Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

A Unified Bayesian Model of Scripts, Frames and Language

Francis Ferraro[1] **Benjamin Van Durme**[1,2]
[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
Johns Hopkins University

Abstract

We present the first probabilistic model to capture all levels of the Minsky Frame structure, with the goal of corpus-based induction of scenario definitions. Our model unifies prior efforts in discourse-level modeling with that of Fillmore's related notion of frame, as captured in sentence-level, FrameNet semantic parses; as part of this, we resurrect the coupling among Minsky's frames, Schank's scripts and Fillmore's frames, as originally laid out by those authors. Empirically, our approach yields improved scenario representations, reflected quantitatively in lower surprisal and more coherent latent scenarios.

Introduction

Frames or scripts describe prototypically complex situations in terms of certain events, actions, actors and other pieces of information we expect to be involved. These theories posit that for many situations we encounter, there is a **template** with a number of **slots** that need to be filled in order to understand the situation. For example, we partially describe a BOMBING situation with a `Detonation` action, along with those involved, e.g., BOMBERS and VICTIMS.

Corpus statistics can be used to induce approximate, probabilistic versions of these templates using verb cooccurrences and automatically generated syntactic dependency parses (Cheung, Poon, and Vanderwende 2013; Bamman, O'Connor, and Smith 2013; Chambers 2013, i.a.). These parses can serve as a limited proxy for sentence meaning, owing to information conveyed via the syntax/semantics interface. They do not however fully (explicitly) represent a semantic analysis (Rudinger and Van Durme 2014).

Fillmore's notion of frame semantics ties a notion akin to Minsky's frames to individual *lexical items* (Fillmore 1976; 1982). Word meaning is defined in terms of the roles words play in situations they typically *invoke*, and in how they interact with other lexical items.

In the following we present a probabilistic model which unifies discourse-level Minskian frames with Fillmore's frame semantics. Despite the historical and intellectual connections between these theories, previous empirical efforts have focused on just one or the other: this model is the first to make the connection explicit. We show how current efforts in discourse modeling, and semantic frame induction and identification can be combined in a single model to capture what classic AI theory posited. Quantitatively, by using a frame-semantic parser pre-trained on FrameNet (Baker, Fillmore, and Lowe 1998), we show that incorporating frame information provides both a better fit to held-out data and improved coherence (Mimno et al. 2011). Our unified probabilistic model provides a principled mathematical way of restating Minsky's argument for the four frame levels, and our results show that our model is a legitimate way to capture what Minsky proposed.

Frames Background

Classic Theories

Minsky, along with a number of contemporaries, believed in schematizing common situations and experiences into *"chunks"*, or *frames*. These frames contain world knowledge that would allow artificial intelligence systems to encounter various occurrences and react appropriately. For Minsky, frames were data structures, with *slots*, to "[represent] a stereotyped situation." Some slots and conditions could have default values; entities (references to an "object" in the world) and pointers to other frames could fill slots.

Minsky (1974) outlined four different "levels" of frames:

Surface Syntactic Frames *"Mainly verb and noun structures. Prepositional and word-order indicator conventions."*

Surface Semantic Frames *"Action-centered meanings of words. Qualifiers and relations concerning participants, instruments, trajectories and strategies, goals, consequences and side-effects."*

Thematic Frames *"Scenarios concerned with topics, activities, portraits, setting."*

Narrative Frames *"Skeleton forms for typical stories, explanations, and arguments. Conventions about foci, protagonists, plot forms, development, etc., designed to help*

