Learning to Bridge Colloquial and Formal Language Applied to Linking and Search of E-Commerce Data

Ivan Vulić, Susana Zoghbi, and Marie-Francine Moens
Department of Computer Science
KU Leuven, Belgium
{ivan.vulic, susana.zoghbi, marie-francine.moens}@cs.kuleuven.be

ABSTRACT

We study the problem of linking information between different idiomatic usages of the same language, for example, colloquial and formal language. We propose a novel probabilistic topic model called multi-idiomatic LDA (MiLDA). Its modeling principles follow the intuition that certain words are shared between two idioms of the same language, while other words are non-shared, that is, idiom-specific. We demonstrate the ability of our model to learn relations between cross-idiomatic topics in a dataset containing product descriptions and reviews. We intrinsically evaluate our model by the perplexity measure. Following that, as an extrinsic evaluation, we present the utility of the new MiLDA topic model in a recently proposed IR task of linking Pinterest pins (given in colloquial English on the users' side) to online webshops (given in formal English on the retailers' side). We show that our multi-idiomatic model outperforms the standard monolingual LDA model and the pure bilingual LDA model both in terms of perplexity and MAP scores in the IR task.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-Information filtering

Keywords

topic models; unstructured data; user interests; recommendation systems; user-generated data; personalized linking

1. INTRODUCTION

Colloquialisms are words or phrases employed in conversational or informal language, but not in formal writing. As more users become producers of content, a large part of the Web, like blogs, reviews and social media sites, may be full of colloquial expressions. In contrast, the language used elsewhere on the Web is more formal. News sites, online retailers or sites for expert knowledge are typical examples. For a given news article, an event may be described using formal language. On the other hand, users are free to comment on the article. Users' comments refer to the same event, but the language tends to be more informal. A similar situation applies to online retailers, where products tend to be described in a formal way. The reviews refer to the same product, but the language use is rather informal.

In this work, the goal is to bridge these two *idioms* of what is essentially the same language: colloquial and formal. This is, for instance, particularly important for retrieval tasks in e-commerce. For example, to find products, users may issue queries on retail sites like `Amazon.com` and `eBay.com`. A common problem is that the language and words chosen by the user may differ significantly from those in the product description. Suppose a user is looking for a particular kind of lamp, one that looks like a mushroom. To this user, it may seem like *"mushroom lamp"* would be a good query to find it. Note that the phrase *"mushroom lamp"* is actually found on social media sites to describe such items. However, on the retail site, the same product is described as *"Bramble Toadstool Nightlight Red"*. It is clear that users and retail sites may be talking about the same objects, but they choose different words to describe them. Although the product is relevant for the query, it may not be retrieved under a traditional document representation, such as bag-of-words. The idea is to develop a model that is able to learn how these two "languages" are related, and consequently has the ability to link knowledge from the users' side which uses the colloquial language to the target side given in the formal language.

The contributions of this paper are as follows. We propose, describe and evaluate a novel unsupervised topic model called multi-idiomatic LDA (MiLDA) which is able to deal with such multi-idiomatic data, taking into account both the knowledge of shared and non-shared words across two different idioms of the same language. We train our model on a dataset that contains `Amazon.com`'s product descriptions paired with the corresponding product reviews, and then test it in the task of linking users' Pinterest pins to relevant webshops [7]. After the inference on the test dataset, we demonstrate that our new model obtains lower perplexity as well as better mean average precision (MAP) scores than standard monolingual LDA (which does not distinguish between different idioms in the same language) and bilingual LDA (which treats two different idioms of the same language as two completely separate languages) which were trained on the same training collection. Moreover, we also outperform the best results previously reported in [7].

2. RELATED WORK

Monolingual and multilingual probabilistic topic models have been proven as a powerful unsupervised toolkit to an-

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609543.

alyze large text collections. However, no prior work has focused on capturing the evidence coming from multi-idiomatic data such as products descriptions (given in a formal language) which are naturally linked to their reviews (given in a colloquial user-centered language). On one hand, standard monolingual topic models such as LDA [1] do not distinguish between two different idioms of the same language at all. On the other hand, standard multilingual topic models such as bilingual or polylingual LDA [2, 4, 3] treat two different idioms of the same language as two totally separate languages. They do not take into account that a significant portion of words and phrases does not exhibit idiomatic usage and is shared across two idioms. The utility of multilingual topic models has been demonstrated when they are trained on multilingual aligned Wikipedia articles, but here we show that they have limited capability of dealing with different, multi-idiomatic data.

