

Modeling concept dependencies in a scientific corpus

JONATHAN GORDON, LINHONG ZHU, ARAM GALSTYAN, PREM NATARAJAN, GULLY BURNS

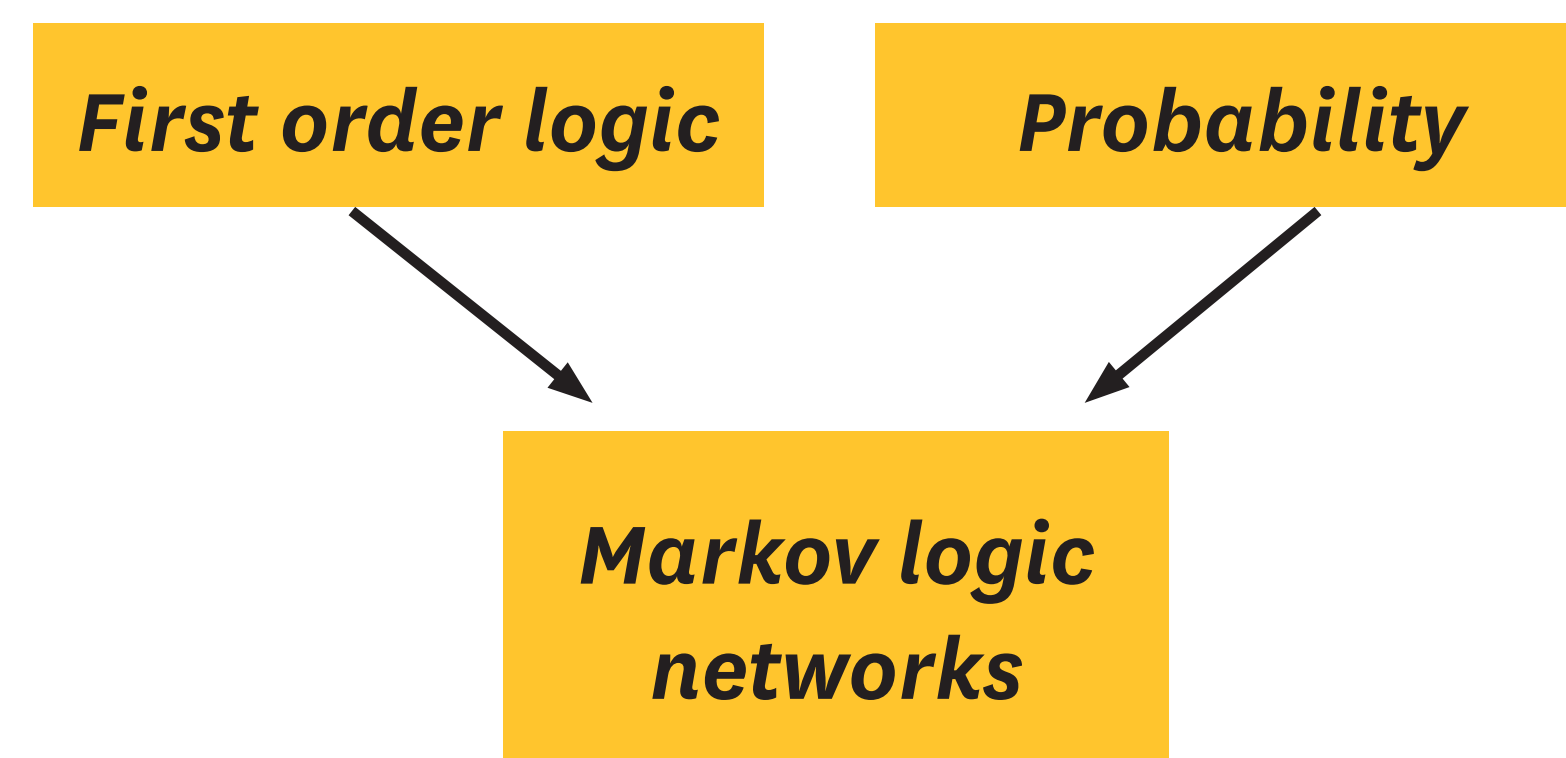
Motivation We want to help you learn scientific & technical concepts. How could we generate a reading list like an expert does?

We need to discover what concepts help you to understand others.

“What do I need to know about before I start reading about ... Markov logic networks?”

Approach From a scientific corpus, generate a **concept graph**. Concepts can be learned using latent Dirichlet allocation (LDA).

Each document is linked to the concepts it discusses, and each concept is linked to the other concepts it depends upon.



Data: ACL Anthology We use the text of papers from the ACL Anthology Network (2013), with additional automatic and manual text clean up.

We inferred a 300-topic Mallet LDA topic model over the bigrams.

Relevant phrases:
machine translation, translation system, mt system, transfer rules, mt systems, lexical transfer, analysis transfer, translation process, transfer generation, transfer component, analysis synthesis, transfer phase, analysis generation, structural transfer, transfer approach, human translation, transfer grammar, analysis phase, translation systems, transfer process

Human Evaluation We sampled 285 pairs of concepts, biased for coverage of the strongest and weakest dependencies predicted by the different methods.

They were evaluated by 8 human judges with varying levels of domain expertise. Agreement was higher for domain experts, but in all cases moderate.

Judges are asked:

Would Topic 1 help you to understand Topic 2?

Would Topic 2 help you to understand Topic 1?

– *I don't know*

– *Not at all*

– *Somewhat*

– *Very much*

Information-theoretic measures

Cross-entropy

If most instances of concept c_i can be explained by occurrences of concept c_j – but not vice versa – predict c_i depends on c_j .

Information flow

Predict c_i depends on c_j if c_i receives less navigation traffic than c_j and the traffic from c_i to c_j is stronger than that to another non-dependent concept c_k .

Use random walks over the concept co-occurrence graph to approximate human navigation.

Baseline methods

Word similarity

More similar concepts are more likely to be connected by dependency relations. We compute the Jaccard similarity coefficient based on the top-20 words in a concept's word distribution.

Hierarchy

How close does identifying hierarchical relations come to identifying dependency relations? We tried agglomerative clustering over the concept co-occurrence graph.

Citation-based

If documents that are highly related to c_j are cited by most instances of c_i , c_i may depend on c_j . We adapt the method of Wang et al. (2013).

Most of the time, documents that mention *MLNs* also mention *Probability*. The reverse is not true. **So, maybe *MLNs* depends on *Probability*!**

	Top 20	Top 150			All scores > 0		
	Prec.	Prec.	Rec.	f1	Prec.	Rec.	f1
Cross entropy	0.851	0.765	0.358	0.487	0.693	0.670	0.681
Information flow	0.793	0.696	0.311	0.429	0.693	0.323	0.441
Word similarity	0.808	0.768	0.382	0.511	0.768	0.382	0.511
Hierarchy	0.680	0.692	0.297	0.416	0.686	0.638	0.661
Cite	0.693	0.718	0.343	0.465	0.693	0.670	0.681
Random	0.659	0.661	0.580	0.500	0.658	1.000	0.794

Results

The results verify the feasibility of automatic approaches for inferring concepts and their dependencies.

Word similarity is a strong baseline, but when we compare the strongest edges predicted by each method (Top 20 in table), the cross-entropy method is most precise.

Annotations, concept graphs, and implementations: <http://techknacq.isi.edu>