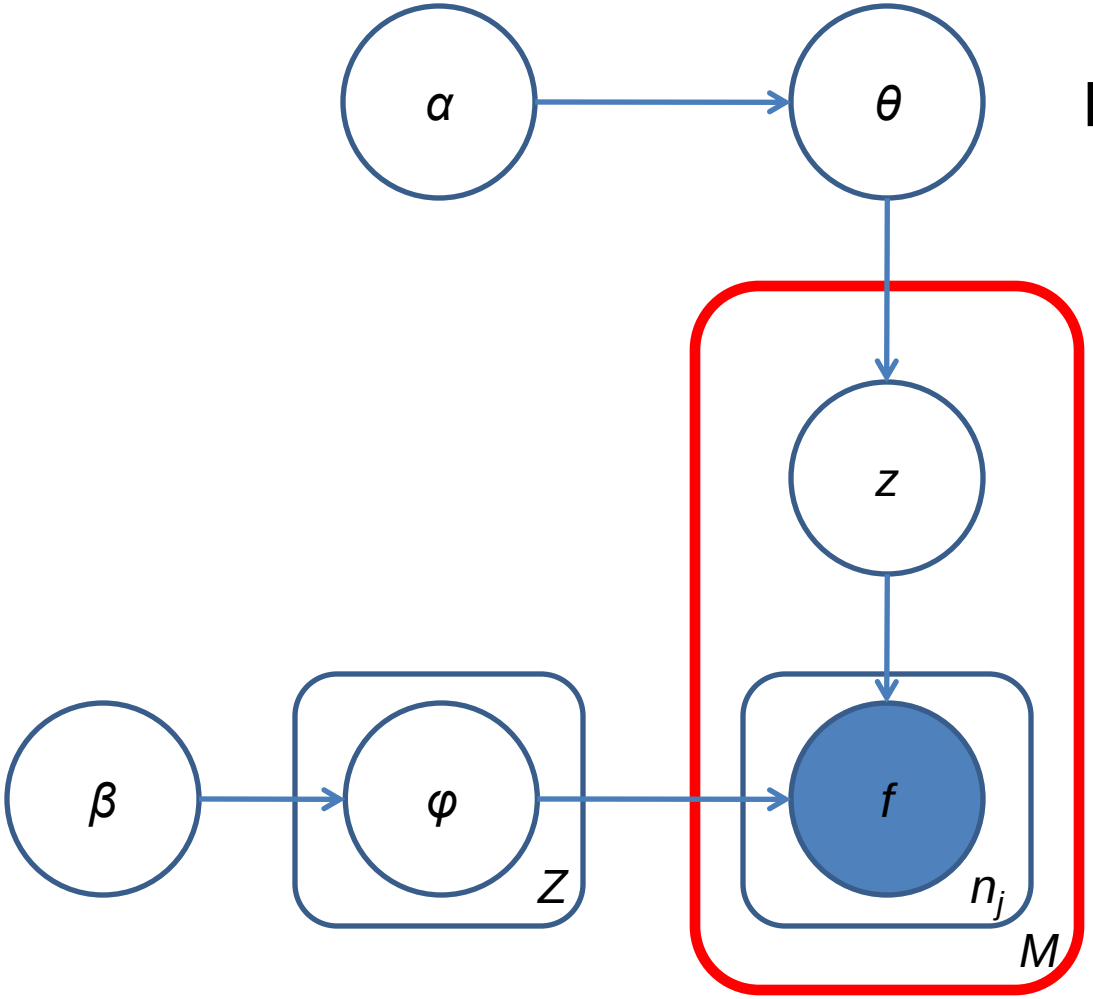
	the	green	ing	ly	ein	homme	...
ball	98	24	0	0	85	0	
throwing	0	0	1	0	0	25	
big	278	0	0	0	62	0	
quickly	0	0	0	1	0	0	
loves	0	0	0	0	0	78	
...							

