Information Sciences 367–368 (2016) 573–599

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Latent Dirichlet allocation for linking user-generated content and e-commerce data

Susana Zoghbi*, Ivan Vulić, Marie-Francine Moens

Computer Science Department, KU Leuven, Celestinenlaan 200A, Heverlee, Belgium

ARTICLE INFO

Article history:
Received 25 May 2015
Revised 12 May 2016
Accepted 29 May 2016
Available online 15 June 2016

MSC:
00-01
99-00

Keywords:
Multi-idiomatic text processing
Probabilistic topic models
Linking and recommendation
IR For e-commerce
Pinterest data
Amazon data

ABSTRACT

Automatic linking of online content improves navigation possibilities for end users. We focus on linking content generated by users to other relevant sites. In particular, we study the problem of linking information between different usages of the same language, e.g., colloquial and formal *idioms* or the language of consumers versus the language of sellers. The challenge is that the same items are described using very distinct vocabularies. As a case study, we investigate a new task of linking textual *Pinterest.com* pins (colloquial) to online webshops (formal). Given this task, our key insight is that we can learn associations between formal and informal language by utilizing aligned data and probabilistic modeling. Specifically, we thoroughly evaluate three different modeling paradigms based on probabilistic topic modeling: monolingual latent Dirichlet allocation (LDA), bilingual LDA (BiLDA) and a novel multi-idiomatic LDA model (MiLDA). We compare these to the unigram model with Dirichlet prior. Our results for all three topic models reveal the usefulness of modeling the hidden thematic structure of the data through topics, as opposed to the linking model based solely on the standard unigram. Moreover, our proposed MiLDA model is able to deal with intrinsic multi-idiomatic data by considering the shared vocabulary between the aligned document pairs. The proposed MiLDA obtains the largest stability (less variation with changes in parameters) and highest mean average precision scores in the linking task.

1. Introduction

The Web is driven by links between content. Search engines rank relevant documents to an information need based largely on content and how items are linked. Often, links are explicit. However, since the Web is a very large space, all similar content is not explicitly linked. There is an unprecedented amount of user-generated content; people often use social media sites to express things they are interested in, creating a large wealth of information. In the absence of explicit links between similar Web items, we want to automatically link sources that refer to related content. This can improve the ability to find relevant data and to meet users' information needs.

We focus on linking content generated by users to other relevant sites. On social media sites, it is common practice to suggest or recommend items and relationships (friends). Many techniques are used. However, in most cases, suggested items live within the same source or ecosystem where the user has posted the information. For example, Facebook suggests other Facebook pages, or products that have signed up for Facebook ads. Twitter suggests to follow other Twitter users or

* Corresponding author.
E-mail address: susana.zoghbi@cs.kuleuven.be, susana.zoghbi@gmail.com (S. Zoghbi).

http://dx.doi.org/10.1016/j.ins.2016.05.047

to retweet. Pinterest[1] suggests to follow pin boards and Pinterest users. Furthermore, online stores like Amazon and eBay also perform product recommendations within their collection of known items.

For this work, we propose to link items between quite different sources or ecosystems. For example, a tweet about a popular soccer match may be linked to other related articles elsewhere on the Web, like sports sites. This allows for increased connectivity between items, so that we explicitly bridge latent relationships, creating links that did not exist before and effectively identifying similar items that live in different ecosystems.

Our experiments will focus on linking items from the social site Pinterest.com to online products. This application is helpful for both users and businesses. For example, a Pinterest user might post an item that she would like to buy, but may not know where to buy it. In this case, it is useful to have a system that can automatically recognize the content of the pin and suggest online stores outside of Pinterest where the item (or similar ones) can be bought. Similarly, an online store might wish to find people interested in products resembling the ones in the store.

For this linking task, the challenge is that the language from social ecosystems is full of colloquialisms, informal expressions and subtle or implicit references. For example a pin on Pinterest says *love the whole matrix look*. It expresses a positive sentiment for a long black leather coat and pants, as the outfits worn by the characters in the movie "The Matrix". However, this is not explicitly stated. Instead, we need to infer that *matrix look* refers to a certain style of clothing. As more users become content producers, a large portion of the Web fills up with such language expressions. In contrast, elsewhere on the Web, the language usage is more formal, references are clear and explicit. News sites, online retailers and sites for expert knowledge are good examples.

