Carnegie Mellon University

Knowledge-based Word Sense Disambiguation using Topic Models

Devendra Singh Chaplot

Ruslan Salakhutdinov

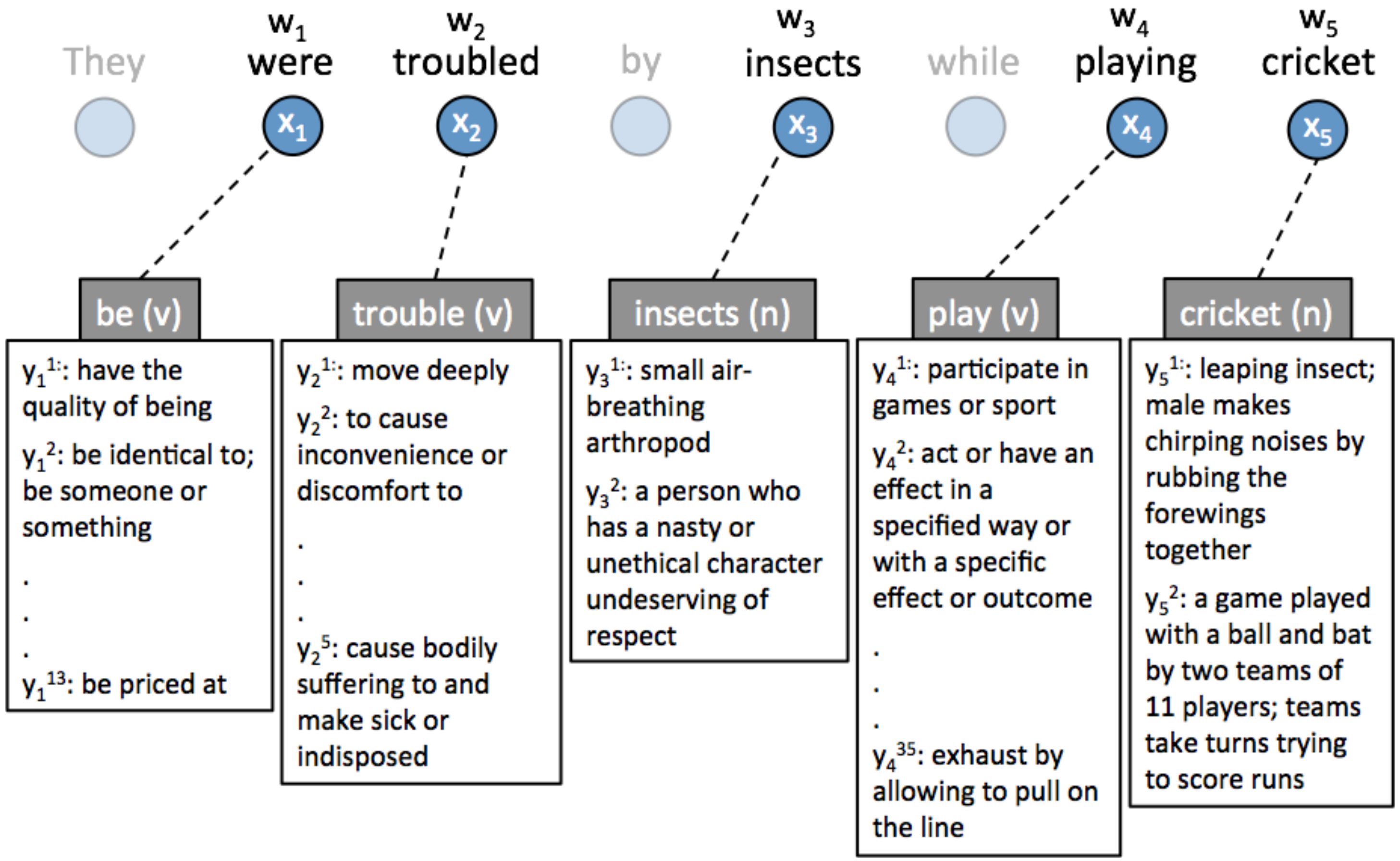
Word Sense Disambiguation

Word sense disambiguation (WSD) is defined as the problem of computationally determining which "sense" of a word is activated by the use of the word in a particular context. [Navigli, 2009]

Problem definition

Input: Raw text

Output: Sense of all content words in the given text



Word Sense Disambiguation

Word sense disambiguation (WSD) is defined as the problem of computationally determining which "sense" of a word is activated by the use of the word in a particular context. [Navigli, 2009]

Motivation

- AI-Complete problem [Mallery, 1988]
- Importance in NLP: Sentiment Analysis, Machine Translation, Information Retrieval, Text summarization, Text Entailment, …

Why unsupervised?

- Supervised WSD performs well but needs sense tagged corpora – usually domain specific
- Obtaining sense tagged corpora is costly in terms of time and money
- Unsupervised approaches are preferred for their resource consciousness and robustness

Context for disambiguating a word?

- Context: discourse that surrounds a language unit and helps to determine its interpretation.

- What should be the context for WSD?
  - Sentence in which the target word occurs [Chaplot et al. 2015]
  - Window of k words around the target word? [Agirre et al. 2014]

Hypothesis

Whole document as context for Word Sense Disambiguation

Hypothesis

Whole document as context for Word Sense Disambiguation

" He forgot the *chips* at the counter. "

Hypothesis

Whole document as context for Word Sense Disambiguation

" He forgot the *chips* at the counter. "

- "***chips***" - potato chips, micro chips, gambling chips?