Aligned context words

Basic model structure

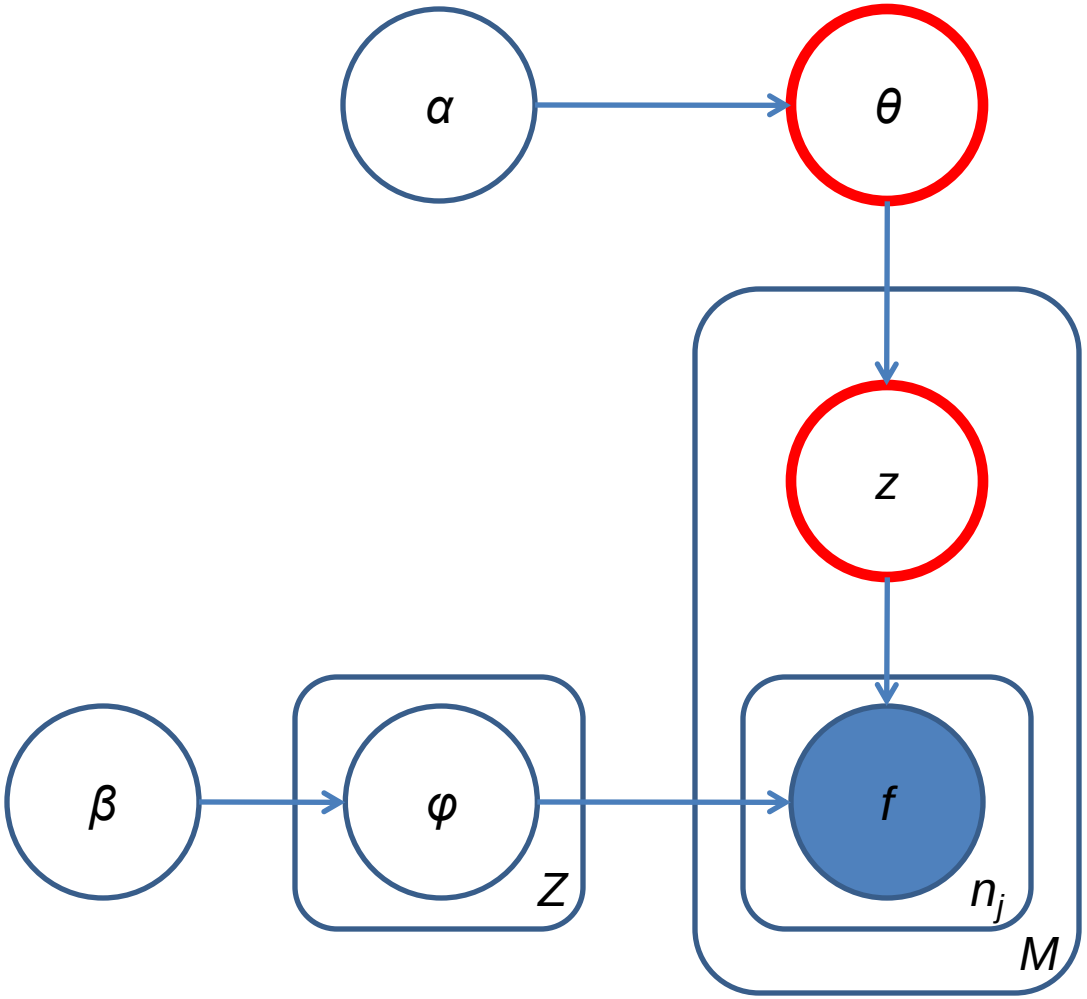


Basic model structure



For each word type i

Basic model structure

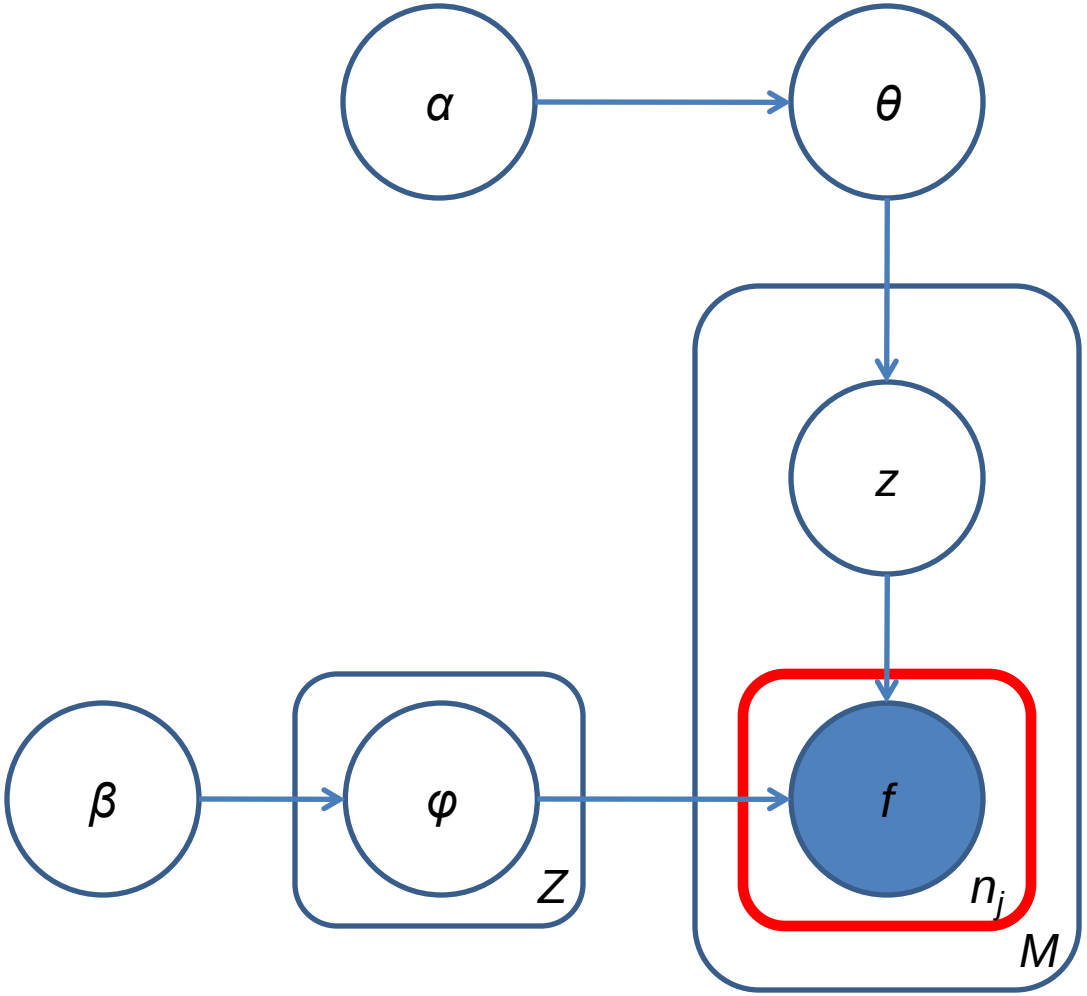


For each word type i

choose a class z_i

(conditioned on θ)