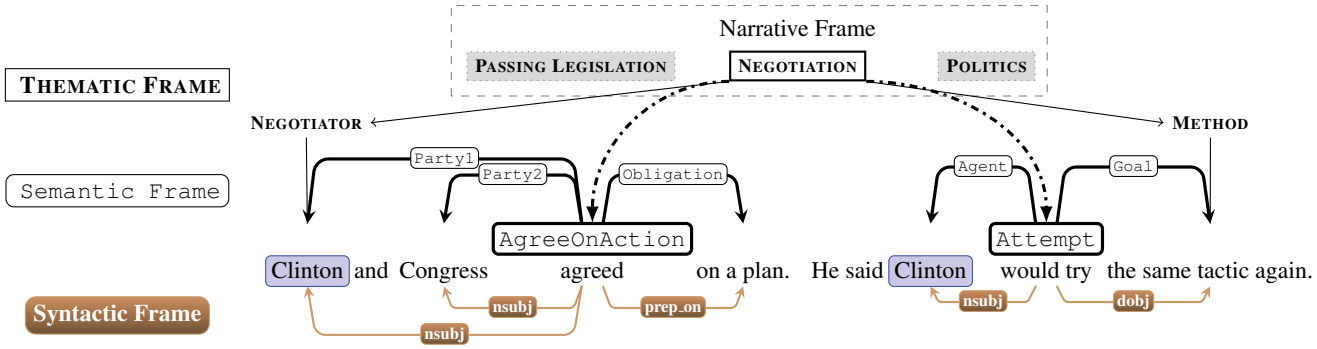


Figure 1: An interpretation of Minsky's four frame levels on two newswire sentences. The syntactic (below) and surface semantic (above) frames provide the lowest-level intrasentential analyses of this abbreviated document (Latin font). The NEGOTIATION template (thematic frame) fills two of its *slots*, NEGOTIATOR and METHOD, intersententially, with "Clinton" and "tactic," using predicate and dependency information from some combination of the syntactic and surface semantic frames. Here, "Clinton" is highlighted twice stressing that thematic frames may both produce and rely on information across sentences. The narrative frame invokes the NEGOTIATION thematic frame, though related themes PASSING LEGISLATION and POLITICS may appear elsewhere in the document. Our model follows this interpretation. (Adapted from the automatically labeled version of NYT_ENG_19980330.0346 in Ferraro et al. (2014).)

> *a listener construct a new, instantiated Thematic Frame in his own mind."*

Figure 1 illustrates an interpretation of these four levels on newswire automatically tagged with syntactic and semantic frames, and example thematic and narrative frames.

These hierarchical levels require attention to different aspects of language; as one changes levels, details highly relevant to one may become "displaced" by more appropriate aspects of another. Information important for the syntactic level may be relevant to, e.g., the thematic or narrative level through an abstracted or "coarsened" version. We assume the lower-level syntactic and surface semantic frames are localized analyses, restricted to sentences, while the higher-level thematic and narrative frames allow for a global analysis, aggregating information across sentences.

While many people are familiar with Schank and Abelson (1977)'s formulations of scripts, the connection between frames and scripts is at times forgotten:

> ... a frame is a general name for a class of knowledge organizing techniques that guide and enable understanding. Two types of frames that are necessary are SCRIPTS and PLANS. Scripts and plans are used to understand and generate stories and actions – Schank (1975).

Schankian scripts are thus a distinct sub-type of Minskian frames. Broadly, scripts introduce a mechanism for ordering events within frames. For simplicity our model does not encode order, though it provides a framework for future efforts to incorporate ordering, perhaps utilizing some prior ordering efforts. We discuss this later on.

Fillmore's case grammar and frame semantics (Fillmore 1967; 1976; 1982) posit that word meaning is defined in terms of the roles they play in situations they typically *invoke*, and then in how they interact with other lexical items. We can think of Fillmore as being 'Minsky over words,' where Fillmore's ideas can be realized within the broader development of frames during the 1970s:

> [frames are] certain schemata or frameworks of concepts or terms which link together as a system, which impose structure or coherence on some aspect of human experience, and which may contain elements which are simultaneously parts of other such frameworks. – Fillmore (1975).

The FrameNet Project (Baker, Fillmore, and Lowe 1998) is an ongoing effort to implement Fillmore's frames.

Contemporary Efforts

There have been various styles of models in the spirit of this work, though none capture all four levels of the Minsky hierarchy. The most similar are the three concurrent Bayesian template models (Bamman, O'Connor, and Smith 2013; Chambers 2013; Cheung, Poon, and Vanderwende 2013). Like this work, the former two view documents as collections of prespecified entities and mentions. They similarly incorporate narrative, thematic and syntactic levels, as documents are modeled as mixtures over templates relying on syntactic information. Subsequent work from Bamman and colleagues has refined event participant descriptions or ascribing temporal attributes to atomic events, rather than exploring hierarchical event substructure, as we do (Bamman, Underwood, and Smith 2014; Bamman and Smith 2014). None of these efforts have incorporated separate semantic and syntactic Minskian frames.

Cheung, Poon, and Vanderwende (2013) model ordering of syntactic clauses, grouping predicates into latent events, and a predicate's arguments to event slots. A latent "frame" assignment stratifies templates more coherently across the clauses and throughout the document. In the Minskian terminology used here, they have two layers of thematic frame, but, as above, no layer of semantic frame.

A number of other efforts in learning semantic frames consider syntactic information, though there has not been a presentation incorporating both narrative and thematic components (Titov and Klementiev 2011; Materna 2013; Modi, Titov, and Klementiev 2012; Lorenzo and Cerisara 2012; Bejan 2008; Modi and Titov 2014). Temporal scripts have been learned with graph algorithms (Regneri, Koller, and Pinkal 2010), Bayesian model merging (Orr et al. 2014), and permutation priors (Frermann, Titov, and Pinkal 2014), i.a.. These incorporate a rich narrative level, though without thematic frames: the narrative level deals directly with the semantic or syntactic frames.

