A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation using First-Order Logic

David Andrzejewski (LLNL)
Xiaojin Zhu (Wisconsin)
Mark Craven (Wisconsin)
Benjamin Recht (Wisconsin)

Lawrence Livermore National Laboratory
Livermore, CA (USA)

University of Wisconsin-Madison
Madison, WI (USA)

Summary

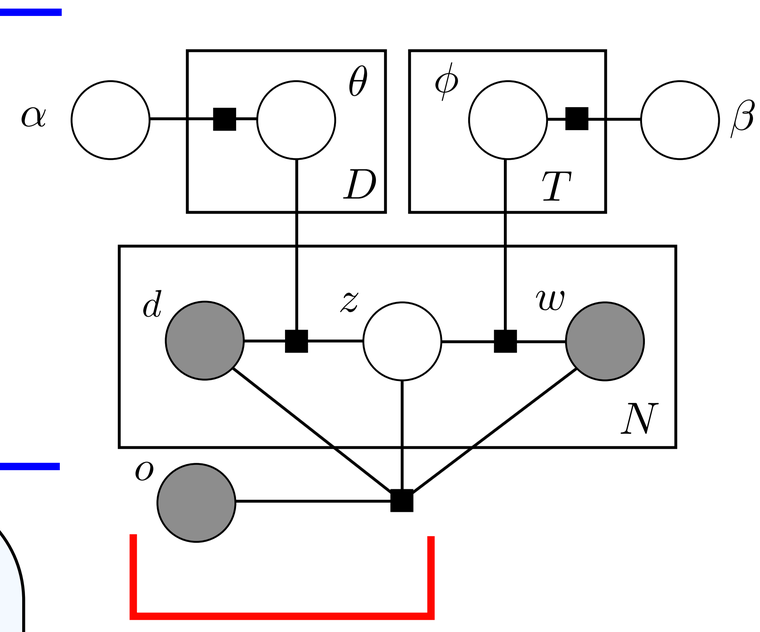
Latent topic models have emerged as a versatile tool for data exploration. Researchers often extend the base Latent Dirichlet Allocation (LDA) [1] model in order to capture domain knowledge or side information relevant to a particular application, but constructing these extensions is non-trivial. We propose a general framework for encoding domain knowledge as First-Order Logic (FOL), allowing users to adapt topic modeling to their needs.

The Model

First-Order Logic latent Dirichlet ALLocation(Fold-all)

Latent Dirichlet Allocation (LDA)

Topic-word	$\phi_z(w) = P(w\|z)$
Doc weights	$\theta_d(z) = P(z\|d)$
Topic assign	$\mathbf{z} = z_1 \dots z_N$

