representation of words and documents.

We train Context-GPU and set θ and σ to 0.7 and 0.8 respectively based on a grid search of values in $[0.5, 0.6, 0.7, 0.8, 0.9]$ using 5-fold cross validation. We set the maximum number of Gibbs sampling iterations to 1500. We compare Context-GPU with the following baselines:

- LDA. We use the LDA implementation in MALLET[6] with the default settings and perform hyperparameter optimization every 200 iterations.
- Generalized Pólya urn (GPU) model (**?**). We implemented this algorithm by modifying the LDA implementation in the MALLET library.
- TopicVec (**?**). We use the available implementation[7] with the default configuration, increasing the maximum iteration number.
- TPM (**?**). We implemented the Topical Phrase Model which extracts medical topics using both MedTagger and a hierarchy of Pitman-Yor processes. It outperformed other topical phrase extraction models.

Topic coherence

We assess the generated topics by evaluation of their topic coherence. We adopt the topic coherence measure proposed in Mimno et al. (**?**), which relies on the co-occurrence statistics collected from the analyzed corpus; this allows us to directly measure the coherence of topics with topical phrases (e.g. *short_of_breath*).

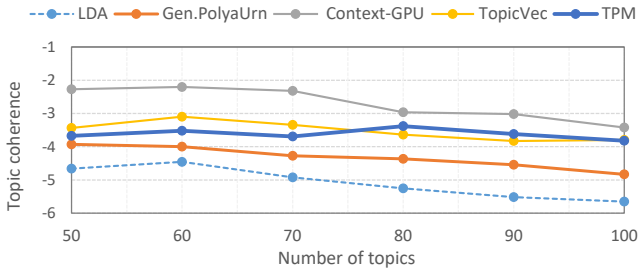

Figure 1: Topic coherences vs. number of topics.

In our evaluations, we compute the topic coherence on the top 10 words/phrases using the implementation provided in the Palmetto library[8] (**?**). In Figure **??** we report the topic coherence computed by averaging the coherences resulting for each topic. A peak of coherence is obtained around 60/70 topics for every model, suggesting a potentially suitable number of topics to discriminate the documents. GPU with only local context incorporated outperforms LDA, but its performance is worse compared to TopicVec or TPM. Context-GPU gives superior results over all the baseline models, in particular around 60 and 70 topics. This shows that additionally incorporating global context is essential to achieve better topic coherence results compared to only considering local context. Also, our proposed Context-GPU only involves simple modifications to GPU, but it appears to be more effective than more complicated ways of incorporating word embeddings into topics models (such as TopicVec) or assuming word generation following the HPYP process (such as TPM).

To extract topical phrases from text, we have explored a few different ways in learning word/phrase representations such as learning directly from our data using SVD, training a combined Wikipedia/clinical report data using SGNS or FastText. In Figure **??** we compare these word/phrase embedding learning results over our Context-GPU. We can observe that SVD and SGNS perform similarly in most cases and SVD even slightly outperforms SGNS when the topic number is set to 80 or 90. FastText outperforms the other two word/phrase embedding learning methods especially when the topic number is lower than 80. This shows that FastText built on character n-grams is more effective in capturing phase sub-structures.

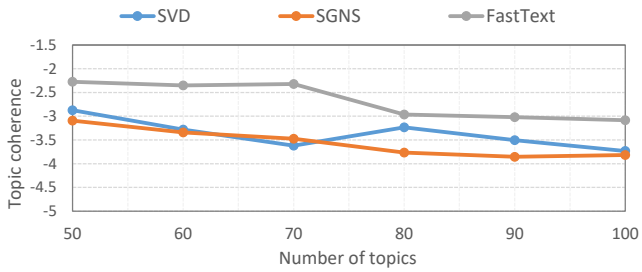

Figure 2: Context-GPU with different word/phrase embedding learning methods vs. number of topics.

Finally, we compare in Figure **??** the execution time needed to train the models, excluding the constant time required by each model to load the embeddings. We did not plot the training time for TPM in the figure as it required significantly more time (over 12 hours) compared to all the other models, showing that modeling phrase generation using HPYP is very expensive. For the remaining models, TopicVec is computationally more complex than the others. Both GPU and Context-GPU have no noticeable difference and they both required three-fold the training time of LDA. Overall, Context-GPU appears to be more effective compared to TopicVec and TPM.