Basic model structure



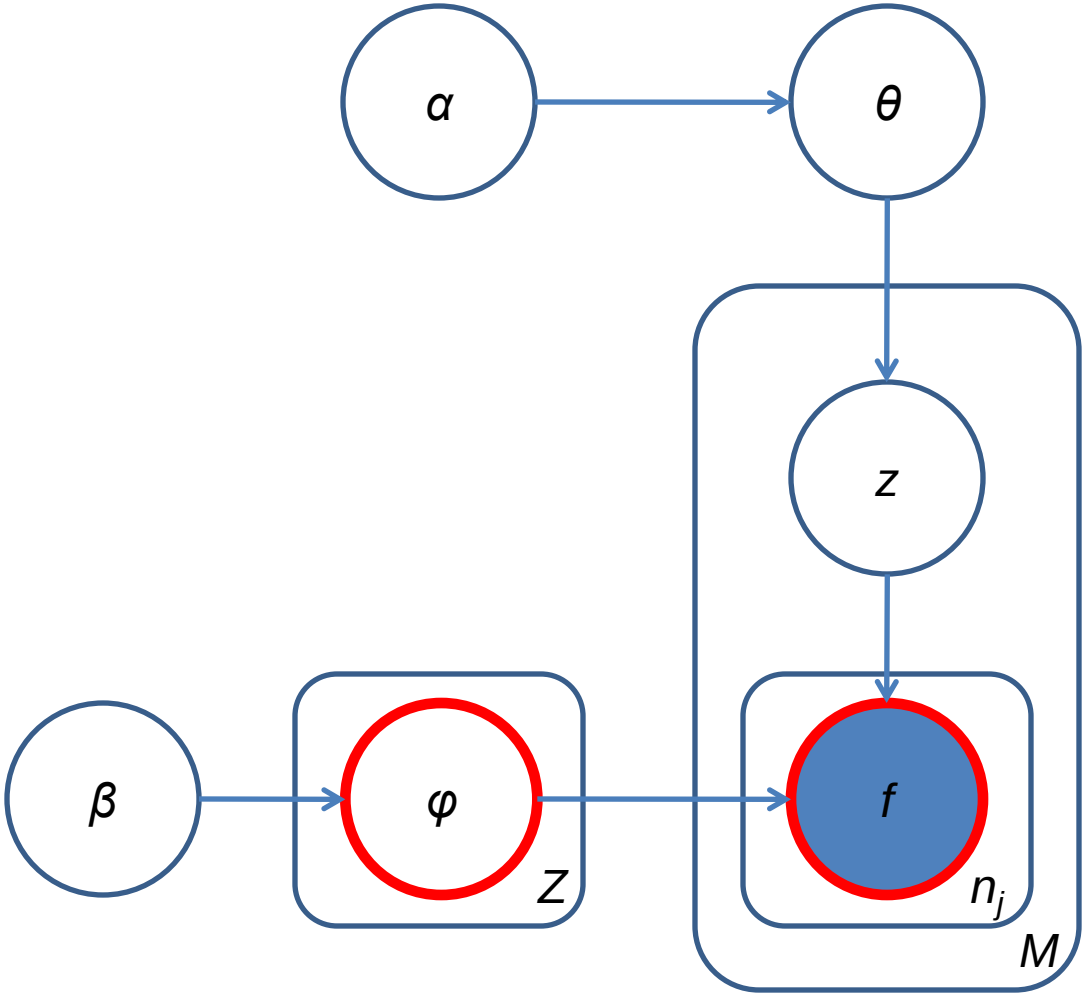
For each word type i

choose a class z_i

(conditioned on θ)

For each word token j

Basic model structure



For each word type i

choose a class z_i

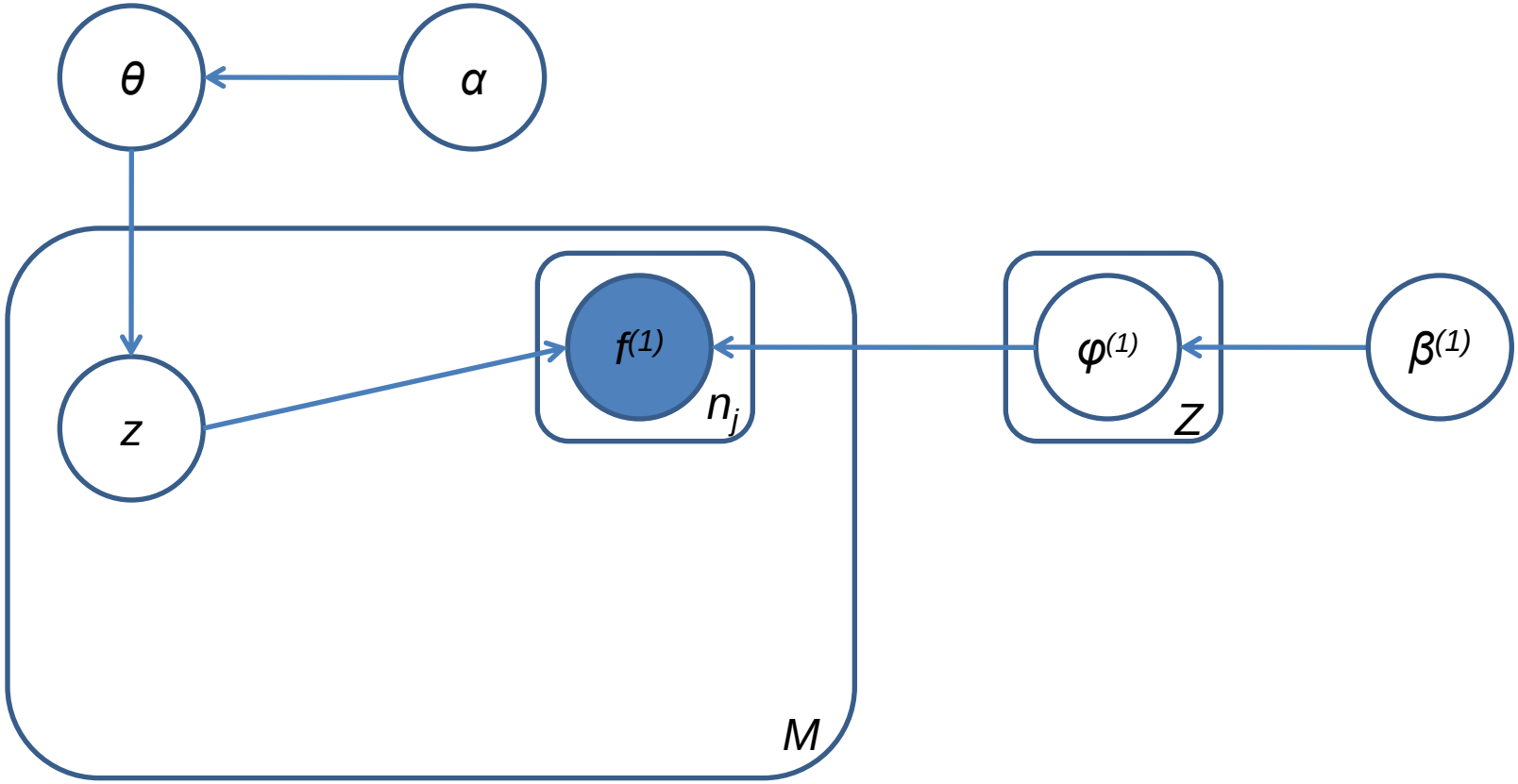
(conditioned on θ)