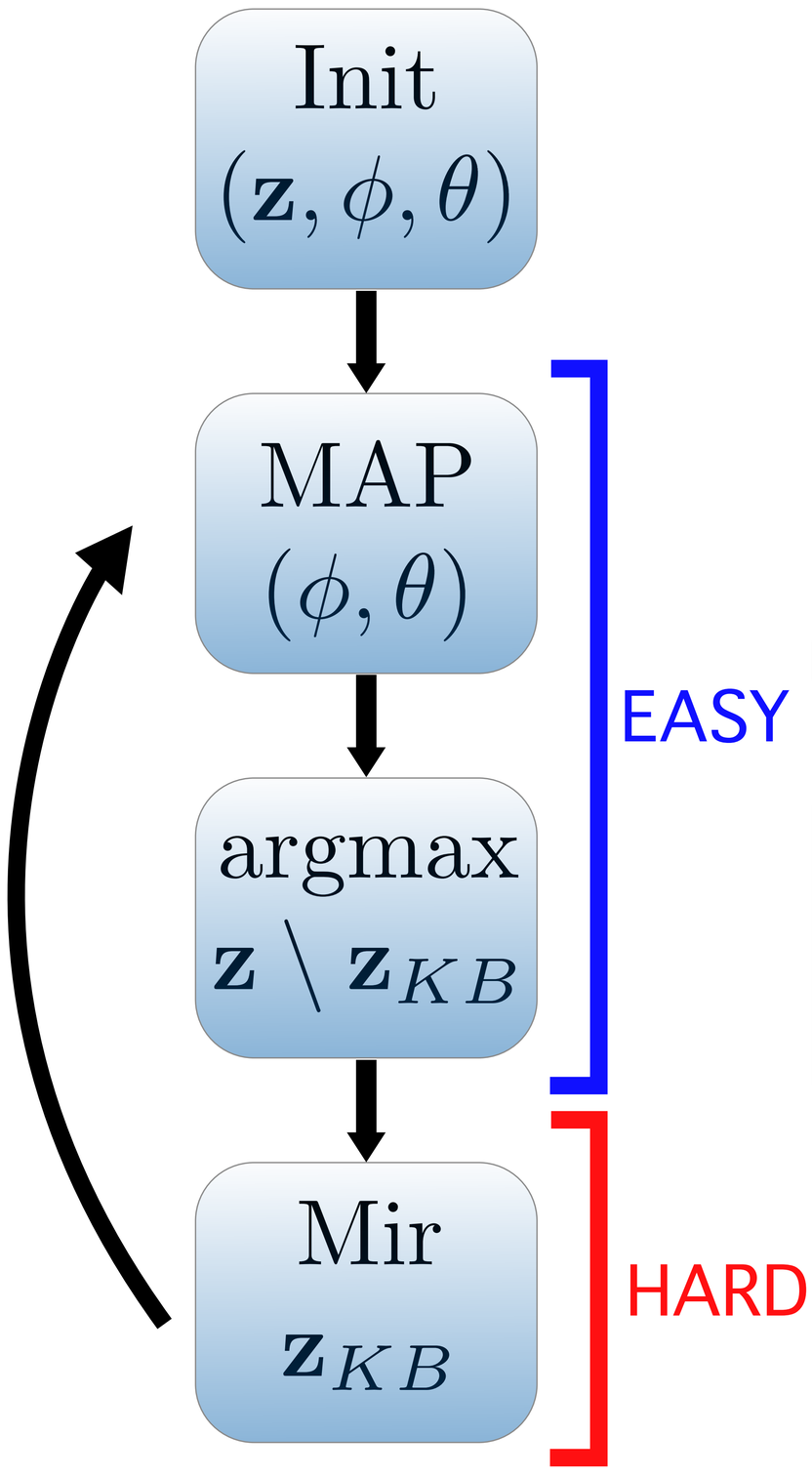


Side Information and Logic Factor (MLN)

Rule	ψ_ℓ
Weight	λ_ℓ
Groundings	$G(\psi_\ell)$
Indicator	$\delta_g(\mathbf{z}, \mathbf{w}, \mathbf{d}, \mathbf{o})$

The user encodes domain knowledge into a knowledge base (KB) of weighted logic rules. As in Markov Logic Networks (MLNs) [2], these rules are then grounded, and each grounding is associated with a potential function. The learned topics will then reflect both the document-word statistics (as in LDA) and the user-defined rules (as in MLNs). This model can be considered an instance of a Hybrid MLN [3], specialized for topic modeling.

Scalable MAP Inference



The combinatorial explosion in rule groundings is a challenge, as in general MLNs. We use a continuous relaxation scheme along with random sampling of groundings to do stochastic Mirror Descent (Mir) over the latent topic assignments affected by logic rules.

Biological Concept Expansion (Human Development Genes)

A biological expert is interested in expanding a set of seed terms related to human developmental biology. The table below shows blind relevance accuracy judgments on the Top 50 recovered words for both standard LDA and several different KBs. Exploiting expert knowledge and sentence annotations (FULL-KB) leads to improved results.

Given	neuro dendro(cyte), glia, synapse, neural crest
LDA	disease, disorders, schizophrenia, syndrome, abnormal
SEED	myelin, astrocytes, oligodendrocytes
FULL KB	dendritic, forebrain, hindbrain, microglial, motoneurons, neuroblasts, neurogenesis, retinal

	KBs				
	FULL-KB	INCL	EXCL	SEED	LDA
Neural	**0.59**	**0.57**	**0.54**	**0.54**	0.31
Embryo	**0.24**	**0.24**	**0.23**	**0.23**	0.07
Blood	**0.46**	**0.47**	0.40	0.39	0.13
Gast.	**0.18**	**0.18**	**0.16**	**0.16**	0.00
Cardiac	**0.36**	**0.37**	**0.34**	**0.35**	0.08
Limb	**0.18**	**0.18**	**0.15**	**0.14**	0.09

References

[1] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research (JMLR) 3 (Mar. 2003), 993-1022.

[2] Domingos, P. and Lowd, D. (2009). Markov Logic: An Interface Layer for Artificial Intelligence. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan and Claypool Publishers.

[3] Wang, J. and Domingos, P. (2008) Hybrid Markov Logic Networks. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 1106--1111. AAAI Press.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-CONF-466413), with additional support from NSF IIS-0953219, AFOSR FA9550-09-1-0313, and NIH/NLM R01 LM07050. We would like to thank Ron Stewart for his participation in the HDG experiments. LLNL-POST-488374

Source Code

https://github.com/davidandrzej/LogicLDA

Example Rules

Rule Type	Logical form	Meaning
Seed word	$\mathtt{W}(i, embryo) \Rightarrow \mathtt{Z}(i, 3)$	Assign "embryo" to topic 3
Doc label	$\mathtt{D}(i, j) \wedge \mathtt{HasLabel}(j, +) \Rightarrow \neg\mathtt{Z}(i, 3)$	Do not use topic 3 in docs with label +
Cannot-link	$\mathtt{W}(i, neural) \wedge \mathtt{W}(j, disorder) \Rightarrow \neg\mathtt{Z}(i, t) \vee \neg\mathtt{Z}(j, t)$	Do not assign "neural" and "disorder" to the same topic
Must-link	$\mathtt{W}(i, neural) \wedge \mathtt{W}(j, cortex) \Rightarrow \neg\mathtt{Z}(i, t) \vee \mathtt{Z}(j, t)$	Assign "neural" and "cortex" to the same topic
Bigram (1)	$\mathtt{W}(i, stem) \wedge \mathtt{W}(i+1, cell) \Rightarrow \mathtt{Z}(i, 1)$	Assign "stem" in "stem cell" to topic 1
Bigram (2)	$\mathtt{W}(i-1, stem) \wedge \mathtt{W}(i, cell) \Rightarrow \mathtt{Z}(i, 1)$	Assign "cell" in "stem cell" to topic 1
Sent Incl	$\mathtt{Sentence}(i, i_1, \dots) \wedge \neg\mathtt{Z}(i_1, 6) \wedge \dots \Rightarrow \neg\mathtt{Z}(i, 0)$	Topic 6 is required for topic 0 to occur in the same sentence
Sent Excl	$\mathtt{S}(i, s) \wedge \mathtt{S}(j, s) \wedge \mathtt{Z}(i, 7) \Rightarrow \neg\mathtt{Z}(j, 0)$	Topic 7 prevents topic 0 from occurring in the same sentence