Universität des Saarlandes
WiSe 2016/17
Computational Linguistics

We tried to investigate the potentials of Latent Dirichelet Allocation with Gibbs sampling in the framework of scripts.
For this purpose, we tested the model on the InScript corpus (Mody et al.2016), a crowdsourced collection of simple stories instantiating 10 scenarios (e.g. 'GETTING A HAIRCUT', 'BAKING A CAKE'). The corpus is in the current folder.
Our work is divided in 3 steps:

1. all_scenarios: we tested the model on the whole corpus to see how well it could retrieve 10 topics corresponding to the 10 scenarios.

2. cake_documents: we were particularly interested in investigating whether LDA can be used to retrieve a script's internal structure.
Modi et al. also created script templates that described script-specific event labels and participant labels for each scenario (e.g., event labels in BAKING A CAKE: get_ingredients, put_cake_oven, etc. and participant labels: ingredients, oven, etc.), which were used to annotate the stories. They annotated event-denoting verbs in the stories with the event labels and participant-denoting NPs with the participant labels.
The idea was that, assuming a number of topics equal to the number of labels identified in the corpus for a specific scenario, LDA could retrieve a similar structure to the one that was manually annotated in the corpus, namely identify words with the same label as belonging to the same topic.
For this purpose, we tested the model on the 97 stories from the CAKE scenario (37 labels), which seemed to us to display a more uniform, script-like structure across stories. The labels for the CAKE scenario can be manually inspected in: InScript/templates/Cake_scenario_readme.pdf

3. cake_sentences: since the model might perform better when the documents display a strong preference for a topic (it was at least partially successful in step 1 but not in step 2), we tested it on the corpus of all sentences from the 97 stories in the CAKE scenario, assuming that speakers tend to organize the text with a sequence similar to the one of the steps of a script.
The idea was that, assuming a number of topics equal to the number of steps in a scenario, LDA could identify words describing the same step as belonging to the same 'topic'.
In this case, we didn't assume the number of steps to be equal to the number of labels.
Wanzare et al. (2016) crowdsourced the DeScript Corpus, a corpus of event-sequence descriptions (EDSs) for a number of well-known activities (including BAKING A CAKE). 320 English native speakers were asked to write down one EDS for each scenario, namely a minimum of 5 and a maximum of 16 steps which would instantiate each script. Semi-supervised clustering algorithms were employed to group event descriptions into paraphrase sets. Each paraphrase set was manually labeled (e.g. choose_recipe, buy_ingredients, etc. for the scenario BAKING A CAKE). We considered the number of paraphrase sets to be more representative of the number of steps to retrieve in the cake scenario (27).

Francesca Capuano