Processus de Dirichlet profonds pour le topic modeling

Miguel Palencia-Olivar[*,**], Stéphane Bonnevay[*,**]
Alexandre Aussem[***], Bruno Canitia[*]

[*]Lizeo IT, 42 quai Rambaud, 69002 Lyon, France
miguel.palencia-olivar, stephane.bonnevay, bruno.canitia@lizeo-group.com
[**]Laboratoire ERIC, Université de Lyon, 5 Av. Pierre Mendès France, 69500 Bron, France
[***]LIRIS, Université de Lyon, 25 Av. Pierre de Coubertin, 69100 Villeurbanne, France
alexandre.aussem@univ-lyon1.fr

Résumé. Cet article présente deux nouveaux modèles : l'Embedded Dirichlet Process et l'Embedded Hierarchical Dirichlet Process. Ces méthodes sont des extensions non-paramétriques de l'Embedded Topic Model (ETM) qui permettent d'apprendre simultanément le nombre de thématiques, des représentations latentes de documents, des embeddings de thématiques et des embeddings de mots. Pour ce faire, nous remplaçons l'*a priori* logit-normal de l'ETM par des processus de Dirichlet dans un cadre d'inférence par autoencodage variationnel amorti. Nous testons nos modèles sur deux jeux de données : 20 Newsgroups et Humanitarian Assistance and Disaster Relief. Nos modèles présentent l'avantage de maintenir une faible perplexité tout en fournissant des représentations sémantiques parlantes qui surclassent celles des autres méthodes de l'état de l'art. Enfin, les topics sont extraits dans un contexte multilingue, et ce sans sacrifice d'un *a priori* de type Dirichlet.

1 Introduction

Largement utilisés dans l'industrie et dans le monde universitaire (Boyd-Graber et al. (2017)), les *topic models* font partie des outils de référence pour l'exploration non supervisée de corpus. Depuis son introduction, *l'Allocation de Dirichlet Latente (LDA)* (Blei et al. (2003)) a été utilisée comme canevas de base pour de nombreux topic models dont les destinations varient. Ces destinations différentes se reflètent notamment sur les hypothèses sous-jacentes à ces modèles, donnant ainsi autant de cas d'usages divers. La LDA est un modèle à deux niveaux qui suppose que dans un corpus, les documents - indépendants deux-à-deux - sont des mélanges de topics qui sont eux-mêmes des mélanges de multinomiales appliqués au niveau des mots. Ces mots sont également considérés indépendants deux-à-deux. Les processus d'inférence (Blei et al. (2017); Griffiths et Steyvers (2004)) visent à déterminer les proportions de mélange pour chaque niveau. Ces processus d'inférence sont à sélectionner en fonction des objectifs poursuivis, et notamment la masse de donnés à traiter (Blei et al. (2017)). Dans cet article, nous nous concentrons sur l'inférence variationnelle neuronale (Kingma et Welling (2014)), car elle permet de bénéficier à la fois des propriétés et des avancées du *Deep Learning* et des topic models probabilistes, tout en conservant une bonne interprétabilité des résultats.

Or, et malgré ces caractéristiques, le nombre de topics reste en général considéré comme un hyper-paramètre. Par conséquent, l'utilisateur est obligé d'effectuer plusieurs essais pour réussir à sélectionner un modèle, car il est impossible de deviner en amont le nombre de sujets dans une quantité massive de texte. Les ré-exécutions sont relativement faciles à réaliser pour l'analyse de petits corpus, mais elles posent des problèmes de coûts importants lorsque les expériences sont réalisées à grande échelle. Pour remédier à ce problème, nous avons proposé dans Palencia-Olivar et al. (2021) deux nouveaux modèles : l'Embedded Dirichlet Process (EDP) et l'Embedded Hierarchical Dirichlet Process (EHDP). Ces deux topic models basés sur le framework VAE infèrent automatiquement le nombre de topics à partir des données grâce à des distributions Beta et Gamma. Nos modèles présentent l'avantage de maintenir une faible perplexité tout en fournissant des représentations parlantes de document, de topics et de mots qui surpassent d'autres méthodes de l'état de l'art, tout en évitant de coûteuses ré-exécutions. Nous nous concentrons principalement sur le maintien d'un compromis entre le pouvoir prédictif et l'interprétabilité.
Cet article est organisé comme suit : la section 2 présente succintement notre état de l'art, en mettant l'accent sur l'Embedded Topic Model (ETM) de Dieng et al. (2019) et les processus de Dirichlet. Nous présentons nos modèles en section 3, ainsi que leurs objectifs d'optimisation. La section 4 est consacrée aux études empiriques. Enfin, la section 5 présente nos conclusions et les orientations futures de nos recherches.