In [7], a new task of linking users' Pinterest pins to webshops has been proposed. Combining Pinterest data -given in the colloquial user-generated language- with Amazon webshops -given in the formal language- implies the need for multi-idiomatic text processing. [7] already demonstrated the utility of topical representations for linking models in this task, but they reported only preliminary results where a monolingual LDA model was trained on the target collection to improve overall retrieval results. In contrast, in this paper, we target to learn true "cross-idiomatic" topics on a training collection consisting of Amazon's product descriptions aligned to their respective reviews, and then infer these topics on the target collection for retrieval.

3. MULTI-IDIOMATIC TOPIC MODEL

Intuition. Assume you have a collection of document pairs. Each pair comprises a product description coupled with the corresponding users' reviews. The product description in the pair is written in formal language (e.g. written by retailers or experts) and the other in colloquial language (e.g., written by laymen or users). Given such collection, the goal is to find how these two language idioms (colloquial vs. formal) relate to each other. This knowledge might prove its potential in a task such as linking user-generated data (e.g., Pinterest pins) to online webshops [7].

Given this setup, our new *multi-idiomatic LDA (MiLDA)* model takes into account two main points: (1) A pair of documents shares the same distribution over topics (i.e., in essence, they talk about the same product). This is conceptually similar to a requirement from bilingual topic modeling [2, 4, 3], where a pair of news articles or Wikipedia articles discusses the same topics in two different languages; (2) A portion of words in the colloquial idiom differs from those in the formal idiom and vice versa. Moreover, a portion of words is shared between the two language idioms, and each document may be observed as a bag of shared and non-shared words combined together. This is conceptually different from [2, 4, 3], where it was assumed that each language has a unique set of words and no words are shared. In that case, given two languages, these bilingual topic models induce two unique sets of per-topic word distributions, each for one language. Here, we introduce the third set of per-topic word distributions, taking into account the knowledge of shared words. As a result, each latent "cross-idiomatic" topic is represented as a mixture of: (i) its idiom-specific per-topic word distributions over non-shared words; and (ii) idiom-shared per-topic word distributions (i.e., distributions

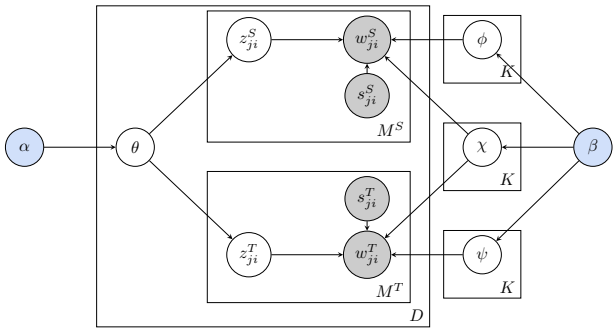


Figure 1: Graphical representation of the multi-idiomatic LDA (MiLDA) model in plate notation.

over shared words). **Description of the Model.** Fig. 1, shows the plate representation of our new MiLDA model. To address point (1) above, we consider that both documents in a pair have the same topic distribution θ, which is sampled from a symmetric Dirichlet with hyperparameter α. To address point (2), we consider three sets of per-topic word distributions: one unique to the colloquial idiom, (ϕ); one unique to the formal idiom, (ψ); and one common to both idioms, (χ). These distributions are independently drawn from a symmetric Dirichlet distribution with hyper parameter β.

Algorithm 3.1: GENERATIVE STORY FOR MILDA()

initialize: (1) set the number of topics K;
(2) set values for Dirichlet priors α and β;
(3) set values for s^S_{ji} and s^T_{ji};
sample K times $\phi \sim Dirichlet(\beta)$
sample K times $\psi \sim Dirichlet(\beta)$
sample K times $\chi \sim Dirichlet(\beta)$
for each document pair $d_j = \{d^S_j, d^T_j\}$
do
- sample $\theta_j \sim Dirichlet(\alpha)$
- **for each** word position $i \in d^S_j$
  - **do**
    - sample $z^S_{ji} \sim Multinomial(\theta)$
    - **if** $s^S_{ji} = 1$
      - sample $w^S_{ji} \sim Multinomial(\chi, z^S_{ji})$
    - **if** $s^S_{ji} = 0$
      - sample $w^S_{ji} \sim Multinomial(\phi, z^S_{ji})$
- **for each** word position $i \in d^T_j$
  - **do**
    - sample $z^T_{ji} \sim Multinomial(\theta)$
    - **if** $s^T_{ji} = 1$
      - sample $w^T_{ji} \sim Multinomial(\chi, z^T_{ji})$
    - **if** $s^T_{ji} = 0$
      - sample $w^T_{ji} \sim Multinomial(\psi, z^T_{ji})$