In this work, we want to bridge these two *idioms*[2]. Our hypothesis is that people talk about the same objects using different words depending on the context, degree of formality required, expertise level, etc. We ask questions like *Can we learn associations between two different language expressions or idioms?Can we learn objects' attributes using examples of item descriptions by different people in different contexts? Can we automatically link similar items found on very different sources?*

To address these questions, one of our key insights is that we can use pairs of *aligned documents* that discuss similar items, and employ different vocabularies. For example, product descriptions aligned with the corresponding users reviews provide a set where the descriptions tend to use a rather formal vocabulary and the reviews are presented informally. Other examples include news articles aligned with users' comments about the articles, scientific articles aligned with layman's explanations, or historical events aligned with discussions by the public. Leveraging aligned cross-idiomatic information is a novel approach to this problem.

While the Web is full of such documents, to our knowledge there are no datasets readily available to learn how different expressions of language are associated. For this work, we collected a novel dataset of pairs of aligned documents, where each pair consists of a document written in formal language and one written in colloquial language. In particular, we present a collection of product descriptions (the seller language) and product reviews (consumer language) from Amazon.com. Based on such corpus[3], we show that we can learn to bridge two expressions of the language.

To learn, we require advanced textual representations, where the semantics of related words are captured. The common unigram model (i.e., representing a text by the single words it contains) fails to encode the relationship between multiple semantically related terms. There exist many techniques that can be applied to this problem. One approach uses linguistic associations to find related words from a thesaurus like WordNet. However, such a thesaurus does not contain many popular and informal expressions like the ones used on social media or users reviews. Another approach computes term-to-term similarities. However, each term is encoded as an atomic word and concepts are not fully captured (please refer to Section 2 for an in-depth survey of the literature). In contrast, the family of Latent Dirichlet Allocation (LDA) models allows to softly cluster related terms into topics. Each document in the corpus may be represented as a soft collection of terms. This yields a richer and more flexible representation for our task. In particular, under the monolingual LDA model, a pair of aligned documents can be concatenated into one unique document and we can learn common topics that span over one common vocabulary.

Furthermore, LDA-based models [4,13,37] allow us to elegantly produce a mathematical representation of our main intuition, i.e., for a pair of aligned documents, people essentially talk about the same items using different words. Concretely, using ideas from the bilingual latent Dirichlet allocation (BiLDA), we can model the pair alignment by a shared topic distribution (i.e., similar topics are discussed). Under the BiLDA model, the consumer and the seller idioms are considered distinct languages. This allows to explicitly model the differences in the language; in contrast with the monolingual LDA that treats consumer and seller idioms as one common language. While the assumption of two completely distinct languages does not quite hold in our case (since seller and consumer idioms have many terms in common as well as many unique terms), we will see that BiLDA yields to some improvements over LDA for this task, as it considers a shared topic distribution and it explicitly models the differences in language usage between the two sources.

As an alternative to the monolingual LDA and bilingual LDA, we propose a novel topic model that aims to represent the multi-idiomatic nature of the data with more flexibility. We name this model multi-idiomatic latent Dirichlet allocation (MiLDA). Specifically, we propose to retain the assumption from BiLDA that aligned documents have the same topic dis-

[1] If the reader is not familiar with Pinterest.com, please refer to Section 8.

[2] The word idiom may be defined as the language peculiar to a people or to a district, community, or class. We use "multi-idiomatic" to denote multiple usages in language between different Web ecosystems.

[3] Datasets available at https://people.cs.kuleuven.be/~susana.zoghbi/

tribution as they discuss the same items. However, unlike BiLDA we explicitly model not only the differences in language but also the similarities. To this end, we propose to partition the vocabulary into three vocabulary distributions: one shared vocabulary that captures the similarities between the sources and two non-shared vocabularies that are unique to each of the sources: seller and consumer.