Hypothesis

Whole document as context for Word Sense Disambiguation

" He forgot the *chips* at the counter. "

- "***chips***" - potato chips, micro chips, gambling chips?
- The presence of other words like '**casino**' and '**gambler**' in the document would indicate the sense of **poker chips**

Hypothesis

Whole document as context for Word Sense Disambiguation

" He forgot the *chips* at the counter. "

- "***chips***" - potato chips, micro chips, gambling chips?
- The presence of other words like '**casino**' and '**gambler**' in the document would indicate the sense of **poker chips**
- The presence of other words like '**electronic**' and '**silicon**' in the document indicate the sense of **micro chip**

Proposed Method

- Using the whole document as context for Word Sense Disambiguation

Proposed Method

- Using the whole document as context for Word Sense Disambiguation
  - One sense per discourse [Gale et al. 1992]

Proposed Method

- Using the whole document as context for Word Sense Disambiguation
  - One sense per discourse [Gale et al. 1992]
  - Need to control computational complexity

Proposed Method

- Using the whole document as context for Word Sense Disambiguation
  - One sense per discourse [Gale et al. 1992]
  - Need to control computational complexity
- Leverage the formalism of Latent Dirichlet Allocation (LDA) [6] to model the whole document

Method – Key Ideas

- Documents have a distribution of synsets: Replace topics in LDA by synsets

Method – Key Ideas

- Documents have a distribution of synsets: Replace topics in LDA by synsets
- Within each synset, some words are more frequent than others - Non uniform prior for word distribution for synset

player, participant
a person who participates in or is skilled at some game

player, **actor**
a theatrical performer

Carnegie Mellon University

Method – Key Ideas

- Documents have a distribution of synsets: Replace topics in LDA by synsets
- Within each synset, some words are more frequent than others - Non uniform prior for word distribution for synset

player, participant
a person who participates in or is skilled at some game

player, **actor**
a theatrical performer

- Some synsets tend to co-occur - Logistic Normal Sense Model

river
a large natural stream of water

bank
sloping land (especially the slope beside a body of water)

bank
a financial institution that accepts deposits and channels the money into lending activities

MACHINE LEARNING DEPARTMENT

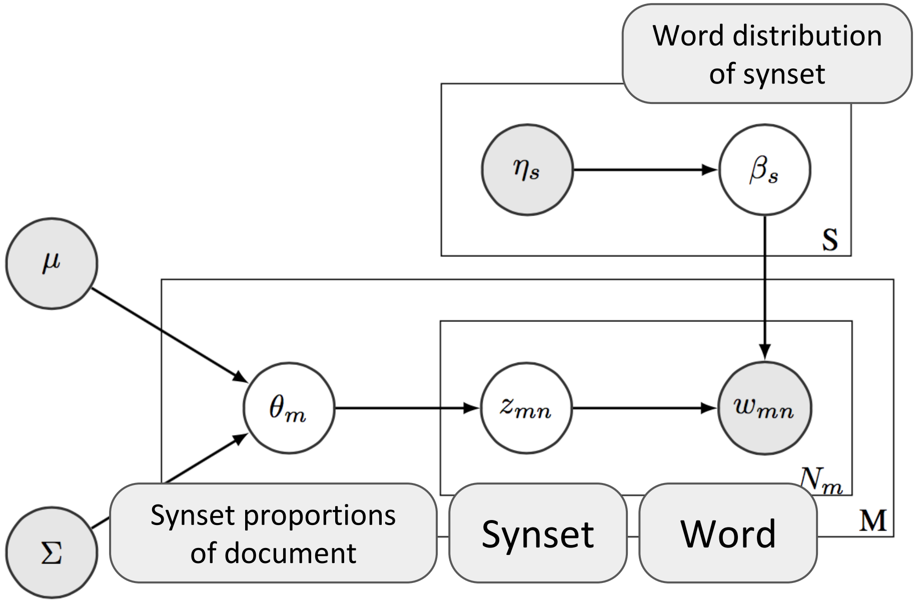
Carnegie Mellon University

Method – Generative Process

Our method assumes a corpus is generated by the following process:

1. For each synset, $s \in \{1, \ldots, S\}$
   (a) Draw word distribution of the synset $\beta_s \sim \mathrm{Dir}(\eta_s)$
2. For each document, $m \in \{1, \ldots, M\}$
   (a) Draw $\alpha_m \sim \mathcal{N}(\mu, \Sigma)$
   (b) Draw synset proportions $\theta_m \sim \mathrm{f}(\alpha_m)$
   (c) For each word in the document, $n \in \{1, \ldots, N_m\}$
      i. Draw synset assignment $z_{mn} \sim \mathrm{Mult}(\theta_m)$
      ii. Draw word from assigned synset $w_{mn} \sim \mathrm{Mult}(\beta_{z_{mn}})$

$$\text{where } f(\boldsymbol{\alpha}) = \frac{exp(\boldsymbol{\alpha})}{\sum_i \alpha_i}$$



Carnegie Mellon University

Word distribution in Synsets

research_worker%1:18:00::
(research worker, researcher, investigator)

cell%1:03:00::
(cell)

cistron%1:08:00::
(gene, cistron, factor)

intervention%1:04:02::
(treatment, intervention)

cancer%1:26:00::
(cancer, malignant neoplastic disease)

Word distribution in Synsets

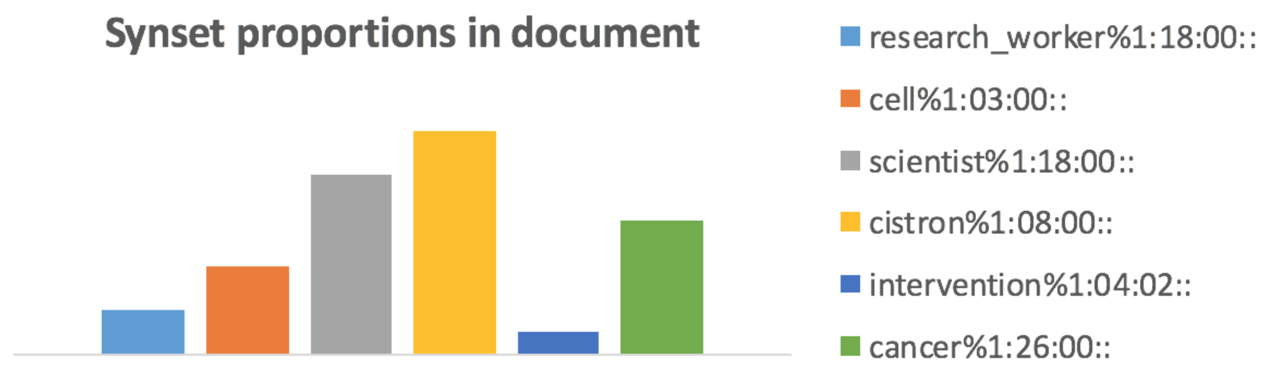
research_worker%1:18:00::
(research worker, researcher, investigator)

cell%1:03:00::
(cell)

cistron%1:08:00::
(gene, cistron, factor)

intervention%1:04:02::
(treatment, intervention)

cancer%1:26:00::
(cancer, malignant neoplastic disease)