2 État de l'art

2.1 L'Embedded Topic Model

L'Embedded Topic Model (ETM ; Dieng et al. (2019)) étend la LDA pour inclure des embeddings de mots et des embeddings de topics dans son processus génératif. L'algorithme apprend ces embeddings à partir de données. Les embeddings sont visualisables dans le même espace par construction du modèle [1]. Il est également possible - mais facultatif - d'utiliser des embeddings pré-entraînés, y compris pour des mots qui n'apparaissent pas dans le jeu de données considéré ; le modèle adaptera les embeddings pour les mots supplémentaires en fonction de leur voisinage. L'ETM partage les hypothèses de mélange de LDA, à ceci près que, contrairement à cette dernière, les mots et les sujets peuvent être corrélés. Le passage à l'échelle de l'ETM est assuré par l'usage d'un VAE (Kingma et Welling (2014)), au prix d'un *a priori* logit-normal au lieu d'un *a priori* de Dirichlet afin d'être compatible avec l'astuce de reparamétrisation nécessaire au fonctionnement du VAE.

2.2 Processus de Dirichlet et construction *stick-breaking*

La construction *stick-breaking* est souvent utilisée dans les modèles basés sur les *Processus de Dirichlet* (Teh et al. (2006)). Ishwaran et James (2001) définissent un *a priori stick-breaking* comme étant une mesure aléatoire de la forme $G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\zeta_k}(\cdot)$, où δ_{ζ_k} est une mesure discrète concentrée à $\zeta_k \sim G_0$, un tirage d'une distribution de base G_0 (une fonction δ de Dirac par exemple). Les termes π_k sont des poids aléatoires qui ne dépendent pas de G_0, choisis de

1. La matrice mots-topics est en effet un modèle log-linéaire, donnant ainsi lieu à une décomposition dont l'esprit est similaire à celle réalisée dans le cadre d'une ACP.

telle sorte à ce que $0 \leq \pi_k \leq 1$, et $\sum_k \pi_k = 1$ presque sûrement. La procédure suivante permet l'échantillonnage des pondérations π_k :

$$\pi_k = \begin{cases} v_k \; if \; k = 1 \\ v_k \Pi_{j<k} \left(1 - v_j\right) \; for \; k > 1 \end{cases} \tag{1}$$

avec $v_k \sim Beta\,(\alpha, \beta)$. Lorsque $\alpha = 1$, la distribution *Beta* est la construction stick-breaking pour le processus de Dirichlet. Cette distribution est également appelé distribution *Griffiths, Engen et McCloskey* (GEM). La distribution GEM prend un seul paramètre de concentration α_0 qui est égal au deuxième paramètre de forme (β) de la distribution Bêta. Nalisnick et Smyth (2017) ont adapté le stick-breaking pour les processus de Dirichlet afin qu'il fonctionne avec un *VAE*. Du fait des contraintes imposées par l'astuce de la reparamétrisation, cette adaptation utilise une distribution de Kumaraswamy au lieu d'une distribution Beta.

2.3 Reparamétrisation implicite

L'astuce de la reparamétrisation de Kingma et Welling (2014) est essentielle au calcul de gradients à faible variance pour les variables aléatoires continues dans le cadre du framework VAE. Cette technique fonctionne mieux avec les distributions de type gaussienne, à l'exclusion des distributions Gamma, Bêta, et Dirichlet. Figurnov et al. (2019) ont proposé une alternative dite de *reparamétrisation implicite*. En se basant sur les techniques de différenciation implicite et automatique, celle-ci permet non seulement d'utiliser des distributions basées sur la distribution Gamma [2], mais aussi un échantillonnage plus rapide.