We use a superscript S or T to differentiate the idioms or languages, e.g., colloquial vs. formal, or more generally source (S) vs. target (T). s^S_{ji} is a precomputed indicator that reveals whether the word at position i is shared ($s^S_{ji} = 1$) or unique ($s^S_{ji} = 0$) to this idiom. As a simple heuristic, we assume that all words which occur on both sides of the given document collection are shared words. As a consequence in our model, s^S_{ji} and s^T_{ji} are fully observed because once we see the full corpus, it is trivial to determine whether a word position contains a shared word or not. Finally, algorithm 3.1 shows the full generative story of the MiLDA model.

Training and Output. We train our MiLDA model using Gibbs sampling [5]. For the source language S, if $s^S_{ji} = 0$:

$$P(z^S_{ji} = k|\Theta, s^S_{ji} = 0) \propto \frac{n^S_{j,k,\neg i} + n^T_{j,k} + \alpha}{n^S_{j,\cdot,\neg i} + n^T_{j,\cdot} + K\alpha} \cdot \frac{v^S_{k,w^S_{ji},\neg} + \beta}{v^S_{k,\cdot,\neg} + |V^S|\beta} \quad (1)$$

where $\Theta = (\mathbf{z}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^T, \mathbf{w}^S, \alpha, \beta)$. Whereas, if $s^S_{ji} = 1$:

$$P(z_{ji}^S = k|\Theta, s_{ji}^S = 1) \propto \frac{n_{j,k,\neg i}^S + n_{j,k}^T + \alpha}{n_{j,\cdot,\neg i}^S + n_{j,\cdot}^T + K\alpha} \cdot \frac{v_{k,w_{ji},\neg}^C + \beta}{v_{k,\cdot,\neg}^C + |V^C|\beta} \quad (2)$$

Analogous equations can be derived for documents in the target language. V^S and V^T denote vocabularies of non-shared words for the respective source and the target idioms, while V^C is a vocabulary of words shared between the two idioms. $n_{j,k}$ denotes the number of times topic k is assigned within document d_j; while $n_{j,k,\neg i}$ has the same meaning but not counting the current assignment to w_{ji}. $v_{k,w_{ji},\neg}$ counts the number of times a word w_{ji} has been assigned to topic k, not counting the current position. When a dot (".") appears in the subscript of a variable, it means that we sum over all the possible values of the corresponding variable[1].

After the burn-in period, we obtain the document-topic distribution, $P(z_k|d_j)$, which indicates the probability of each topic k in a document d_j.

$$P(z_k|d_j) = \frac{n_{j,k}^S + n_{j,k}^T + \alpha}{n_{j,\cdot}^S + n_{j,\cdot}^T + K\alpha} \quad (3)$$

We also obtain three per-topic distributions: one for the colloquial (or source) idiom words; one for the formal (or target) idiom words; and one for the shared words between the languages. Each of these indicate the probability of a word given a topic, as shown by eq. (4):

$$P(w_i^L|z_k) = \frac{v_{k,w_i^L}^L + \beta}{v_{k,\cdot}^L + |V^L|\beta} \quad (4)$$

where the index L may refer to S, T or C.

4. LINKING PINS TO WEBSHOPS - EXPERIMENTAL SETUP

Evaluation Task: Linking Pins to Webshops. To evaluate the utility of the new multi-idiomatic topic model, we perform the same retrieval task as proposed in [7]. The task is to link users' pins posted on the social media site `Pinterest.com` given by their textual description to a set of relevant webshops where the user might search for or buy the "pinned" product. It is essentially a retrieval task in which, given the pin as a query [7], one obtains a list of relevant webshops.

Target Collection and Queries. The target collection is obtained from [7] and consists of 19,955 product descriptions from `Amazon.com` grouped into 1,171 webshops. One product may belong to more than one webshop. The query set consists of 50 users' pins in text format. An example query is "Be daring, go all out in red! Modern Jessica Rabbit".

Training Collection. We use a training dataset that organically aligns documents written in colloquial and formal language. The training dataset consists of 15,566 aligned document pairs, where each pair consists of: (1) a description of a product acquired from `Amazon.com` (formal language), (2) a merged collection of top 10 most helpful users' reviews for that particular product (colloquial language).