University

Word distribution in Synsets

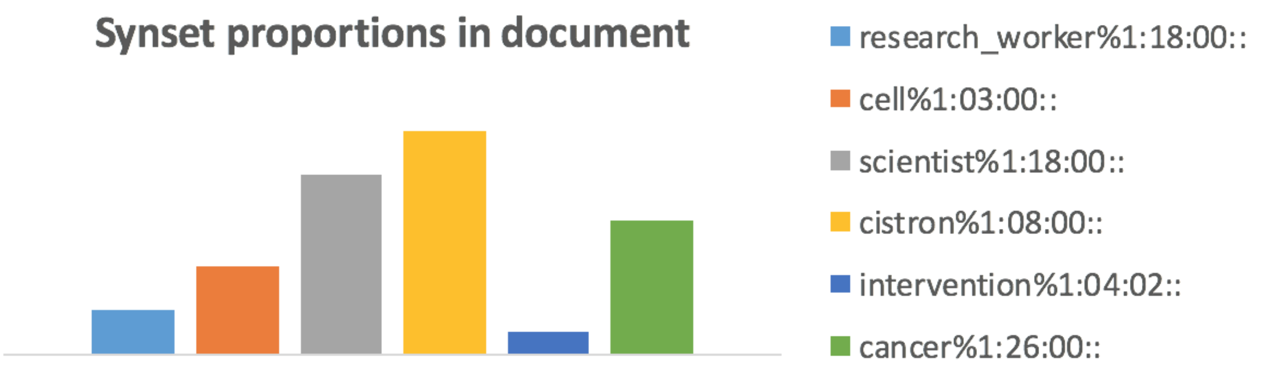
research_worker%1:18:00::
(research worker, researcher, investigator)

cell%1:03:00::
(cell)

cistron%1:08:00::
(gene, cistron, factor)

intervention%1:04:02::
(treatment, intervention)

cancer%1:26:00::
(cancer, malignant neoplastic disease)



Medical scientists are starting to uncover a handful of genes which, if damaged, unleash the chaotic growth of cells that characterizes cancer. Scientists say the discovery of these genes in recent months is painting a new and startling picture of how cancer develops. An emerging understanding of the genes is expected to produce an array of new strategies for future cancer treatment and prevention . That is for the future. Already, scientists are developing tests based on the newly identified genes that, for the first time, can predict whether an otherwise healthy individual is likely to get cancer. “It's a super-exciting set of discoveries”, says Bert Vogelstein, a Johns Hopkins University researcher.

Method - Priors

Method - Priors

- η_{sv} - Frequency of word v in synset s obtained from WordNet

player: 20, participant: 1
a person who participates in or is skilled at some game

player: 1, actor: 14
a theatrical performer

Method - Priors

- η_{sv} - Frequency of word v in synset s obtained from WordNet

player: 20, participant: 1
a person who participates in or is skilled at some game

player: 1, actor: 14
a theatrical performer

- Σ_{ij}^{-1} - Negative similarity between synset i and synset j obtained from WordNet.

river
a large natural stream of water

0.11

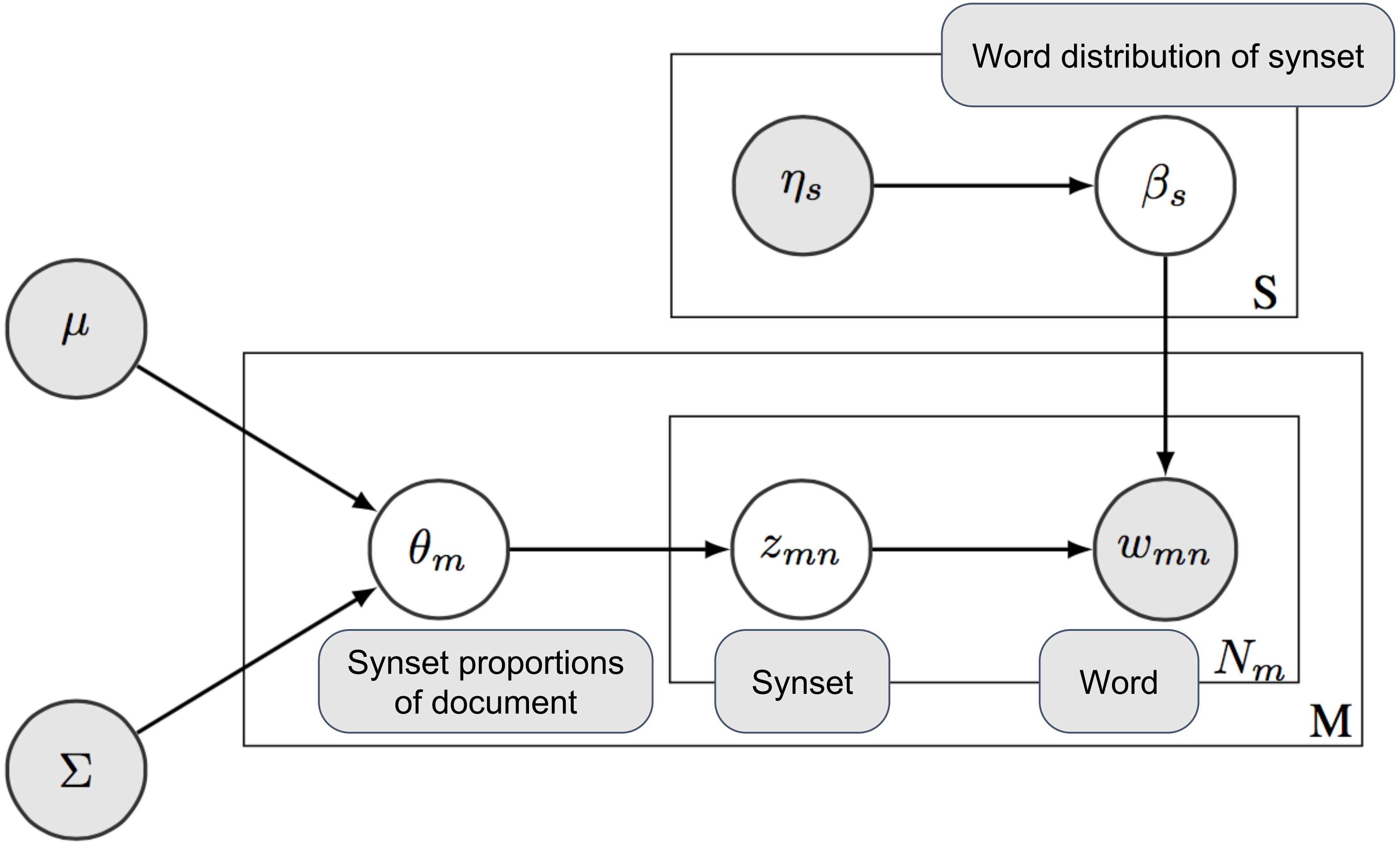
bank
sloping land (especially the slope beside a body of water)