For each word token j

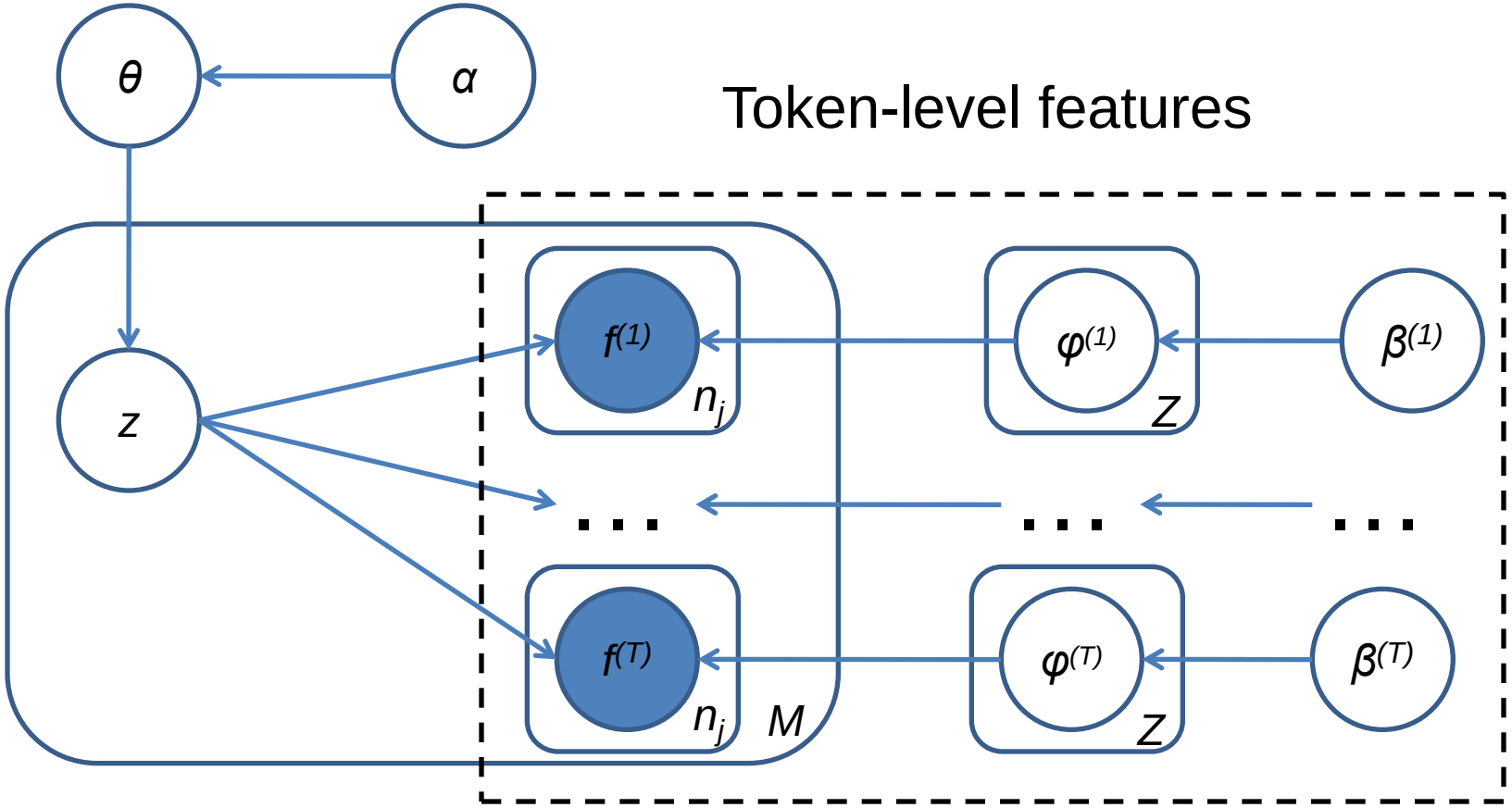
choose a feature f_{ij}

(conditioned on φ_i)