While other efforts have focused on both generative and discriminative models for less-than-supervised frame induction (Minkov and Zettlemoyer 2012; Huang and Riloff 2013; Patwardhan and Riloff 2009, i.a.), of particular note are those incorporating event "triggers," reminiscent of Rosenfeld's trigger language models (Rosenfeld 1994; 1996; Van Durme and Lall 2009). Some of those efforts have identified which verbs trigger events (Chen et al. 2011, working between the syntactic and semantic levels), while others have focused on discourse relation (Maslennikov and Chua 2007, working between the narrative and syntactic levels).

Multiple efforts have formulated global (document-level) and local (sentence-level) constraints for supervised graphical models. Reichart and Barzilay (2012)'s factor graph with global and local potentials presents an extensive narrative level that incorporates both thematic and syntactic levels, but excludes the semantic. Both Liao and Grishman (2010) and Li, Ji, and Huang (2013) encode the thematic, semantic and syntactic levels, but no narrative level.

The Penn Discourse Treebank (Prasad et al. 2008, PDTB) provides both explicit and implicit discourse and causality annotations atop original syntactic annotations of the WSJ portion of the Penn Treebank. As PDTB annotations are both cross-sentential and intrasentential discourse relations, we can view the PDTB as a type of thematic frame. Although with some additional effort Minsky's surface semantic frames could be incorporated—e.g., by aligning PDTB with shallow semantic annotations, such as from PropBank—the narrative level is missing.

Unlabeled Induction with Frames

A contribution of our work is to rekindle the joint notion of a "frame" shared among Minsky, Fillmore and Schank, and position it within state-of-the-art probabilistic modeling. We combine high-performing tools in NLP with a large collection of documents in order to induce a probabilistic version of what Minsky, broadly, called scenario definitions.

Our model, detailed formally in Figure 3 and informally in Figure 1, captures the *"ingredients"* of a frame structure at all frame levels posited by Minsky (1974): Surface Syntactic (syntactic dependencies), Surface Semantic (FrameNet semantic parses), Thematic (templates), and Narrative (document-level mixtures over templates). Prior work has either conflated multiple levels together, or otherwise ignored levels entirely: inclusion of these levels as distinct model components is novel to this work.

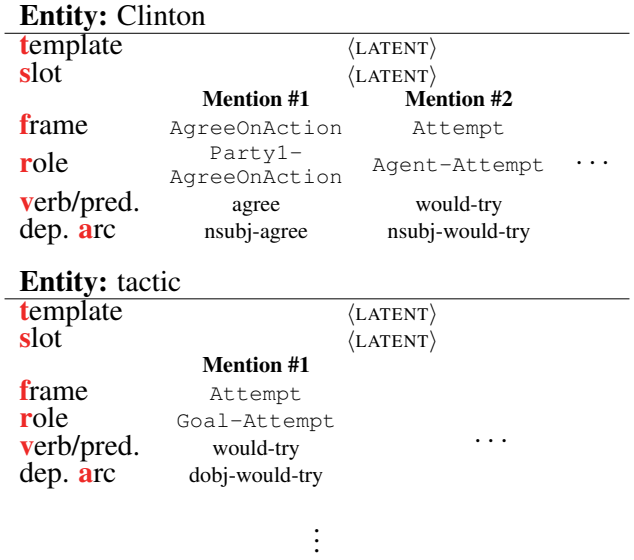


Figure 2: A view of the observed semantic and syntactic levels, as well as the latent thematic level, on the example document in Figure 1. Notice how entities do not have to be animate. The highlighted variables (**t**, **s**, **f**, **r**, **v** and **a**) correspond to those in Figure 3.

Available Observations

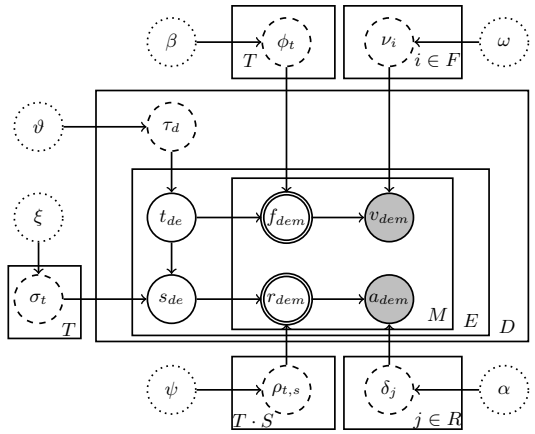
Following recent efforts we assume that both coreference resolution and a syntactic analysis have been performed on our documents as part of corpus processing. To learn a model, we assume an automatically produced semantic frame analysis, such as from FrameNet, too; we treat this as latent during heldout evaluation.

