Batch Size r	2	5	10
SEM-HMM	42.2%	45.1%	46.0%
SEM-HMM Approx.	43.3%	43.5%	44.3%
BMM + EM	41.1%	41.2%	42.1%
BMM	41.0%	39.5%	39.1%
Conditional Frequency			36.2% 27.3%

Table 1: The average accuracy on the OMICS domains

Example 1	Example 2
Hear the doorbell.	Listen for the doorbell.
Walk to the door.	Go towards the door.
Open the door.	Open the door.
Allow the people in.	Greet the vistor.
Close the door.	See what the visitor wants.
	Say goodbye to the visitor.
	Close the door.

Table 2: Examples from the OMICS "Answer the Doorbell" task with event triggers underlined

use EM for parameter estimation and instead only incrementally updates the parameters starting from the raw document counts. Further, it learns a standard HMM, that is, with no λ transitions. This is very similar to the Bayesian Model Merging approach for HMMs (**?**). The fourth baseline is the same as above, but uses our EM algorithm for parameter estimation without λ transitions. It is referred to as "BMM + EM."

The Open Minds Indoor Common Sense (OMICS) corpus was developed by the Honda Research Institute and is based upon the Open Mind Common Sense project (**?**). It describes 175 common household tasks with each task having 14 to 122 narratives describing, in short sentences, the necessary steps to complete it. Each narrative consists of temporally ordered, simple sentences from a single author that describe a plan to accomplish a task. Examples from the "Answer the Doorbell" task can be found in Table 2. The OMICS corpus has 9044 individual narratives and its short and relatively consistent language lends itself to relatively easy event extraction.

The 84 domains with at least 50 narratives and 3 event types were used for evaluation. For each domain, forty percent of the narratives were withheld for testing, each with one randomly-chosen event omitted. The model was evaluated on the proportion of correctly predicted events given the remaining sequence. On average each domain has 21.7 event types with a standard deviation of 4.6. Further, the average narrative length across domains is 3.8 with standard deviation of 1.7. This implies that only a frcation of the event types are present in any given narrative. There is a high degree of omission of events and many different ways of accomplishing each task. Hence, the prediction task is reasonably difficult, as evidenced by the simple baselines. Neither the frequency of events nor simple temporal structure is enough to accurately fill in the gaps which indicates that most sophisticated modeling such as SEM-HMM is needed.

The average accuracy across the 84 domains for each method is found in Table 1. On average our method significantly out-performed all the baselines, with the average improvement in accuracy across OMICS tasks between SEM-HMM and each baseline being statistically significant at a .01 level across all pairs and on sizes of $r = 5$ and $r = 10$ using one-sided paired t-tests. For $r = 2$ improvement was not statistically greater than zero. We see that the results improve with batch size r until $r = 10$ for SEM-HMM and BMM+EM, but they decrease with batch size for BMM without EM. Both of the methods which use EM depend on statistics to be robust and hence need a larger r value to be accurate. However for BMM, a smaller r size means it reconciles a couple of documents with the current model in each iteration which ultimately helps guide the structure search. The accuracy for "SEM-HMM Approx." is close to the exact version at each batch level, while only taking half the time on average.

5 Conclusions

In this paper, we have given the first formal treatment of scripts as HMMs with missing observations. We adapted the HMM inference and parameter estimation procedures to scripts and developed a new structure learning algorithm, SEM-HMM, based on the EM procedure. It improves upon BMM by allowing for λ transitions and by incorporating maximum likelihood parameter estimation via EM. We showed that our algorithm is effective in learning scripts from documents and performs better than other baselines on sequence prediction tasks. Thanks to the assumption of missing observations, the graphical structure of the scripts is usually sparse and intuitive. Future work includes learning from more natural text such as newspaper articles, enriching the representations to include objects and relations, and integrating HMM inference into text understanding.

Acknowledgments

We would like to thank Nate Chambers, Frank Ferraro, and Ben Van Durme for their helpful comments, criticism, and feedback. Also we would like to thank the SCALE 2013 workshop. This work was supported by the DARPA and AFRL under contract No. FA8750-13-2-0033. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA, the AFRL, or the US government. Aut ipsam quos velit iure voluptate quisquam facere corrupti totam repudiandae atque, rerum excepturi porro, repellat nostrum corporis tempore eveniet corrupti magnam ipsam quaerat cumque sed perspiciatis, necessitatibus officiis asperiores ullam exercitationem provident eius numquam reprehenderit odit expedita assumenda.Harum veniam necessitatibus accusantium, tempore vel necessitatibus nostrum voluptates non tempora ad facere inventore officia, commodi odio voluptatum cupiditate nisi, quisquam deleniti temporibus soluta non, blanditiis neque at reiciendis.Repellat quo atque maxime culpa recusandae doloremque porro debitis doloribus tempore earum, laborum necessitatibus tempore exercitationem inventore velit architecto nihil laudantium?Incidunt alias provident sed nesciunt aut reiciendis exercitationem dolore vel volup-

tatum quae,