We evaluate these models experimentally by performing the task that we set out to do: to link content between Web sources with significantly different language usage. An example of this task is to link user-generated content from social media (e.g., Pinterest) to e-commerce data (e.g., Amazon). We propose to perform this linking as an information retrieval task, where posts from social media are used as queries and products are used as a target collection. To our knowledge, this has not been studied before, and presents a novel approach to product recommendation between different ecosystems. Our results indicate that the LDA-based models outperform the unigram model. Additionally, our proposed MiLDA achieves higher scores and more stability when compared to LDA and BiLDA.

Our full experimental approach consists of four phases: training, inference, retrieval and evaluation. During training, we use the data we have collected from Amazon consisting of product descriptions aligned with users' reviews as a collection to learn topics containing both seller and consumer expressions. We train using four different models: the unigram model with Dirichlet prior, monolingual LDA, bilingual LDA and the new multi-idiomatic LDA. We then infer (inference phase) our learned representations onto a target collection that consist of webshops (simulated collections of products) and individual products. For retrieval, we use 100 randomly selected pins from Pinterest as queries and retrieve a ranked list of webshops or products from Amazon. Using ground truth generated by human annotators, we quantitatively evaluate the output of our system with common information retrieval measurements.

More concretely, the contributions of this article are as follows:

- We propose, perform and assess a novel task that links related items between significantly different sources where the context and expression of language differ, for example, social media sites and online retailers.
- We propose, collect and analyze a novel corpus of aligned data (product descriptions and reviews) to train models to bridge consumer and seller vocabularies.
- To successfully link items from different sources in a cross-idiomatic setting, we require advanced textual representations. We perform a thorough theoretical and didactic comparative study of different modelling approaches.
- We propose, describe, formally derive and evaluate a novel unsupervised topic model called multi-idiomatic LDA (MiLDA) which is able to deal with intrinsic multi-idiomatic data, taking into account both the knowledge of shared and non-shared vocabulary across two different idioms of the same language.
- We perform a systematic empirical comparison and analysis of different probabilistic topic models, and discuss the different modeling approaches and premises in the context of the given data considering its "monolinguality", "multilinguality" or "multi-idiomacy".

The rest of this article is organized as follows. Section 2 presents related work. Section 3 depicts an overview of the full information retrieval pipeline in this context. Section 4 formally describes the task. Section 5 and Section 6 present a detailed theoretical and comparative study of these models, starting from the simple unigram model with Dirichlet prior, following with the monolingual and bilingual LDA and ending with our newly proposed multi-idiomatic LDA. Section 7 presents the retrieval model. Section 8 describes the datasets. Experiments are described in Section 9 and results are presented in Section 10. The article ends with discussions, Section 11, and conclusions and future work, Section 12.

2. Related work

On one hand this work introduces the novel task of linking pins to webshops, which is related to the tasks of automated hyperlinking and content-based recommendation. On the other hand we propose a novel methodology of cross-idiom modelling for attaining this goal.

From the early days of the Internet one has dreamed of automatically generating hyperlinks [38] because automatic linking of content improves navigation possibilities for the user. Given its difficulty, the problem of automated hyperlinking still receives research interest today. For example, automatic linking is becoming popular as reading aids of texts that are difficult to understand, where complex terminology is linked to simpler definitions and explanations [32]. The problem of automatic linking of content on the Web has also been studied in the context of wikification, e.g., [18,34,35,43] or entity linking [5,12,22,23], where the goal is to link salient terms (typically named entities) from unnanotated raw text to target knowledge bases such as Wikipedia to provide additional background knowledge to the given text and consequently enrich the reader's experience. In another recent example Aggarwal et al. [1] present a tool for automating Web-navigation support that computes semantic similarity between the user goal and the website information. The problem we deal with in this article is substantially different. We focus on linking items, posted by Pinterest users, to online webshops. Pinterest's goal is to connect everyone in the world through the "things" they find interesting, making it a valuable study object for sociologists, linguists, online retailers, marketeers and others. Statistics of user characteristics and characteristics of followers (e.g., gender, age, nationality) are presented in [20]. The act of pinning and commenting is considered analogous to indexing and abstracting [58]. This author also remarks that the social collecting activities may offer a new way of viewing user-centered categories, much like the one provided by earlier work with folksonomies and tagging. Click behavior of Pinterest users has already been studied extensively, showing that a higher proportion of Pinterest users click through to e-commerce sites, and when they go there, they spend significantly more money than people who come from sites like Facebook [20]. In this respect, the new task of automatically linking Pinterest data with relevant webshops data that we propose in this article is of great practical value.