A document is a bag of entities (coreference chains), with each entity having one or more mentions. Each entity mention is syntactically governed through a typed dependency arc (a) to a verb lemma (v). Each verb evokes a surface semantic frame (f), which is related to the entity mention through a frame role (r). Like many other research efforts, we observe syntax, but assume that syntactic dependencies should be predicate specific. Beyond linguistic arguments for this, we, like Chambers (2013), have found results to be more human interpretable when r and a are typed by their corresponding frame or verb (Ruppenhofer et al. 2006, § 3.2). If a mention is a subject of "attempt," we set its arc a to `nsubj-attempt`, rather than just `nsubj`.

We observe at most the syntactic and semantic levels. The thematic and narrative levels are latent and are handled by our generative model. Figure 2 demonstrates this on a portion of the Figure 1 document.

Generative Story

Our observations and latent assignments are discrete; we place conjugate Dirichlet priors with symmetric hyperparameters on each. See Figure 3 for a formal diagram and variable gloss table. The narrative frame of a document d is represented as a mixture over the set of templates T (Minsky's thematic frames), $\tau_d \sim \text{Dir}(\vartheta)$.



(a) Shaded nodes are always observed, while double-edged nodes may or may not be; all others are latent. Solid-edged nodes such as t_{de} have collapsed priors (dashed edges, e.g.: τ_d) with optimized hyperparameters (dotted edges, e.g.: ϑ).

Variable	Meaning	Minsky
τ_d	document's dist. of templates (themes)	Narrative
σ_t	dist. of template-specific slots	Thematic
ϕ_t	dist. of template-specific semantic frames	Semantic
$\rho_{t,s}$	dist. of slot-specific semantic roles	Semantic
ν_i	dist. of semantic frame's syntactic realization	Syntactic
δ_j	dist. of semantic role's syntactic realization	Syntactic
$t_{d,e}$	template of entity e	Thematic
$s_{d,e}$	template-specific slot of entity	Thematic
$f_{d,e,m}$	semantic frame governing mention	Semantic
$r_{d,e,m}$	mention's semantic role	Semantic
$v_{d,e,m}$	governing predicate of mention	Syntactic
$a_{d,e,m}$	predicate-typed dependency of mention	Syntactic

(b) Brief meaning gloss of the model's variables, with the corresponding Minsky frame levels, given a document d, each of its corefence chains e, and each mention m of e. For simplicity, the hyperparameters (the dotted nodes in Figure 3a) are omitted.

Figure 3: Our unified probabilistic model.

Each template t, such as representing NEGOTIATION, is represented by a distribution σ_t over S unique slots, such as the NEGOTIATOR, and a distribution ϕ_t over F semantic frames (which will come from FrameNet). Both sets of distributions have Dirichlet priors, $\sigma_t \sim \text{Dir}(\xi)$, $\phi_t \sim \text{Dir}(\beta)$.

Every semantic frame i has a distribution over verb lemmas, $\nu_i \sim \text{Dir}(\omega)$, and each slot has a distribution $\rho_{t,s}$ over R frame roles, $\rho_{t,s} \sim \text{Dir}(\psi)$. Just as every semantic frame has a distribution over verb lemmas, every role j has a distribution over syntactic relations $\delta_j \sim \text{Dir}(\alpha)$.

An entity e is assigned to a single (latent) template $t_{d,e}$ and slot $s_{d,e}$, where $t_{d,e} \sim \text{Cat}(\tau_d)$ and $s_{d,e} \sim \text{Cat}\left(\sigma_{t_{d,e}}\right)$. For every mention m of e, the entity template $t_{d,e}$ directly influences the selection of the mention's frame assignment $f_{d,e,m} \sim \text{Cat}\left(\phi_{t_{d,e}}\right)$, and the slot $s_{d,e}$ directly influences the frame role $r_{d,e,m} \sim \text{Cat}\left(\rho_{s_{d,e}}\right)$. For instance, in Figure 2 we could replace Clinton's ⟨LATENT⟩ template and slot values with NEGOTIATION and NEGOTIATOR, respectively. The semantic frames `AgreeOnAction` and `Attempt` would both be attributed to the NEGOTIATION template, while the corresponding roles would be attributed to the NEGOTIATION-*specific* slot NEGOTIATOR.

Finally, the syntactic verb and syntactic relation surface forms are chosen given the frame and role, respectively: $v_{d,e,m} \sim \text{Cat}\left(\nu_{f_{d,e,m}}\right)$, and $a_{d,e,m} \sim \text{Cat}\left(\delta_{r_{d,e,m}}\right)$. For instance, in Figure 2, "agree" is attributed to `AgreeOnAction` and "nsubj-agree" is attributed to the typed semantic role `Party1-AgreeOnAction`.

Model Discussion Recall that we observe *typed dependencies*: for the syntactic subject of the verb "attempt,' the dependency is "nsubj-attempt." However, our model views these as separate observations without any direct (statistical) influence between the two. In the past, this typed predicate/dependency coupling has not been modeled directly (Chambers 2013; Cheung, Poon, and Vanderwende 2013); it is an open question how to decouple the dependencies and verb predicates and still model the motivating intuition at scale. While Lorenzo and Cerisara (2012) use separate distributions for each verb and Bamman and Smith (2014) use an exponential family parametrization, we operate at different scales: Lorenzo and Cerisara use fewer verb types, while Bamman and Smith use a significantly reduced relation set.