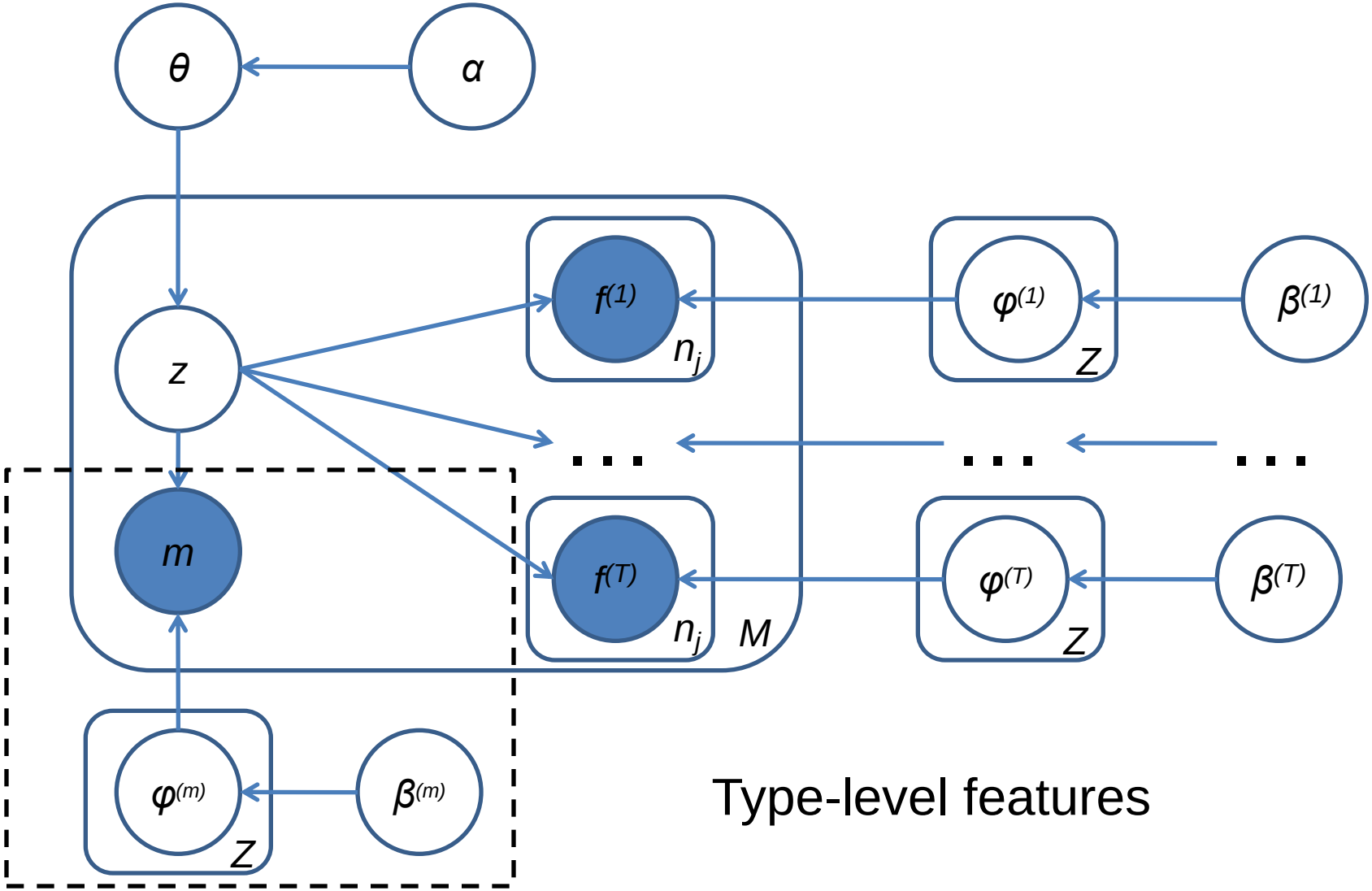
Extended model



Extended model



Extended model



Extended model features

Morphology

- Suffixes extracted from morfessor

Extended model features

Morphology

- Suffixes extracted from morfessor
- Extended morphology features from

  Haghighi & Klein (2006)

  – *initial-capital, contains-hyphen, contains-digit* plus *contains-punctuation*

Extended model features

Alignments

- Bidirectional Giza++ alignments
- Left and right context words of the aligned token as features

It was a bright cold day in April

Extended model features

Alignments

- Bidirectional Giza++ alignments
- Left and right context words of the aligned token as features

It was a bright cold day in April

Într- o zi senină şi friguroasă de aprilie

Extended model features

Alignments

- Bidirectional Giza++ alignments
- Left and right context words of the aligned token as features

It was a bright cold day in April

Într- o zi senină şi friguroasă de aprilie

Inference

- Collapsed Gibbs sampler

Inference

- Collapsed Gibbs sampler
- Main difference:
  - Features of each word type are co-dependent
  - Each feature probability must include all the counts due to previous features

Inference

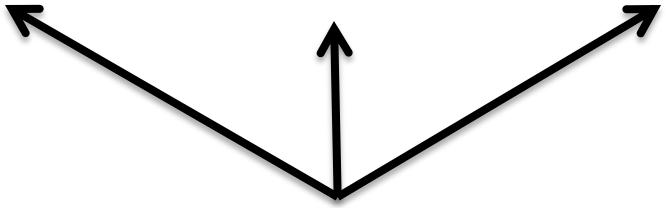
- Collapsed Gibbs sampler
- Main difference:
  - Features of each word type are co-dependent
  - Each feature probability must include all the counts due to previous features
- Independence assumption for all kinds of features (context, morphology, alignments)

Bayesian Mixture Model

	the	green	ing	ly	ein	homme	...
ball	98	24	0	0	85	0	
throwing	0	0	1	0	0	25	
big	278	0	0	0	62	0	
quickly	0	0	0	1	0	0	
loves	0	0	0	0	0	78	
...							