This novel linking task has some resemblance to content-based recommendation systems in e-commerce where the personalized recommender system receives information from a customer about which products she is interested in, and recommends products that are likely to fit her needs [31,52]. In many of the recommendation systems similar items are suggested as items that customers often bought together. In this work in order to better personalize the recommendation, we work with *unstructured* textual content data as found on social network sites such as Pinterest, and we completely automatically link a user's post, in our case a pin, to a relevant webshop. It is sometimes acknowledged that the Web has transformed products into search goods.

In analogy with the evaluation of recommendation techniques that are evaluated as information retrieval models [3,11], in this paper automated hyperlinking is evaluated as a retrieval problem, where relevant documents are ranked according to the personal interest of the user.

With regard to the methodologies for automated linking of textual content, many existing approaches are rather straightforward. For example, approaches to hyperlinking are often restricted to clustering of content [8] or linking via a shared terminology or common named entities [7], which in our case are not useful as the e-commerce consumers use a quite different language than the sellers. To deal with vocabulary differences, several approaches have been proposed, including text normalization, query reformulation, search results clustering, and automatic query expansion. The interested reader may refer to [60] for a more in-depth treatment. Automatic query expansion expands the original query with the aim to produce a query that is more likely to retrieve relevant documents. There are many ways to generate expansion candidates for queries [9]. One approach is to rely on linguistic associations, to find synonyms and related words of a query word from a thesaurus, usually WordNet. A thesaurus like WordNet has not caught up with consumer terms in use in social media. Mining user query logs may also be used to generate query expansion features. Given the difficulty of having access to query logs or to any other form of user feedback, we did not compare to such an approach in this paper.

In the past, several methods have been studied that group semantically related words into statistical concepts or topics. The most prominent methods regard latent semantic analysis (LSA) [15], probabilistic latent semantic analysis (pLSA) [24] and latent Dirichlet allocation (LSA) [4]. Whereas LSA models topics as the results of the factorization of the document-term matrix of a document collection, pLSA and LDA are generative probabilistic models that model each document of a collection as a mixture of topics and a topic as a mixture of words. The Web navigation tool mentioned earlier [1] uses the technique of LSA to compute semantic similarity between the user goal and the website information. To bridge the vocabularies of different domains, pLSA has been used in [54,55]. Compared to the above models, LDA-based models have over the years proven their superiority for probabilistic topic modelling by the use of priors as hyperparameters and their easy way for inferring the topic distributions of documents not seen during training. The LDA models are often trained with Markov Chain Monte Carlo sampling techniques such as Gibbs sampling, or variational inference, which are less prone to getting stuck in local maxima than the expectation-maximization based learning used in [54,55]. LDA-based models have been used for modelling the semantic topics of user-generated content. For instance, they have become popular as representation medium of Twitter and microblog messages [25,33]. For instance, [28] explored topic models for analyzing disaster-related Twitter data and [33] investigated how to improve topic models given the short and messy texts on tweets. Regarding Pinterest data, some work has been done to perform board recommendations [27], and implementations of topic models to understand users' interests [41]. However, there is little work regarding product recommendations in this setting. [46] propose a LDA model called Foreground and Background LDA (FB-LDA), to distill foreground topics and filter out long-standing background topics. The foreground topics can give potential interpretations of the sentiment variations expressed in blogs. [57] utilized LDA to study the problem of diversifying the lists of recommended products obtained from eBay given a user's query. The idea was to balance between the core relevance of the recommended items and the information novelty. More closer to our work, LDA-based models have been used to bridge word usage in cross-domain document classification [2,39,56]. All the above works make use of common monolingual LSA, pLSA or LDA models, i.e., documents of different domains or written in a different language are considered in one training collection. It has been shown that when dealing with bilingual or multilingual content, multilingual topic models trained on document-aligned comparable or parallel corpora [13,36,37] better capture the semantic relationships between words of a different language [47]. As we align in our work a product description with a consumer review, a similar finding will be confirmed when dealing with the different idiomatic expressions of the consumers and sellers language.