3 Modèles

3.1 L'Embedded Dirichlet Process

L' *Embedded Dirichlet Process* (EDP) est un topic model optimisé par un VAE et dont l'un des principaux apports est de déterminer automatiquement le nombre de topics. Par ailleurs, il permet de prendre en compte les liens entre les mots et les liens entre les thématiques sans sacrifice d'un *a priori* de type Dirichlet, contrairement au Correlated Topic Model de Blei et Lafferty (2007). Pour ce faire, le modèle inclut un *a priori* GEM à utiliser en conjonction soit avec une distribution de Kumaraswamy (reparamétrisation explicite), soit avec une distribution Bêta (reparamétrisation implicite). L'*a priori* GEM permet non seulement l'inférence du nombre de topics, mais d'également d'obtenir un modèle plus expressif. Tout comme l'ETM, l'EDP décompose le niveau du mot en un produit matriciel entre les embeddings de topics ϕ et la transposée des embeddings de mots ρ. Il présente les mêmes capacités à déterminer des topics et des espaces d'embedding que l'ETM. Soit $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$ un corpus de D documents, où $\mathbf{w}_d$ est une collection de N_d mots. Pour compléter le processus génératif, nous devons calculer la distribution ci-après.

$$Pr\left(\mathbf{w}_{1:N}, \pi, \hat{\theta}_{1:N} \mid \alpha_0, \Theta, \beta\right) = Pr\left(\pi \mid \alpha_0\right) \Pi_{i=1}^{N} Pr\left(w_i \mid \hat{\theta}_{\mathbf{i}}, \beta\right) Pr\left(\hat{\theta}_{\mathbf{i}} \mid \pi, \Theta\right) \tag{2}$$

2. La loi Bêta peut être échantillonnée à partir d'échantillons de la loi Gamma, voir Figurnov et al. (2019) pour davantage de détails.

où $Pr(\pi \mid \alpha_0) = GEM(\alpha_0)$, $Pr(\theta \mid \pi, \Theta) = G(\theta; \pi, \Theta)$, et $Pr(w \mid \theta, \beta) = \sigma(\theta\beta)$. Par simplification, nous désignons le produit matriciel entre les matrices d'embedding par $\beta = \sigma\left(\rho^T \phi\right)$ où $\sigma(\cdot)$ est la fonction softmax. Nous utilisons une famille de distributions variationnelles dont les paramètres sont déduits à l'aide de perceptrons multicouches pour déterminer une borne inférieure à la log-vraisemblance marginale. Nous désignons cette borne inférieure sous le nom d'*evidence lower bound* (ELBO), qui est une fonction des paramètres du modèle et des paramètres variationnels que nous chercherons à maximiser :

$$\mathcal{L}(\mathbf{w}_{1:N} \mid \Theta, \psi, \beta) = \mathbb{E}_{q_\psi(\nu|\mathbf{w}_{1:N})}\left[\log \Pr(\mathbf{w}_{1:N} \mid \pi, \Theta, \beta)\right] - \mathrm{KL}\left(q_\psi(\nu \mid \mathbf{w}_{1:N}) \parallel \Pr(\nu \mid \alpha_0)\right) \quad (3)$$

où $q_\psi(\cdot)$ représente la famille de distributions variationnelles, ψ représente les paramètres du réseau neuronal et v représente les poids pour la construction stick-breaking. Nous désignons les réseaux neuronaux par NN, et leur passons des sacs de mots $\mathbf{x}$. Nous utilisons l'optimiseur Adam pour ajuster le modèle, à la fois pour les paramètres variationnels et pour les paramètres de distribution.

3.2 L'Embedded Hierarchical Dirichlet Process

Dans le cadre précédent, le paramètre de concentration α_0 est traité comme un hyperparamètre. Plus sa valeur est grande, plus le nombre de bris sera élevé et donc nous obtiendrons davantage de topics, et inversement. Pour limiter leur sur-croissance et éviter leur redondance sémantique, nous cherchons à apprendre la concentration à partir des données. Nous pouvons utiliser la distribution Gamma (δ_1, δ_2) comme *a priori* conjugué pour la distribution GEM. L'ELBO devient alors la suivante :

$$\begin{aligned} \mathcal{L}(\mathbf{w}_{1:N} \mid \Theta, \psi, \beta) &= \mathbb{E}_{q_\psi(\nu|\mathbf{w}_{1:N})}\left[\log Pr(\mathbf{w}_{1:N} \mid \pi, \Theta, \beta)\right] \\ +\mathbb{E}_{q_\psi(\nu|\mathbf{w}_{1:N})q(\alpha|\gamma_1,\gamma_2)}\left[\log Pr(\nu \mid \alpha_0)\right] &- \mathbb{E}_{q_\psi(\nu|\mathbf{w}_{1:N})}\left[\log q_\psi(\nu \mid \mathbf{w}_{1:N})\right] \\ -KL\left(q(\alpha_0 \mid \gamma_1, \gamma_2) \parallel Pr(\alpha_0 \mid \delta_1, \delta_2)\right) & \end{aligned} \quad (4)$$