Inference

We fit the model via Gibbs sampling, collapsing out the priors on all latent and observed variables and optimizing the hyperparameters with fixed-point iteration (Wallach 2008). Posterior inference follows Griffiths and Steyvers (2004). We derive the complete conditionals of the template variables, with the respective priors integrated out; the calculations for slot, frame and role variables are similar.

Deriving the Complete Conditionals In general, for a set of conditionally i.i.d. Categorical variables $z_i|\theta \sim \text{Cat}(\theta)$, where θ has a $\text{Dir}(\alpha)$ prior, the joint probability of all $\mathbf{z}$ is given by the Dirichlet-Multinomial compound distribution $\text{DMC}(\mathbf{z}|\alpha)$:

$$p_\alpha(\mathbf{z}) = \int_\theta p(\mathbf{z}|\theta) p_\alpha(\theta) d\theta \tag{1}$$

$$= \frac{\Gamma\left(\sum_k \alpha_k\right)}{\Gamma\left(\sum_k \left(c(k) + \alpha_k\right)\right)} \prod_k \frac{\Gamma\left(c(k) + \alpha_k\right)}{\Gamma\left(\alpha_k\right)} \tag{2}$$

$$= \text{DMC}(\mathbf{z}|\alpha) \tag{3}$$

where $c(k)$ is the number of z_i with value k. This can be generalized to a gated version: given a collection of i.i.d. M Dirichlet samples $\theta_m \sim \text{Dir}(\alpha)$ and indicator variables y_i, if $z_i|y_i, \theta \overset{i.i.d}{\sim} \text{Cat}(\theta_{y_i})$, then we may consider the collection $[\mathbf{z}]_{\mathbf{y}=m}$ — only those z_i such that $y_i = m$. Then

$$p_\alpha(\mathbf{z}; \mathbf{y}) = \prod_{m=1}^{M} \left(\text{DMC}\left([\mathbf{z}]_{\mathbf{y}=m}|\alpha\right)\right) \tag{4}$$

$$= \prod_{m=1}^{M} \left(\frac{\Gamma\left(\sum_k \alpha_k\right)}{\Gamma\left(\sum_k \left(c(m,k) + \alpha_k\right)\right)} \times \prod_k \frac{\Gamma\left(c(m,k) + \alpha_k\right)}{\Gamma\left(\alpha_k\right)}\right), \tag{5}$$

where $c(m, k)$ is the number of z_i with value k whose corresponding $y_i = m$.

For our unified frames model, the complete conditionals follow the basic form and derivation given by (Griffiths and Steyvers 2004). Note that multiple obsevations are attributable to a single latent choice, e.g., for every entity e, all $\#(f \in e)$ instances of frame $f \in e$ are attributable to the template choice $t_{d,e}$. Due to this model topology, we appeal to the general form of the Gamma factorial expansion: for real x and integral n, $\Gamma(x+n) = \left(\prod_{i=0}^{n-1}(x+i)\right)\Gamma(x)$. The conditional is then $p_{\vartheta,\beta,\xi}\left(t_{d,e} = \hat{t} | \mathbf{t}^{\setminus(d,e)}, \mathbf{s}, \mathbf{f}\right) =$

$$\frac{\mathrm{DMC}(\mathbf{t}|\vartheta)}{\mathrm{DMC}\left(\mathbf{t}^{\setminus t_{d,e}}|\vartheta\right)} \times \frac{\mathrm{DMC}(\mathbf{s}|\xi)}{\mathrm{DMC}\left(\mathbf{s}^{\setminus t_{d,e}}|\xi\right)} \times \frac{\mathrm{DMC}(\mathbf{f}|\beta)}{\mathrm{DMC}\left(\mathbf{f}^{\setminus t_{d,e}}|\beta\right)}. \tag{6}$$

We can substitute the value of each Dirichlet-multinomial compound, and applying the Gamma function expansion, arrive at a proportional value

$$\overbrace{\left(c^{\setminus t_{d,e}}(d,\hat{t}) + \vartheta_{\hat{t}}\right)}^{\text{smoothed template usage}} \times \frac{\overbrace{c^{\setminus t_{d,e}}(\hat{t}, s_{d,e}) + \xi_{s_{d,e}}}^{\text{smoothed template-specific slot frequency}}}{\sum_s c^{\setminus t_{d,e}}(\hat{t}, s) + \xi_s} \times \tag{7}$$

$$\frac{\overbrace{\prod_{f \in e}\left[\prod_{l=0}^{\#(f\in e)-1} c^{\setminus t_{d,e}}(\hat{t}, f) + \beta_f + l\right]}^{\text{smoothed per-template frame frequencies}}}{\sum_f c^{\setminus t_{d,e}}(\hat{t}, f) + \beta_f}$$

Here we use the $\setminus t_{d,e}$ notation to indicate the assignment to $t_{d,e}$ removed from the given quantity. The slot sampling equation is analogous, as are the ones for the frame and role.