Monolingual and multilingual probabilistic topic models have been proven as a powerful unsupervised toolkit to analyze 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, here the language of the seller) which are naturally linked to their reviews (given in a colloquial user-centered language, here the language of the consumer). On one hand, standard monolingual topic models such as LDA [4] 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 [13,36,37] 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 is shared across the seller and the consumer language. In this paper, we provide a successful LDA-based model that provides a solution to this problem.

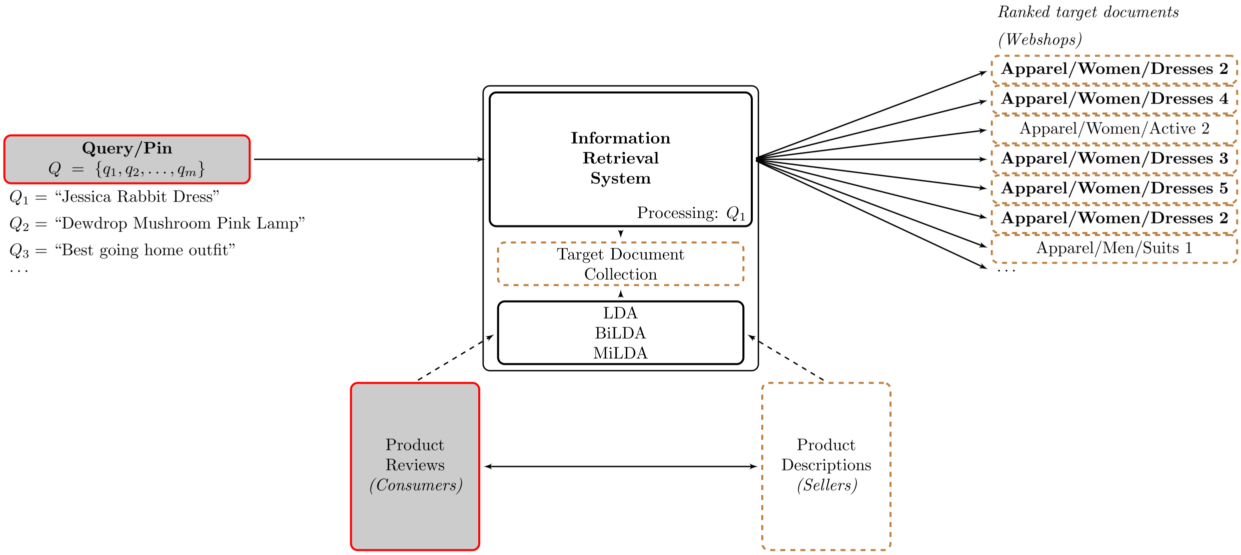

Ranked target documents
(Webshops)
Apparel/Women/Dresses 2
Apparel/Women/Dresses 4
Apparel/Women/Active 2
Apparel/Women/Dresses 3
Apparel/Women/Dresses 5
Apparel/Women/Dresses 2
Apparel/Men/Suits 1
. . .
Query/Pin
$Q = \{q_1, q_2, \ldots, q_m\}$
Q_1 = "Jessica Rabbit Dress"
Q_2 = "Dewdrop Mushroom Pink Lamp"
Q_3 = "Best going home outfit"
. . .
Information Retrieval System
Processing: Q_1
Target Document Collection
LDA
BiLDA
MiLDA
Product Reviews (Consumers)
Product Descriptions (Sellers)

Fig. 1. Information retrieval pipeline overview.

In [48,61,62], the authors have already introduced the new task of linking users' Pinterest pins to webshops, they have reported individual results obtained with different LDA-based models. However, the above works only reported preliminary results. In this article we provide a full theoretical analysis of the models and we significantly extend the experiments to include more queries and a larger target collection.