4 Expériences

4.1 Jeux de données

Les expériences portent sur les jeux de données annotés 20 Newsgroups (20NG) et Humanitarian Assistance and Disaster Relief (HADR) (Horwood (2017)). Notons que HADR est livré avec un lexique que nous utiliserons pour l'estimation qualitative des résultats. Ces deux jeux sont des collections d'articles de 20 sujets (classes) dans le cas de 20 Newsgroups, et de 25 sujets pour HADR. 20 Newsgroups contiennent 18846 articles, tandis que HADR en contient environ 504000 dans différentes langues. Nous ne souhaitons pas reproduire le biais de l'annotateur, mais les classes restent bienvenues pour juger de la qualité de l'extraction. En raison d'une limitation technique pour ce travail [3], nous avons retenu un sous-ensemble aléatoire de 20000 articles de HADR pour nos expérimentations. Dans chaque cas, nous avons utilisé 85% de l'ensemble des données pour les ensembles d'entraînement, 10% pour les ensembles de validation et 5% pour les ensembles de test. Après pré-traitement du vocabulaire, nous retenons

3. La machine utilisée pour les expérimentations ne permettait pas de retenir davantage de données.

des V-vocabulaires de 28307 mots provenant de 20 Newsgroups et de 32794 mots provenant de HADR.

4.2 Paramètres d'entraînement

Nous comparons nos résultats avec ceux des modèles iTM-VAE-Prod, iTM-VAE-G de Ning et al. (2020) et enfin, l'ETM de Dieng et al. (2019). Ces modèles utilisent tous la reparamétrisation explicite. iTM-VAE-Prod est un topic model non-paramétrique avec un *a priori* GEM, mais le modèle n'inclut aucun type de similarité entre topics et mots. Pour une étude plus approfondie, nous avons également adapté iTM-VAE-Prod pour inclure une reparamétrisation implicite. Nous entraînons nos modèles avec une taille de *batch* de 1000 documents, et choisi Adam avec un pas d'apprentissage de $0,002$ dans un cadre d'inférence variationnelle amortie. Enfin, et suivant Dieng et al. (2019), nous avons normalisé les représentations des sacs de mots des documents en les divisant par le nombre de mots pour tenir compte de la longueur des documents. Les encodeurs sont des perceptrons multicouches avec deux couches cachées de 100 neurones chacune. Concernant les *a priori*, nous avons fixé $\alpha = 1$ et $\beta = 5$ pour iTM-VAE-Prod et EDP, $\delta_1 = 1$ et $\delta_2 = 20$ pour iTM-VAE-G et EHDP. Nous inférons pour 50 et 200 topics pour les modèles paramétriques. Tous les paramètres et hyper-paramètres susmentionnés ont été sélectionnés par validation croisée à l'aide des métriques décrites ci-dessous.

4.3 Métriques

En pratique, les analystes utilisent les topic models pour fournir à la fois des représentations de sujets sémantiquement parlantes et une bonne capacité d'inférence sur documents hors ensemble d'entraînement. Cependant, la plupart des topic models ne sont ajustés et sélectionnés que d'un point de vue statistique, la cohérence des thèmes étant calculée périodiquement en raison de son coût élevé. Ceci est particulièrement vrai lorsqu'il s'agit d'indicateurs de type *information mutuelle normalisée*. Dans cette configuration, la cohérence est un indicateur supplémentaire qui est pratiquement séparé du processus d'entraînement. Notre travail se concentrant sur le maintien d'un compromis entre la qualité de l'ajustement et l'interprétabilité, nous sélectionnons nos modèles sur la base d'un rapport cohérence - perplexité pendant l'étape de validation. La perplexité est une métrique commune au topic modeling et, plus globalement, aux modèles de langage [4]. Sa formule est la suivante : $\exp\left(-\frac{1}{D}\sum_{d=1}^{D}\frac{1}{|\mathbf{w}^d|}\log Pr\left(\mathbf{w}^d\right)\right)$. D représente le nombre de documents, et $\mathbf{w}^d$ est le nombre de mots dans le d-ième document. Comme l'ELBO est une borne supérieure de la perplexité, nous l'utilisons pour calculer l'indicateur. Le principal intérêt de la perplexité est d'évaluer le pouvoir prédictif du modèle ; il s'agit en fait de la log-vraisemblance du modèle. Nous calculons la perplexité dans le cadre d'une tâche de complétion de documents. Nous testons alors la capacité du modèle a être « surpris » lors d'une tâche de génération de mots. Comme dans l'article de Dieng et al. (2019), nous considérons que la qualité d'un ajustement de topic model dépend à la fois de la redondance et de la pertinence sémantique du résultat obtenu. Par conséquent, nous calculons la qualité des sujets comme le produit de la diversité des sujets et de leur cohérence. La diversité des sujets est le rapport entre le nombre de tokens uniques parmi les 10 premiers mots de la liste des