0.07

bank
a financial institution that accepts deposits and channels the money into lending activities

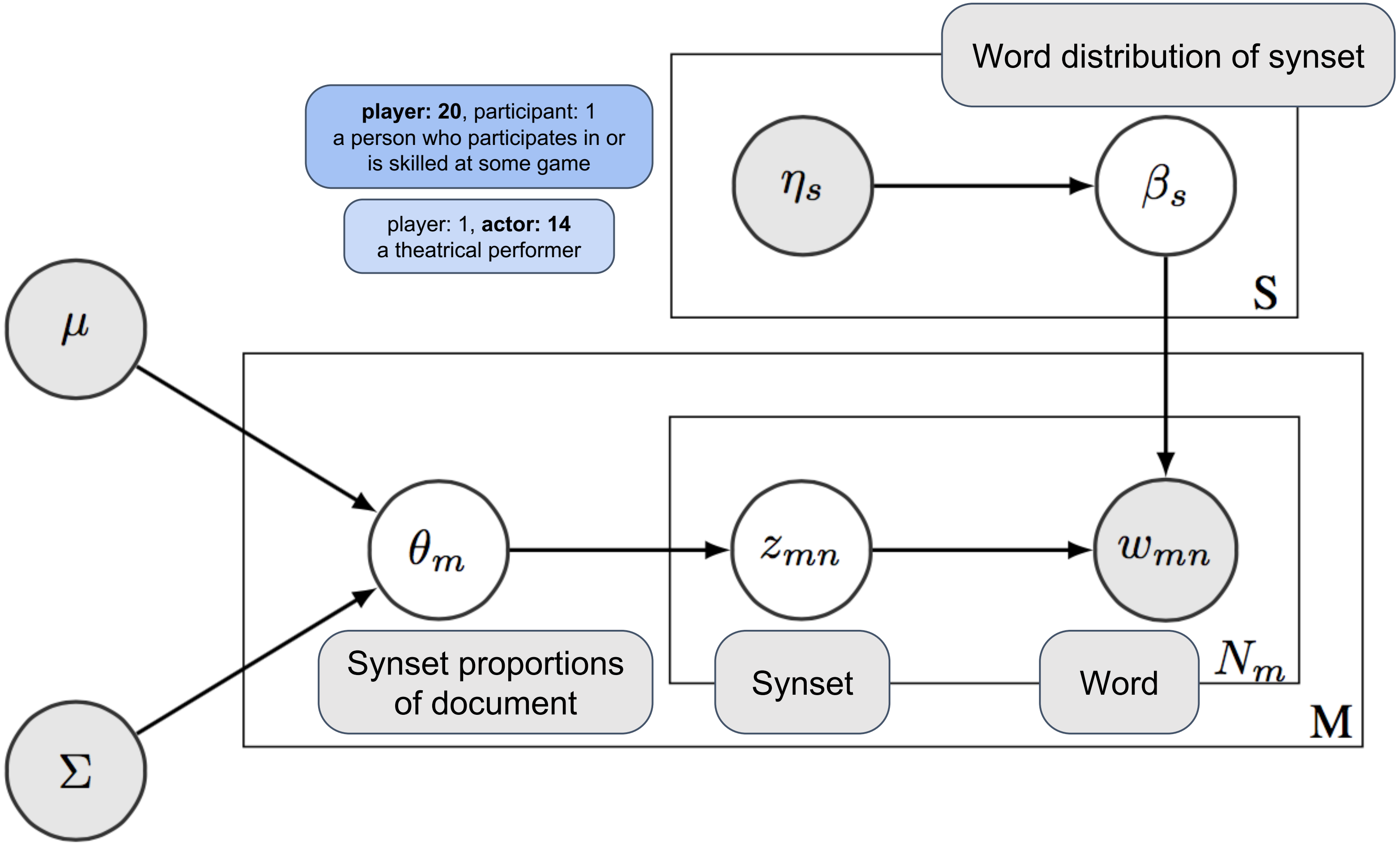
Carnegie Mellon University
MACHINE LEARNING DEPARTMENT