Models for Comparison. We experiment with three different topic models that capture three different paradigmatic approaches to the data: (1) monolingual LDA (which does not exploit the natural links in the training dataset and treats all documents as given in only one language), (2) bilingual LDA (which observes two idioms as two separate languages, and uses the alignment in the training dataset), (3) our new multi-idiomatic LDA. All models have been trained with the same number of topics (K=100, 200, 500, 800, 1000, 1200, 1500) with the same number of iterations (1000) on the same training collection with the same parameter setup, as hyperparameters are set to the standard values $\alpha = 50/K$ and $\beta = 0.01$, according to [6, 5]. The trained models are then inferred on the previously unseen target collection and we test their utility in the task of linking pins to webshops.

Retrieval Models. To perform the actual retrieval, we adapt a well-known LDA-based probabilistic unigram retrieval model from [6], which combines a topical representation (*tr*) with the regular count-based bag-of-words (*bow*) document representation (see [6] for more details and parameter settings):

$$P(Q^S|d_j^T) = \prod_{i=1}^{m} P(q_1^S, \ldots, q_m^S|d_j^T) = \prod_{i=1}^{m} P(q_i^S|d_j^T)$$

$$= \prod_{i=1}^{m} P_{bow+tr}(q_i^S|d_j^T) = \prod_{i=1}^{m} (\lambda P_{bow}(q_i^S|d_j^T) + (1-\lambda) P_{tr}(q_i^S|d_j^T))$$

where Q^S is a query (i.e., a pin) containing m query words $q_1^S, \ldots, q_m^S$ given in the source idiom (i.e., colloquial language), d_j^T a j-th document from the target collection given in the target idiom (i.e., formal language), and λ is the interpolation parameter. The "topical" contribution $P_{tr}(q_i^S|d_j^T)$ is computed as follows: $P_{tr}(q_i^S|d_j^T) = \sum_{k=1}^{K} P(q_i^S|z_k) P(z_k|d_j^T)$. The probability $P(z_k|d_j^T)$ is known after a topic model is inferred on a target document d_j^T. When performing retrieval with MiLDA, we introduce one major difference when computing $P_{tr}(q_i^S|d_j^T)$. Namely, we can propagate the knowledge of shared and non-shared words from the training to our target collection. When the query word q_i^S happens to be a shared word in the training collection, we compute the probability by using the estimated per-topic word distributions for shared words (χ, see eq. (4)). Otherwise, we rely on the idiom-specific per-topic word distributions (ϕ).

5. RESULTS AND DISCUSSION

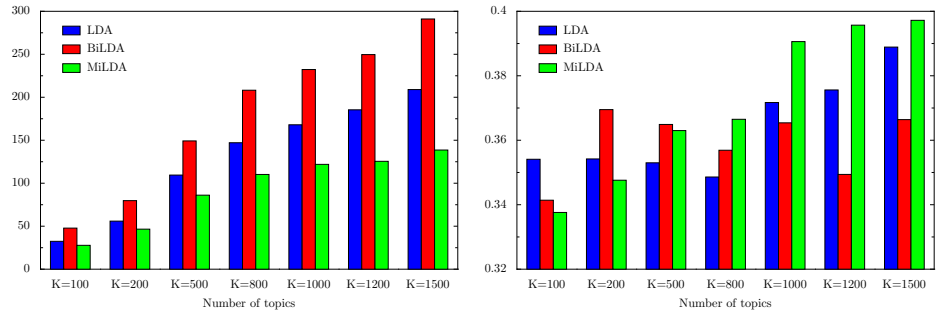
Test 0: Qualitative Evaluation. Tab. 1 shows example lists of the top 5 frequent words in the word distributions for three topics: photography, coffee and tanning. For instance, the shared vocabulary distribution (χ) contains words likely expected in a photography topic, such as *lens*, *focus* and *canon*. In the users' language side, we -as non experts in the subject- learned new words that belong to this topic. For example, *bokeh* means blur or the aesthetic quality of the blur of an image. It refers to "the way the lens renders out-of-focus points of light". For the topic *coffee*, the shared word distribution shows again typical words, such as *espresso*, *press* or *beans*. On the users' side, we again learn new words associated with the topic. *Illy*, *tierra*, *robusta* and *gaggia* are all brands related to coffee or coffee machines. This is also useful for merchants interested in learning about leading or most-talked-about brands.