Implementation Considerations In practice, the iterative multiplication in (7) will run into numerical issues if computed directly. Performing operations step-by-step in log-space is one straightforward solution, though at the cost of implementation efficiency. We found that we were able to get a 40% speed-up within our sampling inner-loop by directly computing the log variant of (6). This involves computing $\log\Gamma(x)$, for which there are numerous publicly available implementations. In our publicly available C++ implementation,[1] we use GSL.

Experiments

Minsky's, Schank's, and Fillmore's motivations were focused on matters of classic AI and cognitive science: the goal was to model human intuitions about everyday affairs (Minsky 1974; Schank 1975; Fillmore 1975). In our experiments, we address the question of how recent statistical approaches bear on the early proposals to discourse understanding, and consciously divorce our model from specific *downstream* tasks. Our evaluations reflect these desiderata. While various applications make use of the notion of an event template, such as MUC (Sundheim 1992; 1996) and ACE (Walker et al. 2006), these tasks are defined by rather limited domains. It is not clear how well these tasks get at the more generalizable background knowledge of importance to the AI pioneers. Moreover, our goal in this paper is to demonstrate how we can bring current efforts in discourse and event modeling closer to Minsky's proposal.[2]

[1] https://github.com/fmof/unified-probabilistic-frames

Data

In the spirit of past efforts to learn general domain narrative schemas, we use 10K training and 1K held-out NYT articles sampled uniformly at random from all years of Concretely Annotated Gigaword (Ferraro et al. 2014). As general newswire, the NYT tends to contain more entities, be longer, less to-the-point, and/or more diverse than previous datasets used for unlabeled template induction.

Processing In our experiments dependency parses and coreference chains are derived via CoreNLP (Manning et al. 2014), and semantic frame analyses from SEMAFOR (Das et al. 2010).

To allay concerns about errant FrameNet annotations, we apply a high-precision filtering step: we only include a mention if (1) its verb v triggers a frame f, and (2) r, one of f's frame roles, points to some span within the mention. We observed in development that these filters compensated for some of the gap in FrameNet and syntactic parsing, albeit by tying frames closely to syntax.

Evaluation Criteria

We examine the effect of frame semantics on learned templates. Quantitatively, we ask if frame semantics result in better (1) model fit (heldout log-likelihood) and (2) coherence (Mimno et al. 2011).

Evaluating held-out log-likelihood makes particular sense in the context of *surprisal* (Attneave 1959; Hale 2001; Levy and Jaeger 2006; Levy 2008). Used successfully to explain people's syntactic processing difficulties, the suprisal of a word w, given prior seen words $\mathbf{h}$ and "extra-sentential context" C (Levy 2011) is as

$$\mathrm{suprisal}\,(w|\mathbf{h}) \propto -\log p(w|\mathbf{h}, C). \tag{8}$$

Surprisal of an entire document d then follows the model's topology and factorization over d. Because our models do not examine sequences of predicate/dependency pairs, the prior history $\mathbf{h}$ is removed from the computation. For this work, we effectively examine semantic and discourse approaches to expanding out this extra-sentential context C, from within a bag-of-words view.

Chang et al. (2009) showed that improvements in topic model held-out log-likelihood do not always correlate with human quality scores. In response, Mimno et al. (2011) developed an automatic *coherence* measure that does correlate

[2] Those tasks' restricted domains mean the *evaluated* templates or relations are constrained not only by the domain, but also by the needs of the "target consumer," and what he or she deems to be "relevant." Nearly 80% of MUC-4 only has one labeled template, despite an average of (at least) three templatable events in the text (Reichart and Barzilay 2012). Further, subleties of evaluation can drastically affect the overall score and end ranking, introducing confounding variables into meta-analyses (Chambers 2013, § 5).

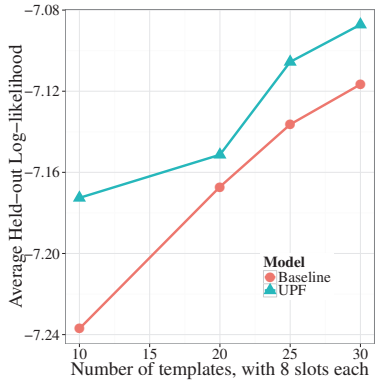


Figure 4: The held-out averaged log-likelihood of our model versus the baseline.