Method – Graphical Model

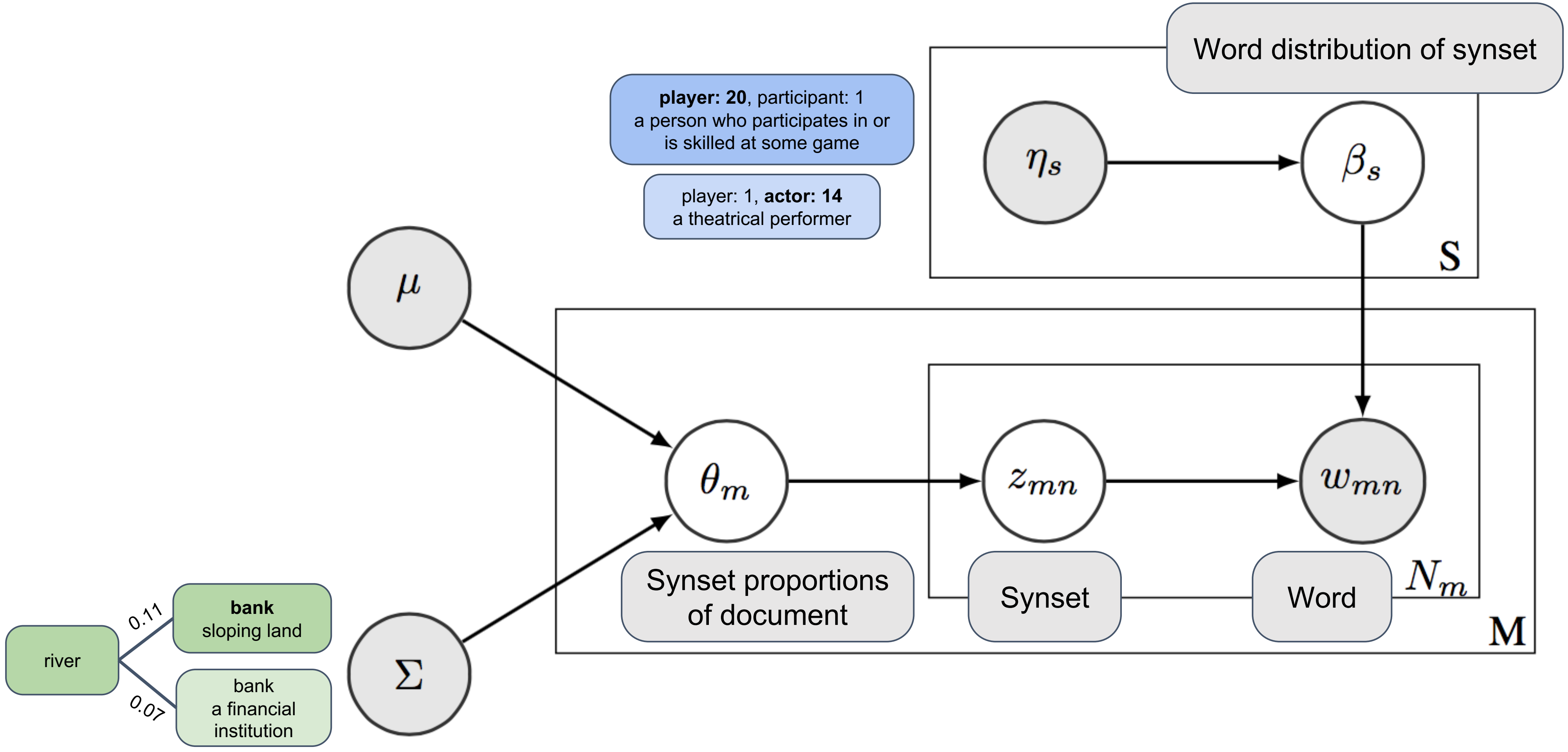


Carnegie Mellon University

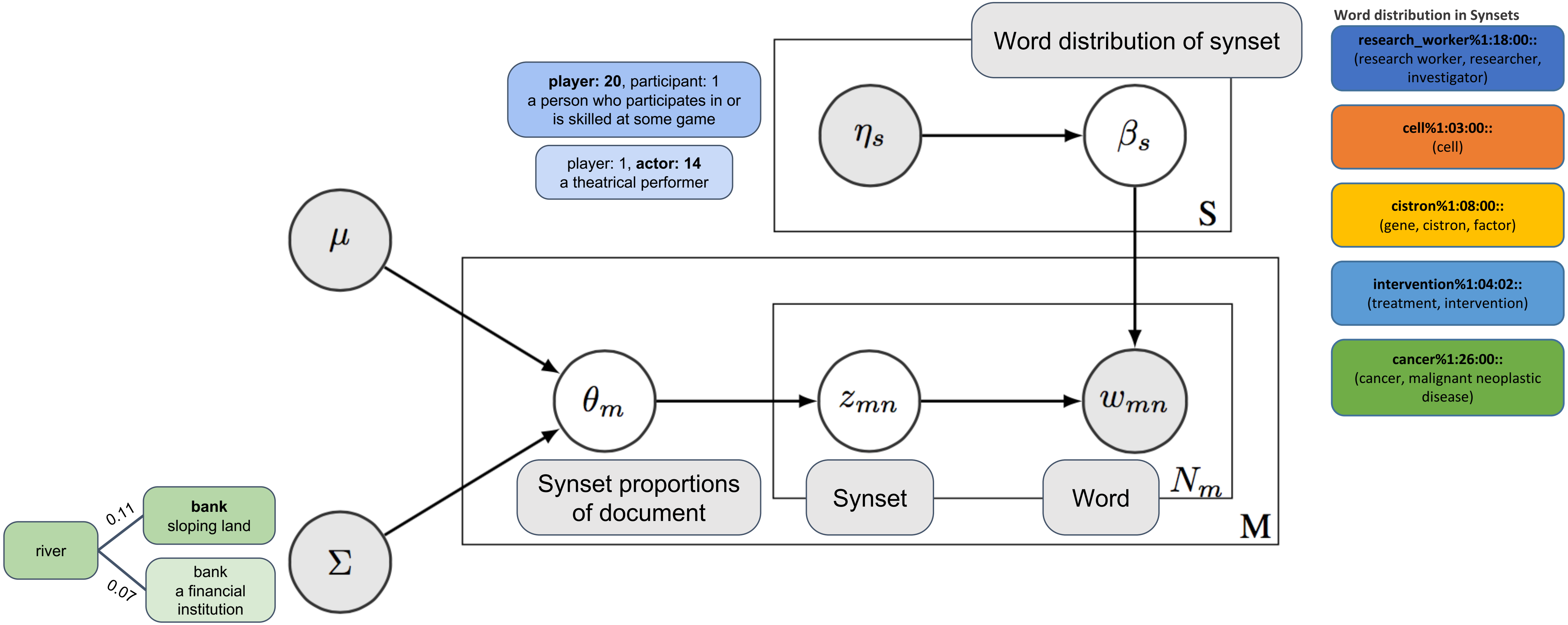
Method – Graphical Model



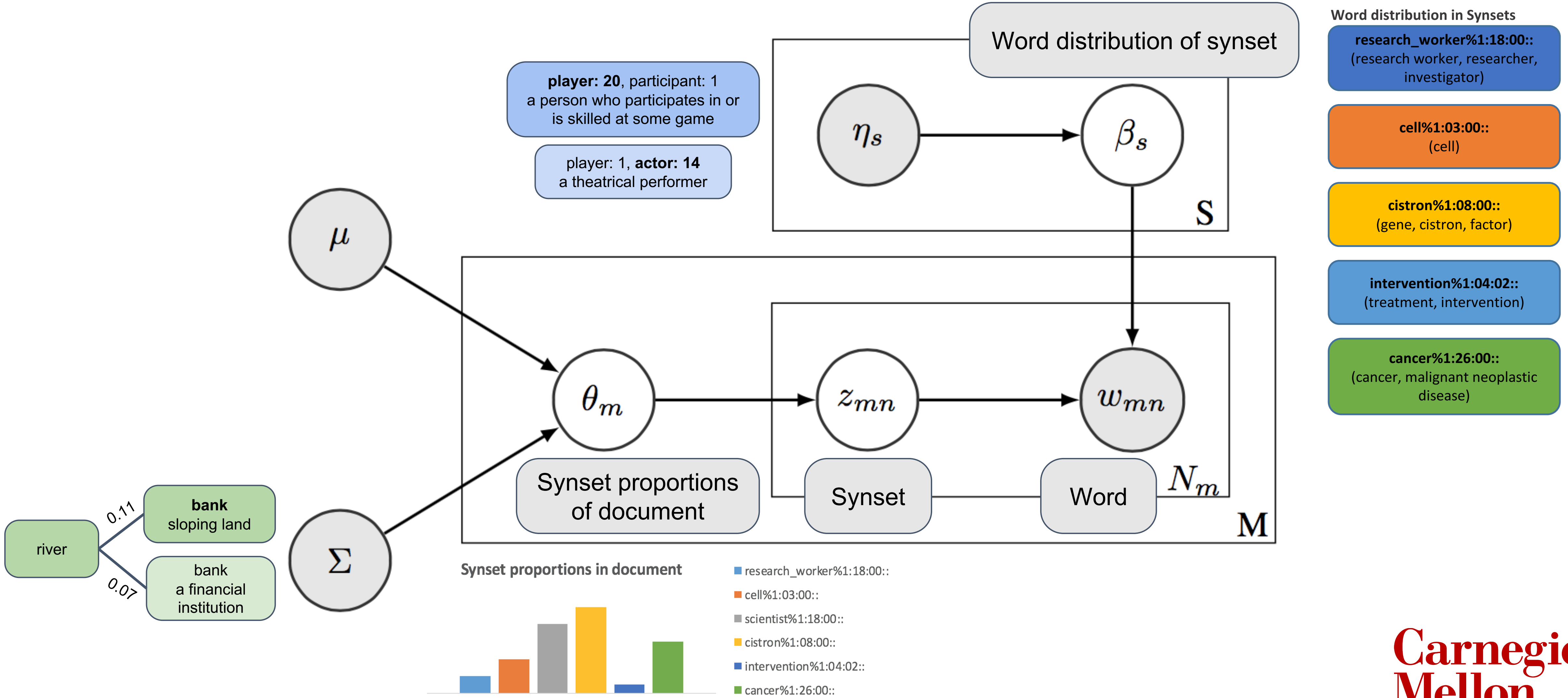
Method – Graphical Model



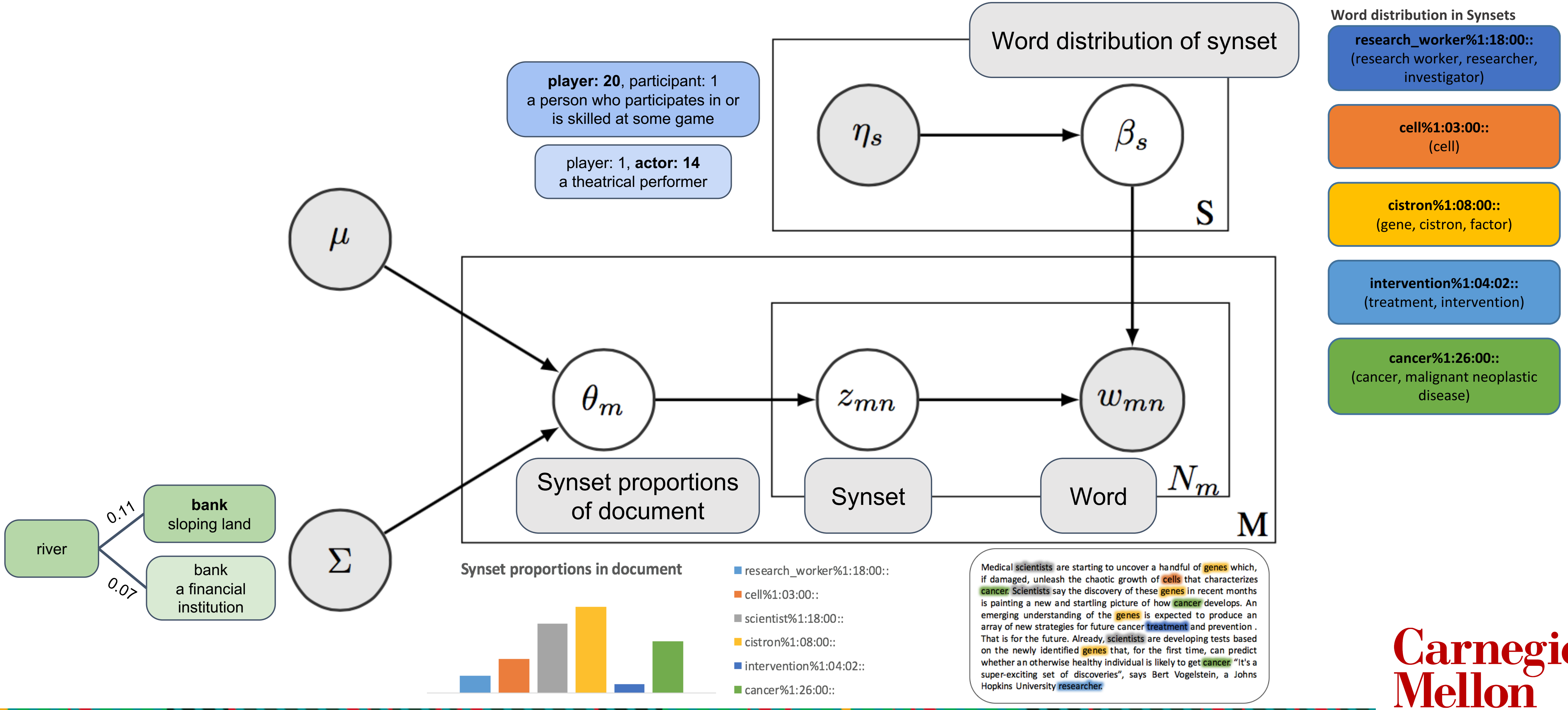
Method – Graphical Model



Method – Graphical Model



Method – Graphical Model



Method – Inference

- Used Gibbs sampler for inference
  - Document-specific word distribution can be collapsed by integrating out β parameters.
  - Document-specific sense distribution can't be integrated out but can be expressed in terms of inverse covariance matrix.

Carnegie Mellon University

Method – Inference

- Used Gibbs sampler for inference
  - Document-specific word distribution can be collapsed by integrating out β parameters.
  - Document-specific sense distribution can't be integrated out but can be expressed in terms of inverse covariance matrix.

$$p(z_{mn} = k|rest) \propto \frac{(\eta_{sv} + n_{sv\_mn}^{SV})}{n_{s\_mn}^{S} + ||\eta_s||_1} exp(\alpha_{mk})$$

$$\begin{aligned} n_{sv}^{SV} &= \sum_{m,n} \{z_{mn} = s, w_{mn} = v\} \\ n_{sm}^{SM} &= \sum_{n} \{z_{mn} = s\} \\ n_{s}^{S} &= \sum_{m} n_{sm}^{SM} \end{aligned}$$

Results

	System	Senseval-2	Senseval-3	SemEval-07	SemEval-13	SemEval-15	All
Knowledge based	Banerjee03	50.6	44.5	32.0	53.6	51.0	48.7