Test I: Perplexity. A standard way to compare the quality of topic models is the perplexity measure, an intrinsic evaluation metric that evaluates the topic model's capability of predicting previously unseen documents [1]. A lower perplexity score implies that the model provides a better explanation and a better representation for unseen documents. Fig. 2(a) compares the perplexity scores for the three models in comparison (LDA vs. BiLDA vs. MiLDA). We

[1]For example $n_{j,\cdot} := \sum_{k^*=1}^{K} n_{j,k^*}$

Table 1: Example of the top 5 words on the per-topic word distributions for K = 500: shared vocabulary distribution, users' (reviews-only) vocabulary distribution

shared vocabulary	users' vocabulary	shared vocabulary	users' vocabulary	shared vocabulary	users' vocabulary
(photography)	(photography)	(coffee)	(coffee)	(tanning)	(tanning)
lens	bokeh	espresso	illy	tan	tanners
gopro	tamron	machine	tierra	skin	rebirthing
focus	primes	press	robusta	lotion	comatose
canon	apertures	coffee	gaggia	self	patchy
light	xti	beans	brikkas	tanning	jergens



(a) Perplexity scores (b) Maximum MAP scores

Figure 2: Evaluation results and comparison of three different topic models in terms of (a) perplexity scores (lower is better), and (b) maximum MAP scores (higher is better) in the task of linking pins to webshops.

may observe that MiLDA consistently outscores the other two models which implies that the true multi-idiomatic text processing as modeled by MiLDA is more beneficial than monolingual (LDA) or multilingual (BiLDA) in this setting.

Test II: Linking Pins to Webshops. In another evaluation test, we measure the ability of the retrieval models from sect. 4 to link relevant webshops to queries/pins given the three different topical representations of target collection documents (again, LDA vs. BiLDA vs. MiLDA). We can again observe that MiLDA outperforms the other two models, which again implies that distinguishing between idiom-specific and idiom-shared words is essential and yields better scores in this application. The maximum MAP scores of 0.396 ($K = 1000$) and 0.398 ($K = 1500$) with MiLDA were obtained by $\lambda = 0.5$. It reveals that both document representations are important for retrieval. Furthermore, since the MAP scores when only the *bow* representation or only the *tr* representation is used in retrieval are 0.341 and 0.359 ($K = 1200$), we observe that combining the two representations leads to a positive synergy in the combined model.

We could further improve the results in a slightly altered retrieval setup, where, instead of retrieving webshops directly, we have first ranked single products according to their relevance to the pin and then provided a score for a webshop as an average over its top B best scoring products. Here, we have not observed any major qualitative change with respect to the relation of MAP scores for LDA-, BiLDA-, and MiLDA-based retrieval models. However, our best scoring MiLDA-based model in this setup produced a MAP score of 0.441 ($K = 1500, B = 5, \lambda = 0.7$) which is better than the previous best reported result in [7] (MAP: 0.414), despite that their LDA was trained directly on the entire target collection.

6. CONCLUSIONS AND FUTURE WORK

We have proposed a new topic model called multi-idiomatic MiLDA which is capable of dealing with multi-idiomatic data and linking information between different idiomatic usages of the same language, for example colloquial and formal language. We showed that we can learn true cross-idiomatic topics on a training collection consisting of Amazon's product descriptions aligned to their respective reviews, and then infer these topics on the target collection for retrieval.

Our results in the task of linking pins to webshops reveal the advantage of the multi-idiomatic MiLDA model over the monolingual LDA model and the bilingual LDA model both in terms of perplexity and MAP.

Since this work on multi-idiomatic text processing is only a start, we strongly believe that our modeling approach will ignite more future applications. For instance, we foresee that the MiLDA model might prove extremely valuable in opinion mining, sentiment analysis, e-commerce applications, social media analysis, dialect mining, or cross-lingual information retrieval for closely related languages.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] W. De Smet and M.-F. Moens. Cross-language linking of news stories on the web using interlingual topic modelling. In *Proc. of the CIKM SWSM Workshop*, pages 57–64, 2009.

[3] D. Mimno, H. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *EMNLP*, pages 880–889, 2009.

[4] X. Ni, J.-T. Sun, J. Hu, and Z. Chen. Mining multilingual topics from Wikipedia. In *WWW*, pages 1155–1156, 2009.

[5] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.

[6] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185, 2006.

[7] S. Zoghbi, I. Vulić, and M.-F. Moens. Are words enough?: A study on text-based representations and retrieval models for linking pins to online shops. In *CIKM UnstructureNLP Workshop*, pages 45–52, 2013.