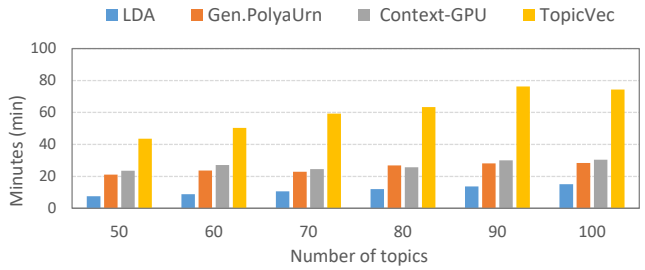

Figure 3: Execution time vs. number of topics.

[6] http://mallet.cs.umass.edu

[7] https://github.com/askerlee/topicvec

[8] https://github.com/dice-group/Palmetto

Topic 1	Topic 2	Topic 3	Topic 4
TopicVec			
carotid	diuresis	dyspnea on exertion	congestive heart failure
coronary artery	torsemide	ejection fraction	fibrillation
magnesium	cardiomyopathy	pulmonary	ejection fraction
saphenous vein graft	shortness of breath	atrial fibrillation	insufficiency
potassium chloride	torsemide 100 mg	diuresed	calcium
coronary artery bypass grafting	spironolactone 25 mg	congestive heart failure	intubation
mitral insufficiency	diuretic	ischemia	thyroid
mitral regurgitation	aldactone	diabetes mellitus	vascular congestion
potassium	pleural effusion	propafenone	tricuspid regurgitation
substernal	pulmonary edema	volume overloaded	right knee
Contex-GPU			
pregnancy	mitral regurgitation	coronary artery disease	congestive heart failure
ultrasound	digoxin	cardiac transplant	pulmonary edema
postpartum hemorrhage	pleural effusion	cardiomyopathy	orthopnea
endometrial biopsy	orthopnea	right coronary artery	nonischemic
total abdominal hysterectomy	dilated cardiomyopathy	pravachol 20 mg	diastolic dysfunction
postpartum	plavix 75 mg	paroxysmal atrial fibrillation	cardiomyopathy
vomiting	shortness of breath	cyclosporine	heart failure
salpingo oophorectomy	dyspnea on exertion	herpes zoster	shortness of breath
physical examination	tachyarrhythmia	fenofibrate tricor	cardiac catheterization
fibroid	pulmonary edema	right coronary artery	atrial fibrillation

Table 1: Topics generated by TopicVec and Context-GPU in 70-topic runs.

Topic Qualitative Assessment

We report in Table **??** some topics generated in a 70-topics run. For the sake of brevity, we report only the topics of TopicVec and Context-GPU since TopicVec gives similar coherence scores as TPM but requires significantly less training time compared to TPM. TopicVec inference learns both word and topic embeddings simultaneously. It allows the model to take into account the local context of words, which in turn, alleviates the lack of global statistic for a term. Both the topics of TopicVec and Context-GPU are able to generate topical phrases. However, in several topics of Context-GPU, we can distinguish a gradual definition of the analyzed themes, which reflect better semantic coherence. For example, in Topic 4, it can be observed a gradual topic refinement under Context-GPU from the general purpose terms such as *felt* or *insufficiency* to more characterizing words/phrases such as *shortness of breath*, *atrial fibrillation*. In addition, we can observe under the same topic symptom and medication, such as *cardiomyopathy* and *plalix 75 mg*. As a result, the overall expressiveness of topics extracted by the Context-GPU outperforms TopicVec, both thanks to their internal coherence and to the improved expressiveness of the adopted words/phrases.

Conclusion

We have described a new approach which aims to effectively combine the local and global context of words and phrases. It first detects high reliable phrases and then generates topics using our proposed Context-aware Pólya urn model. This statistical model combines the word semantic encoded by the context-based and corpus-based embeddings. In particular, we have exploited the LSA and FastText embeddings. The former improved the ties of a word to the corpus themes; the latter allowed a fine-grained use of a word depending on the phrase in which it occurs. An experimental comparison with the state-of-the-art methods has shown an improved coherence of final topics and a decreased computational cost.