	Basile14	63.0	63.7	**56.7**	66.2	64.6	63.7
	Agirre14	60.6	54.1	42.0	59.0	61.2	57.5
	Moro14	67.0	63.5	51.6	**66.4**	**70.3**	65.5
	WSD-TM	**69.0**	**66.9**	55.6	65.3	69.6	**66.9**
Supervised	MFS	66.5	60.4	52.3	62.6	64.2	62.9
	Zhong10	70.8	68.9	58.5	66.3	69.7	68.3
	Melamud16	72.3	68.2	61.5	67.2	71.7	69.4

Comparison of F1 scores with various WSD systems on English all-words datasets of Senseval-2, Senseval-3, SemEval-2007, SemEval-2013, SemEval-2015. WSD-TM corresponds to the proposed method. The best results in each column among knowledge-based systems are marked in bold.

Results

	System	Nouns	Verbs	Adjectives	Adverbs	All
Knowledge based	Banerjee03	54.1	27.9	54.6	60.3	48.7
	Basile14	69.8	51.2	51.7	80.6	63.7
	Agirre14	62.1	38.3	66.8	66.2	57.5
	Moro14	68.6	49.9	73.2	79.8	65.5
	WSD-TM	**69.7**	**51.2**	**76.0**	**80.9**	**66.9**
Supervised	MFS	65.8	45.9	72.7	80.5	62.9
	Zhong10	71.0	53.3	77.1	82.7	68.3
	Melamud16	71.7	55.8	77.2	82.7	69.4

Comparison of F1 scores on different POS tags over all datasets. WSD-TM corresponds to the proposed method. The best results in each column among knowledge-based systems are marked in bold.

Comparison with Prior Work

Method	Word Frequencies (prior)	Sense relationships (contextual)	Key advantages
Random Walk (Agirre-14)	Static PageRank	Personalized PageRank	Utilizes WordNet as a graph
Markov Random Field (Chaplot-15)	Node Potentials	Edge Potentials	Joint Modeling, edge reduction
WSD-Topic Modeling	Non-uniform prior for word distribution of senses	Gaussian prior for sense distribution of documents	Joint Modeling, document context

Analysis

Scientists call the new class of **genes** tumor-suppressors, or simply anti-cancer genes. When functioning normally, they make proteins that hold a **cell's** growth in check. But if the **genes** are damaged -- perhaps by radiation, a chemical or through a chance accident in **cell** division -- their growth-suppressing proteins no longer work, and cells normally under control turn malignant . The newly identified **genes** differ from a family of genes discovered in the early 1980s called oncogenes. Oncogenes must be present for a **cell** to become malignant, but **researchers** have found them in normal as well as in cancerous **cells**, suggesting that oncogenes don't cause **cancer** by themselves. In recent months, researchers have come to believe the two types of **cancer genes** work in concert : An oncogene may turn proliferating **cells** malignant only after the tumor-suppressor **gene** has been damaged. Like all genes, tumor-suppressor genes are inherited in two copies, one from each parent. Either copy can make the **proteins** needed to control **cell** growth, so for **cancer** to arise, both copies must be impaired.