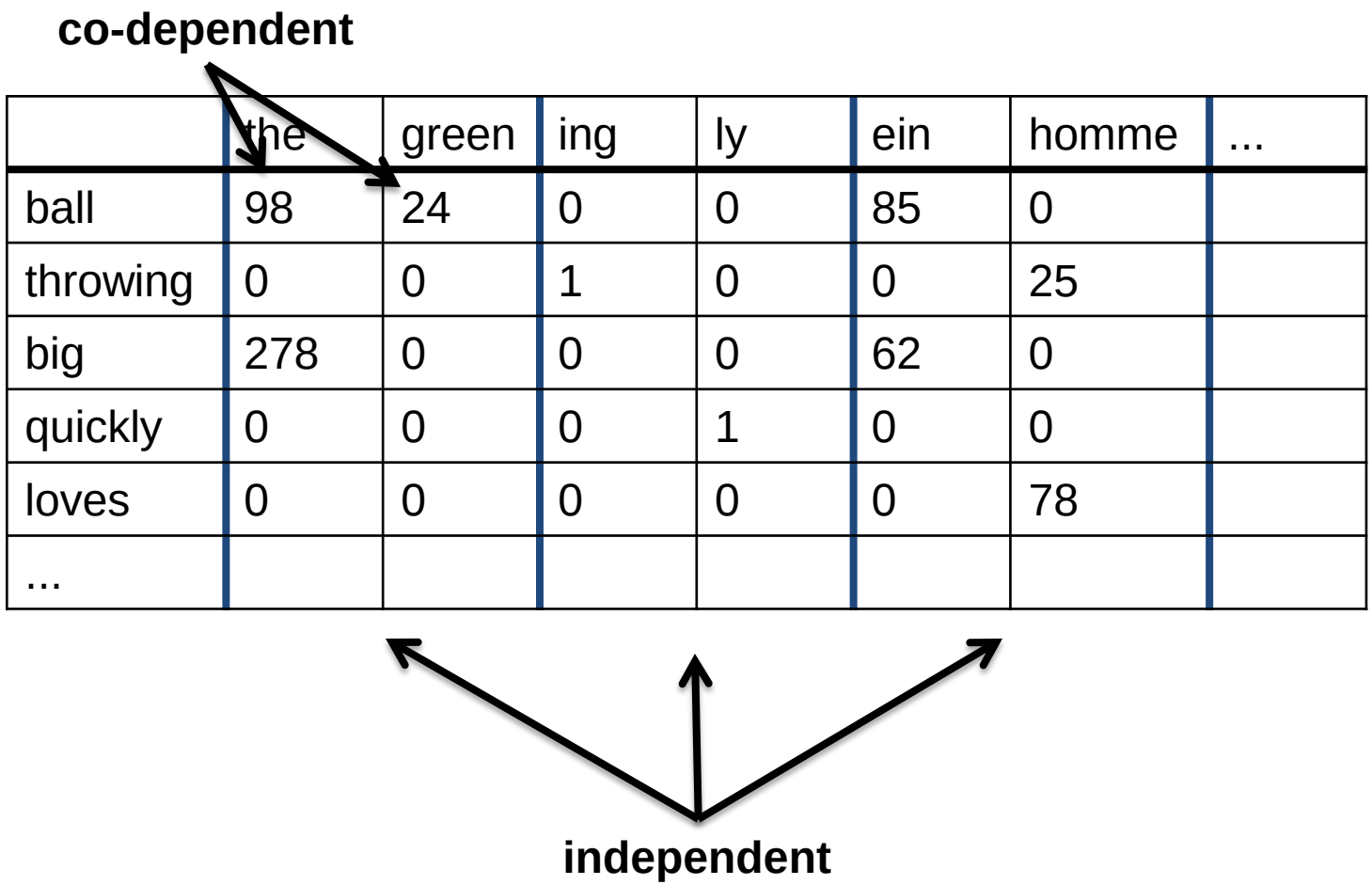
Bayesian Mixture Model

	the	green	ing	ly	ein	homme	...
ball	98	24	0	0	85	0	
throwing	0	0	1	0	0	25	
big	278	0	0	0	62	0	
quickly	0	0	0	1	0	0	
loves	0	0	0	0	0	78	
...							



independent

Bayesian Mixture Model



Experimental setup

- Features
  - [base] 100 most frequent left/right context words
  - [morph] morfessor suffixes
  - [morph(ext)] 4 Haghighi & Klein features
  - [aligns] 100 most frequent context words of the aligned token

Experimental setup

- Features
  - [base] 100 most frequent left/right context words
  - [morph] morfessor suffixes
  - [morph(ext)] 4 Haghighi & Klein features
  - [aligns] 100 most frequent context words of the aligned token
- Hyperparameters
  - Inference using Metropolis-Hastings sampler

Experimental setup

- Features
  - [base] 100 most frequent left/right context words
  - [morph] morfessor suffixes
  - [morph(ext)] 4 Haghighi & Klein features
  - [aligns] 100 most frequent context words of the aligned token
- Hyperparameters
  - Inference using Metropolis-Hastings sampler
- Simulated Annealing

Datasets

- Penn TreeBank (WSJ)
  - 45k sentences + 7k version (wsj-s)
- MULTEXT-East (1984 novel)
  - Parallel corpus
  - 8 languages, ~7k sentences
- CoNLL-X 2006 Shared Task
  - 13 languages

Results

- Many-to-one and V-Measure scores
  - Many-to-one for easy comparison
  - VM less sensitive to number of classes
    - Easy to compare across corpora
    - More stable when mapping to coarse-grained tagsets

Results

- Many-to-one and V-Measure scores
  - Many-to-one for easy comparison
  - VM less sensitive to number of classes
    - Easy to compare across corpora
    - More stable when mapping to coarse-grained tagsets
- Development results on Bulgarian (MULTEXT) and wsj-s

Development results

Bulgarian	**± 1 words VM / M-1**	**± 2 words VM / M-1**
base	58.1 / 70.8	55.4 / 67.6
base (tokens)	48.3 / 62.5	37.0 / 54.4
+morph	58.3 / 74.9	57.4 / 71.9
+morph(ext)	57.8 / 73.7	57.8 / 70.1
+aligns(EN)	58.1 / 72.6	56.7 / 71.1
+aligns(EN)+morph	**59.0** / **75.4**	57.5 / 69.7

English (wsj-s)	**± 1 words VM / M-1**	**± 2 words VM / M-1**
base	63.3 / 64.3	62.4 / 63.3
base (tokens)	48.6 / 57.8	49.3 / 38.3
+morph	66.4 / 66.7	65.1 / 67.2
+morph(ext)	**67.7** / **72.0**	65.6 / 67.0

Development results

Bulgarian	**± 1 words VM / M-1**	**± 2 words VM / M-1**
base	58.1 / 70.8	55.4 / 67.6
base (tokens)	48.3 / 62.5	37.0 / 54.4
+morph	58.3 / 74.9	57.4 / 71.9
+morph(ext)	57.8 / 73.7	57.8 / 70.1
+aligns(EN)	58.1 / 72.6	56.7 / 71.1
+aligns(EN)+morph	**59.0** / **75.4**	57.5 / 69.7

English (wsj-s)	**± 1 words VM / M-1**	**± 2 words VM / M-1**
base	63.3 / 64.3	62.4 / 63.3
base (tokens)	48.6 / 57.8	49.3 / 38.3
+morph	66.4 / 66.7	65.1 / 67.2
+morph(ext)	**67.7** / **72.0**	65.6 / 67.0

- Type-level out-performs token-level

Development results

Bulgarian	± 1 words VM / M-1	± 2 words VM / M-1
base	58.1 / 70.8	55.4 / 67.6
base (tokens)	48.3 / 62.5	37.0 / 54.4
+morph	58.3 / 74.9	57.4 / 71.9
+morph(ext)	57.8 / 73.7	57.8 / 70.1
+aligns(EN)	58.1 / 72.6	56.7 / 71.1
+aligns(EN)+morph	**59.0** / **75.4**	57.5 / 69.7