4. Les topic models probabilistes peuvent également être vus comme des cas particuliers de modèles de langage du fait qu'ils posent une distribution de probabilité sur des mots.

Modèle	**Diversité**				**Cohérence**			
	20NG		HADR		20NG		HADR	
# topics	50	200	50	200	50	200	50	200
ETM	0.47	0.32	0.45	0.28	0.15	-0.06	0.15	0.07
iTM-VAE-G	0.91		**1.0**		0.10		0.16	
EHDP	0.52		**1.0**		**0.38**		**0.32**	
iTME-VAE-Prod implicite	**1.0**		**1.0**		0.04		0.12	
EDP implicite	**1.0**		**1.0**		0.04		0.08	

Modèle	**Qualité**			
	20NG		HADR	
# topics	50	200	50	200
ETM	0.07	-0.02	0.07	0.02
iTM-VAE-G	0.09		0.16	
EHDP	**0.20**		**0.32**	
iTME-VAE-Prod implicite	0.04		0.12	
EDP implicite	0.04		0.08	

TAB. 1 – *Qualité des topics par jeu de données et par nombre de topics. Les meilleurs résultats sont indiqués en gras.*

sujets et le nombre total de mots dans ces 10 premiers. Quant à la cohérence, nous utilisons l'information mutuelle normalisée (NPMI) pour mesurer la co-occurrence des termes dans les corpus.

4.4 Résultats

Les topic models non-paramétriques qui utilisent l'astuce de reparamétrisation explicite ont tous commencé à produire des NaNs dès la deuxième époque de la boucle d'entraînement ; par conséquent, nous les excluons de notre analyse. Cela s'explique par l'avènement d'un *posterior collapse*. Ce phénomène confirme cependant la pertinence de la reparamétrisation implicite car elle permet de construire des modèles probabilistes plus robustes. Nous avons constaté que les modèles sont proches en termes de pouvoir prédictif, aussi, et selon nos critères de sélection, la qualité des topics est prépondérante pour les départager (table 1). Notre EHDP surpasse les autres techniques en termes de qualité, y compris l'EDP.

La table 2 montre les topics extraits de HADR avec EHDP [5]. Bien que notre modèle n'ait pas détecté autant de sujets que l'annotation humaine, il en a fusionné certains. Ainsi, dans HADR, les glissements de terrain, les pluies et les inondations sont trois sujets distincts qui en

5. Dans le topic 1, rly fait référence au chemin de fer. Dans le topic 2, les acronymes correspondent aux Nations-Unies, et dans le topic 3, à des organismes de prédiction météorologique. Enfin, dans le topic 5, drc signifie République démocratique du Congo.

Topic	Liste de mots
India & Bangladesh	tongi, manu, gorai, rly, storey, serjganj, kanaighat
United Nations	gva, dhagva, metzner, masayo, pbp, spaak, pos
Weather	tpc, nws, knhc, outward, forecaster, accumulations, ast
Floods & landslides	floods, landslides, padang, flooding, rain, mudslides, sichuan
Africa	drc, lusaka, monuc, burundi, darfur, congolese, amis
Economic development	development, financing, macroeconomic, management, reduction, usaid, sustainable
Politics & diplomacy	paragraph, decides, resolution, pursuant, vii, welcomes, stresses
French stopwords	les, qui, de, que, à, une, des