Analysis

- S: (n) **cell#1** (any small compartment) *"the cells of a honeycomb"*
- S: (n) **cell#2** ((biology) the basic structural and functional unit of all organisms; they may exist as independent units of life (as in monads) or may form colonies or tissues as in higher plants and animals)
- S: (n) **cell#3**, electric cell#1 (a device that delivers an electric current as the result of a chemical reaction)
- S: (n) **cell#4**, cadre#1 (a small unit serving as part of or as the nucleus of a larger political movement)
- S: (n) cellular telephone#1, cellular phone#1, cellphone#1, **cell#5**, mobile phone#1 (a hand-held mobile radiotelephone for use in an area divided into small sections, each with its own short-range transmitter/receiver)
- S: (n) **cell#6**, cubicle#1 (small room in which a monk or nun lives)
- S: (n) **cell#7**, jail cell#1, prison cell#1 (a room where a prisoner is kept)

Analysis

- S: (n) **cell#1** (any small compartment) *"the cells of a honeycomb"*
- S: (n) **cell#2** ((biology) the basic structural and functional unit of all organisms; they may exist as independent units of life (as in monads) or may form colonies or tissues as in higher plants and animals)
- S: (n) **cell#3**, electric cell#1 (a device that delivers an electric current as the result of a chemical reaction)
- S: (n) **cell#4**, cadre#1 (a small unit serving as part of or as the nucleus of a larger political movement)
- S: (n) cellular telephone#1, cellular phone#1, cellphone#1, **cell#5**, mobile phone#1 (a hand-held mobile radiotelephone for use in an area divided into small sections, each with its own short-range transmitter/receiver)
- S: (n) **cell#6**, cubicle#1 (small room in which a monk or nun lives)
- S: (n) **cell#7**, jail cell#1, prison cell#1 (a room where a prisoner is kept)

Analysis

The similarity of different senses of the word 'cell' with senses of three monosemous words 'scientist', 'researcher' and 'protein'. The correct sense of cell, 'cell#2', has the highest similarity with all the three synsets.

Sense of 'cell'	**Similarity with**		
	scientist#1	researcher#1	protein#1
cell#1	0.100	0.091	0.077
cell#2	**0.200**	**0.167**	**0.100**
cell#3	0.100	0.091	0.077
cell#4	0.100	0.062	0.071
cell#5	0.100	0.077	0.067
cell#6	0.100	0.091	0.077
cell#7	0.100	0.091	0.077

Advantages

- No need to specify the number of synsets
  - Major drawback of LDA is the need to specify number of topics
  - The number of synsets are fixed (equal to number of synsets in the sense repository)

Advantages

- No need to specify the number of synsets
  - Major drawback of LDA is the need to specify number of topics
  - The number of synsets are fixed (equal to number of synsets in the sense repository)
- Synsets are meaningful (topics need not be)
  - Prior for word distribution for each sense is not symmetric: contains equal non-zero entries for only the words contained in corresponding synset

Advantages

- No need to specify the number of synsets
  - Major drawback of LDA is the need to specify number of topics
  - The number of synsets are fixed (equal to number of synsets in the sense repository)
- Synsets are meaningful (topics need not be)
  - Prior for word distribution for each sense is not symmetric: contains equal non-zero entries for only the words contained in corresponding synset
- Leverage the formalism of LDA to model the whole document.
  - Impractical to model the whole document using existing methods as they scale exponentially with number of words in the context.
  - Sentence (~15 words), while document (~600-800 words).

Conclusion & Future Work

- Model the whole document as context using LDA. Incorporate knowledge using different priors.
- State-of-the-art results on WSD benchmark datasets.
- Possible extensions:
  - Adding an additional level in the hierarchy (topics)
  - Incorporating sense tags (using supervised topic models)
- Extending the model to other tasks such as Named-Entity Disambiguation

References

[1] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10.

[2] MALLERY, J. C. (1988). Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers. Ph.D. dissertation. MIT Political Science Department, Cambridge, MA.

[3] Agirre, E., de Lacalle, O. L., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. Computational Linguistics, 40(1), 57-84.

[4] Chaplot, D. S., Bhattacharyya, P., & Paranjape, A. (2015, January). Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser. In AAAI (pp. 2217-2223).

[5] Gale, W. A., Church, K. W., & Yarowsky, D. (1992, February). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language* (pp. 233-237). Association for Computational Linguistics.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of ma- chine Learning research* 3:993–1022.

[7] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity: Measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 38–41.

Knowledge-based Word Sense Disambiguation using Topic Models

Devendra Singh Chaplot, Ruslan Salakhutdinov

Thank you

Appendix

Inference (1)

$$p(\boldsymbol{z}, \boldsymbol{\alpha} | \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\propto \quad p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})\, p(\boldsymbol{\beta}|\boldsymbol{\eta})\, p(\boldsymbol{z}|\boldsymbol{\alpha})\, p(\boldsymbol{\alpha}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Document-specific word distribution can be collapsed by integrating out β parameters.

$$p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\eta}) = \prod_{s=1}^{S} \frac{\prod_v \Gamma(n_{sv}^{SV} + \eta_{sv})}{\Gamma(n_s^S + ||\eta_s||_1)} \frac{\Gamma(||\eta_s||_1)}{\prod_s \Gamma(\eta_{sv})}$$

Inference (2)

Document-specific sense distribution can't be integrated out but can be expressed in terms of inverse covariance matrix.

$$\begin{aligned} p(z_{mn} = k|rest) &= \frac{p(z, w|\alpha, \eta)}{p(z_{-mn}, w|\alpha, \eta)} \\ &\propto p(z, w|\alpha, \eta) \\ &\propto \frac{(\eta_{sv} + n^{SV}_{sv_{-mn}})}{n^S_{s_{-mn}} + ||\eta_s||_1} \exp(\alpha^k_m) \end{aligned}$$

$$\begin{aligned} n^{SV}_{sv} &= \sum_{m,n} \{z_{mn} = s, w_{mn} = v\} \\ n^{SM}_{sm} &= \sum_{n} \{z_{mn} = s\} \\ n^S_s &= \sum_m n^{SM}_{sm} \end{aligned}$$

$$p(\boldsymbol{z}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} \left(\prod_{n=1}^{N_m} \frac{\exp(\alpha^{z_{mn}}_m)}{\sum_{s=1}^{S} \exp(\alpha^s_m)} \right)$$

$$p(\boldsymbol{\alpha_m}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\boldsymbol{\alpha_m}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$