English (wsj-s)	± 1 words VM / M-1	± 2 words VM / M-1
base	63.3 / 64.3	62.4 / 63.3
base (tokens)	48.6 / 57.8	49.3 / 38.3
+morph	66.4 / 66.7	65.1 / 67.2
+morph(ext)	**67.7** / **72.0**	65.6 / 67.0

- Type-level out-performs token-level
- 2 word window doesn't help

Development results

Bulgarian	± 1 words VM / M-1	± 2 words VM / M-1
base	58.1 / 70.8	55.4 / 67.6
base (tokens)	48.3 / 62.5	37.0 / 54.4
+morph	58.3 / 74.9	57.4 / 71.9
+morph(ext)	57.8 / 73.7	57.8 / 70.1
+aligns(EN)	58.1 / 72.6	56.7 / 71.1
+aligns(EN)+morph	**59.0** / **75.4**	57.5 / 69.7

English (wsj-s)	± 1 words VM / M-1	± 2 words VM / M-1
base	63.3 / 64.3	62.4 / 63.3
base (tokens)	48.6 / 57.8	49.3 / 38.3
+morph	66.4 / 66.7	65.1 / 67.2
+morph(ext)	**67.7** / **72.0**	65.6 / 67.0

- Type-level out-performs token-level
- 2 word window doesn't help
- Morphology helps (extended help more)

Development results

Bulgarian	± 1 words VM / M-1	± 2 words VM / M-1
base	58.1 / 70.8	55.4 / 67.6
base (tokens)	48.3 / 62.5	37.0 / 54.4
+morph	58.3 / 74.9	57.4 / 71.9
+morph(ext)	57.8 / 73.7	57.8 / 70.1
+aligns(EN)	58.1 / 72.6	56.7 / 71.1
+aligns(EN)+morph	**59.0** / **75.4**	57.5 / 69.7

English (wsj-s)	± 1 words VM / M-1	± 2 words VM / M-1
base	63.3 / 64.3	62.4 / 63.3
base (tokens)	48.6 / 57.8	49.3 / 38.3
+morph	66.4 / 66.7	65.1 / 67.2
+morph(ext)	**67.7** / **72.0**	65.6 / 67.0

- Type-level out-performs token-level
- 2 word window doesn't help
- Morphology helps (extended help more)
- Alignments help

Alignment results

	BASE		ALIGNMENTS		
	base	+morph	Avg	Best	+morph
Bulgarian	54.4	54.5	53.1	55.2(EN)	**55.7**
Czech	54.2	53.9	52.6	53.8(EN)	**55.4**
English	62.9	63.3	62.5	63.2(HU)	**63.5**
Estonian	52.8	53.3	52.8	53.5(EN)	**54.3**
Hungarian	53.3	54.8	53.3	53.9(RO)	**55.9**
Romanian	53.9	52.3	**56.2**	57.5(ES)	54.5
Slovene	**57.2**	56.7	54.7	55.9(HU)	56.7
Serbian	**49.1**	49	47.3	48.9(CZ)	48.3

- On average they seem to work

Alignment results

	BASE		ALIGNMENTS		
	base	+morph	Avg	Best	+morph
Bulgarian	54.4	54.5	53.1	55.2(EN)	**55.7**
Czech	54.2	53.9	52.6	53.8(EN)	**55.4**
English	62.9	63.3	62.5	63.2(HU)	**63.5**
Estonian	52.8	53.3	52.8	53.5(EN)	**54.3**
Hungarian	53.3	54.8	53.3	53.9(RO)	**55.9**
Romanian	53.9	52.3	**56.2**	57.5(ES)	54.5
Slovene	**57.2**	56.7	54.7	55.9(HU)	56.7
Serbian	**49.1**	49	47.3	48.9(CZ)	48.3

- On average they seem to work
- Best results -only when we know the best matching pair

Alignment results

	BASE		ALIGNMENTS		
	base	+morph	Avg	Best	+morph
Bulgarian	54.4	54.5	53.1	55.2(EN)	**55.7**
Czech	54.2	53.9	52.6	53.8(EN)	**55.4**
English	62.9	63.3	62.5	63.2(HU)	**63.5**
Estonian	52.8	53.3	52.8	53.5(EN)	**54.3**
Hungarian	53.3	54.8	53.3	53.9(RO)	**55.9**
Romanian	53.9	52.3	**56.2**	57.5(ES)	54.5
Slovene	**57.2**	56.7	54.7	55.9(HU)	56.7
Serbian	**49.1**	49	47.3	48.9(CZ)	48.3

- On average they seem to work
- Best results -only when we know the best matching pair
- More than one language?

Final Results

Other systems:

- k-means
  - Simple vector-based clustering

Final Results

Other systems:

- k-means
  - Simple vector-based clustering
- SVD2 (Lamar et al. 2010)
  - Type-based
  - Vector-based, reduced dimensions

Final Results

Other systems:

- k-means
  - Simple vector-based clustering
- SVD2 (Lamar et al. 2010)
  - Type-based
  - Vector-based, reduced dimensions
- Clark (2003)
  - Type-based
  - Morphology modelling
  - Best results on Multext-East

Final Results - WSJ

V-Measure

	k-means	SVD2	clark	Best Pub.*	BHMM
wsj	59.5	58.2	65.6	**68.8**	66.1
wsj-s	56.7	54.3	63.8	62.3	**67.7**

Many-to-one

	k-means	SVD2	clark	Best Pub.*	BHMM
wsj	61.6	64.0	71.2	**76.1**	72.8
wsj-s	60.1	60.7	68.8	70.7	**72.0**

*Christodoulopoulos et al. (2010)

Final Results - MULTEXT

V-Measure

	k-means	SVD2	clark	BHMM
Bulgarian	50.3	41.7	**55.6**	54.5
Czech	48.6	35.5	52.6	**53.9**
English	56.5	52.3	60.5	**63.3**
Estonian	45.3	38.7	44.4	**53.3**
Hungarian	46.7	39.8	48.9	**54.8**
Romanian	45.2	42.1	40.9	**52.3**
Slovene	46.9	39.5	54.9	**56.7**
Serbian	41.4	39.1	**51.0**	49.0