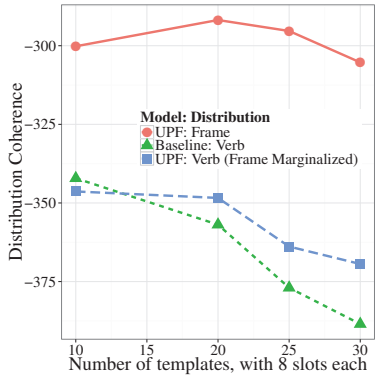


Figure 5: Topic coherence at $M = 20$. For our model (UPF), given a template, we consider frame and verb coherence – the latter marginalizes over frames. Higher is better (more coherent).

(positively) with human quality scores. Despite being developed for topic models, there is nothing inherent in its definition that limits its application to just topic models. Given a list of vocabulary words X, sorted by weight (probability), the coherence score measures the log-relative document frequencies of the M-highest probability elements of X:

$$\text{coherence}(X, M) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(X_{(m)}, X_{(l)}) + 1}{D(X_{(l)})},$$

where $D(\cdot)$ is the number of documents that have at least one occurrence of each of its arguments. We adopt this measure, as the models examined here produce distributions over predicates, frames, and other observations.

Baseline

Our baseline model is a simplification of our proposed model: it does not consider either frame or role information. This way, we can examine the effect of incorporating semantic frames in our unified model. Verbs are drawn directly from the template selection, and the arcs directly from the slots.[3] We note that this is also one of Chambers (2013)'s models; it can be viewed as very similar to Bamman, O'Connor, and Smith (2013)'s generative model. Our evaluation methodology—observing semantic frames only during training—provides a fair comparison between this baseline model and our own.

Evaluation Results

During heldout evalution *only*, we treat frames/roles as latent. For consistency, and because the true number of slots-per-template is unknown, we find a compromise among the number of slots that others have used previously; we use eight slots as we vary the number of templates (Chambers 2013; Reichart and Barzilay 2012; Balasubramanian et al. 2013). We compare per-observation log-likelihood (Fig. 4) on our 1K held-out documents to our baseline; the additional frame information allows the model to better fit held-out data, indicating a lower surprisal.

Our second evaluation is Mimno et al.'s topic coherence, evaluated on both semantic frame and verb distributions. To compute verb coherence in our model (UPF), we marginalize over frame (and role) assignments. In Figure 5 we show coherence at top 20. We find that our model ("UPF: Verb") generally had higher coherence than the baseline ("Baseline: Verb"), even as we varied the number of templates and slots.

Conclusion

We have presented a model for probabilistic frame induction which for the first time explicitly captures all levels laid out by Minsky (1974). In so doing we have combined the notion of Fillmore's frame semantics with a discourse-level notion of a Minsky frame, or Schankian script. We have shown that this leads to improved coherence, and a better explanation of held-out data.

Acknowledgments This work was supported by a National Science Foundation Graduate Research Fellowship (Grant No. DGE-1232825) to F.F., and the Johns Hopkins HLTCOE. We would like to thank members of B.V.D.'s lab, especially Chandler May, Keith Levin, and Travis Wolfe, along with Ryan Cotterell, Matthew Gormley, and four anonymous reviewers for their feedback. Any opinions expressed in this work are those of the authors.

References

Attneave, F. 1959. *Applications of Information Theory to Psychology: A summary of basic concepts, methods, and results*. Holt.

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The Berkeley Framenet Project. In *ACL*.

Balasubramanian, N.; Soderland, S.; Mausam; and Etzioni, O. 2013. Generating coherent event schemas at scale. In *EMNLP*.

Bamman, D., and Smith, N. 2014. Unsupervised discovery of biographical structure from text. *TACL* 2(10):363–376.

[3] We directly draw $v_{d,e,m} \sim \text{Cat}\left(\nu_{t_{d,e}}\right)$ and $a_{d,e,m} \sim \text{Cat}\left(\delta_{s_{d,e}}\right)$, resizing and reindexing the number of predicate and dependency distributions ν_t and δ_s as needed. We remove the discrete variables $f_{d,e,m}$ and $r_{d,e,m}$; the priors ϕ and ρ; and the hyperparameters β and ψ.

Bamman, D.; O'Connor, B.; and Smith, N. A. 2013. Learning latent personas of film characters. In *ACL*.

Bamman, D.; Underwood, T.; and Smith, N. A. 2014. A bayesian mixed effects model of literary character. In *ACL*.

Bejan, C. A. 2008. Unsupervised discovery of event scenarios from texts. In *FLAIRS*.

Chambers, N. 2013. Event schema induction with a probabilistic entity-driven model. In *EMNLP*.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-graber, J. L.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*.

Chen, H.; Benson, E.; Naseem, T.; and Barzilay, R. 2011. In-domain relation discovery with meta-constraints via posterior regularization.