TAB. 2 – *Liste complète des topics extraits par EHDP à partir d'HADR.*

constituent un seul pour l'EHDP. Cette caractéristique est utile pour obtenir des résultats plus résumés, mais elle peut entraver la classification des documents en raison de l'entrelacement des variables. Le topic sur les stopwords français, est un résultat de la sérendipité. Dieng et al. (2019) montrent que l'ETM peut gérer les stopwords et les séparer dans un sujet distinct, mais ETM n'est testé que dans des contextes monolingues. HADR, cependant, est un jeu de données multilingue. Il semblerait que certains documents en français sont restés après que nous ayons filtré le jeu de données. EDHP a quand même réussi à distinguer les stopwords parmi le vocabulaire, tout en les classant par langue, alors que le multilinguisme ne fait pas partie des hypothèses de modélisation. Nous expliquons ce résultat par le fait que nos embeddings sont contextuels, et que les mots français sont beaucoup plus susceptibles d'apparaître dans des documents en français. Malgré son origine accidentelle, ce résultat est intéressant, car le topic modeling et la fouille de textes multilingues restent un problème ouvert (Vulić et others (2015)). En particulier, nous pensons que ces mots peuvent servir de pivots pour relier des mots d'autres langues (rares, notamment), permettant ainsi un topic modeling multilingue facilité.

5 Conclusion et travaux futurs

Dans cet article, nous avons développé deux modèles non paramétriques : l'Embedded Dirichlet Process et sa version hiérarchique, l'Embedded Hierarchical Dirichlet Process. L'EHDP surpasse les autres algorithmes de l'état de l'art, et montre également des signes de robustesse supérieures. En outre, l'EHDP peut gérer les stopwords et effectuer des regroupements dans un environnement multilingue. Enfin, nous notons que l'EHDP tend à sur-synthétiser les textes par rapport aux annotations humaines. Les travaux futurs viseront à ajouter des mécanismes pour inclure le multilinguisme et à obtenir des topics plus précis.

Références

Blei, D. et al. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

Blei, D. M. et al. (2017). Variational Inference : A Review for Statisticians. *Journal of the American Statistical Association 112*(518), 859–877. arXiv : 1601.00670.

Blei, D. M. et J. D. Lafferty (2007). A correlated topic model of Science. *The Annals of Applied Statistics 1*(1), 17–35. arXiv : 0708.3601.

Boyd-Graber, J. et al. (2017). Applications of Topic Models. *Foundations and Trends in Information Retrieval 11*, 143–296.

Dieng, A. B. et al. (2019). Topic Modeling in Embedding Spaces. *arXiv :1907.04907 [cs, stat]*. arXiv : 1907.04907.

Figurnov, M. et al. (2019). Implicit Reparameterization Gradients. *arXiv :1805.08498 [cs, stat]*.

Griffiths, T. L. et M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences 101*(Supplement 1), 5228–5235.

Horwood, G. V. (2017). reliefweb_corpus_raw_20160331.json. In *Humanitarian Assistance and Disaster Relief (HA/DR) Articles and Lexicon*. Harvard Dataverse.

Ishwaran, H. et L. F. James (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 13.

Kingma, D. P. et M. Welling (2014). Auto-Encoding Variational Bayes. *arXiv :1312.6114 [cs, stat]*.

Nalisnick, E. et P. Smyth (2017). Stick-Breaking Variational Autoencoders. *arXiv :1605.06197 [stat]*.

Ning, X. et al. (2020). Nonparametric topic modeling with neural inference. *Neurocomputing 399*, 296–306.

Palencia-Olivar, M., S. Bonnevay, A. Aussem, et B. Canitia (2021). Neural embedded Dirichlet Processes for topic modeling. In *Modeling Decisions for Artificial Intelligence*, pp. 299–310.

Teh, Y. W. et al. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

Vulić, I. et k. . others (2015). Probabilistic topic modeling in multilingual settings : An overview of its methodology and applications. *Information Processing & Management 51*(1), 111–147.

Summary

This paper presents two novel models: the neural Embedded Dirichlet Process and the neural Embedded Hierarchical Dirichlet Process. Both methods extend the Embedded Topic Model (ETM) to nonparametric settings, thus simultaneously learning the number of topics, latent representations of documents, and topic and word embeddings from data. To achieve this, we replace ETM's logistic normal prior with Dirichlet Processes in a variational autoencoding inference setting. Our tests on the 20 Newsgroups and on the Humanitarian Assistance and Disaster Relief datasets show that our models present the advantage of maintaining low perplexity while providing meaningful representations that outperform that of state of the art methods. We obtained our results without having to perform costly reruns to find the number of topics nor having to sacrifice a Dirichlet-like prior.