Final Results - CoNLL

V-Measure

	Ara	Bul	Chi	Cze	Dan	Dut	Ger	Jap	Por	Slo	Spa	Swe	Tur
k-means	**43.3**	53.6	32.6	-	51.7	45.3	58.7	76.1	51.6	52.6	59.5	53.2	40.8
SVD2	27.6	49.0	24.5	-	40.8	36.7	54.1	74.4	45.9	44.0	54.8	47.4	27.4
clark	40.6	**59.6**	31.8	47.1	52.7	52.2	**63.0**	**78.6**	57.4	**53.9**	61.6	**58.9**	36.8
BMMM	42.4	58.8	**42.6**	**48.4**	**59.0**	**54.7**	61.9	77.4	**63.9**	49.4	**63.2**	58.0	**55.4**

Final Results - CoNLL

V-Measure

	Ara	Bul	Chi	Cze	Dan	Dut	Ger	Jap	Por	Slo	Spa	Swe	Tur	Avg
k-means	**43.3**	53.6	32.6	-	51.7	45.3	58.7	76.1	51.6	52.6	59.5	53.2	40.8	51.6
SVD2	27.6	49.0	24.5	-	40.8	36.7	54.1	74.4	45.9	44.0	54.8	47.4	27.4	43.9
clark	40.6	**59.6**	31.8	47.1	52.7	52.2	**63.0**	**78.6**	57.4	**53.9**	61.6	**58.9**	36.8	53.4
BMMM	42.4	58.8	**42.6**	**48.4**	**59.0**	**54.7**	61.9	77.4	**63.9**	49.4	**63.2**	58.0	**55.4**	**55.4**

Final Results - CoNLL

V-Measure

	Ara	Bul	Chi	Cze	Dan	Dut	Ger	Jap	Por	Slo	Spa	Swe	Tur	Avg
k-means	**43.3**	53.6	32.6	-	51.7	45.3	58.7	76.1	51.6	52.6	59.5	53.2	40.8	51.6
SVD2	27.6	49.0	24.5	-	40.8	36.7	54.1	74.4	45.9	44.0	54.8	47.4	27.4	43.9
clark	40.6	**59.6**	31.8	47.1	52.7	52.2	**63.0**	**78.6**	57.4	**53.9**	61.6	**58.9**	36.8	53.4
BMMM	42.4	58.8	**42.6**	**48.4**	**59.0**	**54.7**	61.9	77.4	**63.9**	49.4	**63.2**	58.0	**55.4**	**55.4**

Many-to-one

	Dan	Dut	Ger	Por	Spa	Swe	Avg
k-means	61.6	60.5	67.5	64.4	69.2	62.2	64.2
SVD2	57.6	52.4	64.2	63.1	61.6	58.9	60.8
clark	65.3	67.9	73.9	69.2	71.9	**68.7**	69.5
Best Pub.	66.7	67.3	68.4	75.3	**73.2**	60.6	68.6
BMMM	**71.7**	**71.1**	**74.4**	**76.8**	71.7	68.2	**72.2**

A note on Many-to-one

"Universal tagset" projections

	wsj	wsj-s
BMMM	72.8	72
proj.	**82.2**	**82.5**

A note on Many-to-one

"Universal tagset" projections

	wsj	wsj-s
BMMM	72.8	72
proj.	**82.2**	**82.5**

CoNLL Results

	Ara	Bul	Chi	Cze	Dan	Dut	Jap	Por	Slo	Spa	Swe	Tur	Avg
BMMM	61.5	68.9	69.4	65.7	71.1	71.1	78.5	76.8	56.2	71.7	68.2	58.7	68.2
proj.	91.9	81.8	92.6	70.5	80.8	71.5	87.0	80.9	62.5	81.4	78.4	65.7	**78.7**

A note on Many-to-one

"Universal tagset" projections

	wsj	wsj-s
BMMM	72.8	72
proj.	**82.2**	**82.5**

CoNLL Results

	Ara	Bul	Chi	Cze	Dan	Dut	Jap	Por	Slo	Spa	Swe	Tur	Avg
BMMM	61.5	68.9	69.4	65.7	71.1	71.1	78.5	76.8	56.2	71.7	68.2	58.7	68.2
proj.	91.9	81.8	92.6	70.5	80.8	71.5	87.0	80.9	62.5	81.4	78.4	65.7	**78.7**

Conclusion

- Combine successful ideas from literature
  - Type-based
  - Clustering model

Conclusion

- Combine successful ideas from literature
  - Type-based
  - Clustering model
- Simple generative model

Conclusion

- Combine successful ideas from literature
  - Type-based
  - Clustering model
- Simple generative model
- Easy to incorporate multiple features
  - Context, morphology, alignment
  - Token or type level features

Conclusion

- Combine successful ideas from literature
  - Type-based
  - Clustering model
- Simple generative model
- Easy to incorporate multiple features
  - Context, morphology, alignment
  - Token or type level features
- Competitive results

Thank you!

Code available at:

http://homepages.inf.ed.ac.uk/s0787820/

Gibbs sampling

- Posterior over cluster assignments:

$$P(z_j \mid \mathbf{z}_{-j}, \mathbf{f}, \alpha, \beta) \propto P(z_j \mid \mathbf{z}_{-j}, \alpha, \beta) P(\mathbf{f}_j \mid \mathbf{f}_{-j}, \mathbf{z}, \alpha, \beta)$$

- Prior cluster probability:

$$P(z_j = z \mid \mathbf{z}_{-j}, \alpha) = \frac{n_z + \alpha}{n_\cdot + Z\alpha}$$

- Likelihood:

$$P(\mathbf{f}_j \mid \mathbf{f}_{-j}, z_j = z, \mathbf{z}_{-j}, \beta) = \frac{\prod_{k=1}^{F} \prod_{i=0}^{n_{jk-1}} (n_{jk,z} + i + \beta)}{\prod_{i=0}^{n_{j-1}} (n_{\cdot,z} + i + F\beta)}$$