Cheung, J. C. K.; Poon, H.; and Vanderwende, L. 2013. Probabilistic frame induction. In *NAACL*.

Das, D.; Schneider, N.; Chen, D.; and Smith, N. A. 2010. Probabilistic frame-semantic parsing. In *NAACL*.

Ferraro, F.; Thomas, M.; Gormley, M. R.; Wolfe, T.; Harman, C.; and Van Durme, B. 2014. Concretely Annotated Corpora. In *AKBC*.

Fillmore, C. J. 1967. The case for case. In *Proceedings of the Texas Symposium on Language Universals*. ERIC.

Fillmore, C. J. 1975. An alternative to checklist theories of meaning. In *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*.

Fillmore, C. J. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences* 280(1):20–32.

Fillmore, C. 1982. Frame semantics. *Linguistics in the morning calm* 111–137.

Frermann, L.; Titov, I.; and Pinkal, M. 2014. A hierarchical bayesian model for unsupervised induction of script knowledge. In *EACL*.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *PNAS* 101(Suppl. 1):5228–5235.

Hale, J. 2001. A probabilistic earley parser as a psycholinguistic model. In *NAACL*.

Huang, R., and Riloff, E. 2013. Multi-faceted event recognition with bootstrapped dictionaries. In *NAACL*.

Levy, R., and Jaeger, T. F. 2006. Speakers optimize information density through syntactic reduction. In *NIPS*.

Levy, R. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177.

Levy, R. 2011. Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *ACL*.

Li, Q.; Ji, H.; and Huang, L. 2013. Joint event extraction via structured prediction with global features. In *ACL*.

Liao, S., and Grishman, R. 2010. Using document level cross-event inference to improve event extraction. In *ACL*.

Lorenzo, A., and Cerisara, C. 2012. Unsupervised frame based semantic role induction: application to french and english. In *ACL Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL Demos*, 55–60.

Maslennikov, M., and Chua, T.-S. 2007. A multi-resolution framework for information extraction from free text. In *ACL*, volume 45.

Materna, J. 2013. Parameter estimation for lda-frames. In *HLT-NAACL*, 482–486.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *EMNLP*.

Minkov, E., and Zettlemoyer, L. 2012. Discriminative learning for joint template filling. In *ACL*.

Minsky, M. 1974. A framework for representing knowledge. MIT-AI Laboratory Memo 306.

Modi, A., and Titov, I. 2014. Inducing neural models of script knowledge. In *CoNLL*.

Modi, A.; Titov, I.; and Klementiev, A. 2012. Unsupervised induction of frame-semantic representations. In *NAACL Workshop on the Induction of Linguistic Structure*.

Orr, J. W.; Tadepalli, P.; Doppa, J. R.; Fern, X.; and Dietterich, T. G. 2014. Learning scripts as hidden markov models. In *AAAI*.

Patwardhan, S., and Riloff, E. 2009. A unified model of phrasal and sentential evidence for information extraction. In *EMNLP*.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A. K.; and Webber, B. L. 2008. The Penn Discourse TreeBank 2.0. In *LREC*.

Regneri, M.; Koller, A.; and Pinkal, M. 2010. Learning script knowledge with web experiments. In *ACL*.

Reichart, R., and Barzilay, R. 2012. Multi event extraction guided by global constraints. In *NAACL*.

Rosenfeld, R. 1994. *Adaptive statistical language modeling: A maximum entropy approach*. Ph.D. Dissertation, Computer Science Department, Carnegie Mellon University.

Rosenfeld, R. 1996. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech & Language* 10(3):187–228.

Rudinger, R., and Van Durme, B. 2014. Is the Stanford Dependency Representation Semantic? In *ACL Workshop on EVENTS*.

Ruppenhofer, J.; Ellsworth, M.; Petruck, M. R.; Johnson, C. R.; and Scheffczyk, J. 2006. *FrameNet II: Extended Theory and Practice*. Berkeley, California: ICSI.

Schank, R. C., and Abelson, R. 1977. Scripts. *Plans, Goals, and Understanding. Lawrence, Erlbaum, Hillsdale*.

Schank, R. C. 1975. Using knowledge to understand. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing (TINLAP)*.

Sundheim, B. 1992. Proceedings of the fourth message understanding conference (MUC-4).

Sundheim, B. M. 1996. Overview of results of the MUC-6 evaluation.

Titov, I., and Klementiev, A. 2011. A bayesian model for unsupervised semantic parsing. In *ACL*.

Van Durme, B., and Lall, A. 2009. Streaming pointwise mutual information. In *NIPS*.

Walker, C.; Strassel, S.; Medero, J.; and Maeda, K. 2006. ACE 2005 Multilingual Training Corpus LDC2006T06. DVD. Philadelphia: Linguistic Data Consortium.

Wallach, H. M. 2008. *Structured topic models for language*. Ph.D. Dissertation, University of Cambridge.