3. Overview

The key concepts of the entire information retrieval (IR) pipeline are illustrated by Fig. 1. Shaded rectangles depict text items considered to have been written by users (i.e., the consumers' idiom), while dashed rectangles refer to the sellers' idiom. Each pin posted by the user is treated as a query which is issued against a target document collection. The target documents may be pure textual descriptions of individual products or collections of products called *webshops*, as in the diagram. Examples *pins/queries* are "Jessica Rabbit Dress" or "Best going home outfit". The task of the information retrieval system (IRS) is to rank the target documents according to their relevance to the issued query. Therefore, the IRS outputs a relevance-based ranked list of documents. In this toy example, after processing the query Q1, the IRS outputs webshops labeled as "Apparel/Women/Dresses2" or "Apparel/Women/Dresses4" as the most relevant to the processed query. The names of the webshops which are indeed relevant to the query (as provided in relevance assessments) are given in bold, while all other webshops are considered irrelevant to the query.

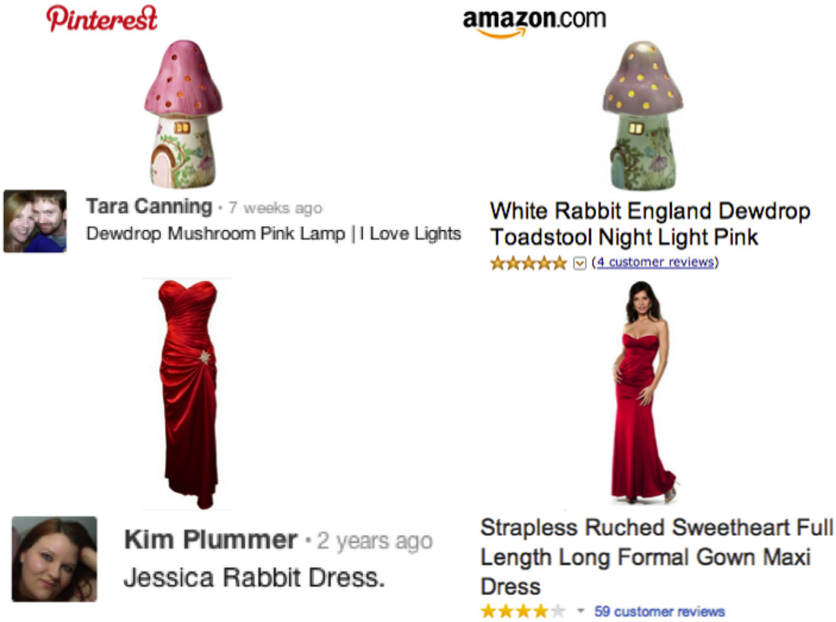
In order to link each pin to the relevant documents, the IRS system has to learn a structured semantic representation for each target document. These semantic representations are induced using three different paradigms from the topic modeling framework (LDA, BiLDA and MiLDA, which we will discuss in great detail in following sections). In short, a topic model is trained on the data set containing product descriptions (retailers' idiom) coupled with users' reviews for the same products (users' idiom), where each topic model exploits the aligned descriptions-reviews data set using different assumptions. The induced topic model is then run on the target collection, and each document may be structurally presented as a distribution over the induced latent topics. In addition, each query word belongs to each topic with a certain probability, and the IRS may quantify the probability that each target webshop generated the query word. That way, the topic modeling framework by means of the LDA, BiLDA, and MiLDA models serves as the semantic link in the IRS between each query and the target collection. To train the topic models, we use Gibbs sampling, an iterative algorithm that allows us to discover the latent semantic document structure. A high level depiction is presented in Algorithm 1. More details are presented later.

4. A cross-idiomatic linking task: from social media to online shops

We propose the task to link multi-idiomatic (yet related) content beyond their native ecosystem. We demonstrate this by linking pins from Pinterest.com to sets of online products from Amazon.com. (The task and techniques may be extended to other sources.) Fig. 2 shows visual examples of items that may be linked between different environments. We see that a user has expressed her interest in a long red formal dress, but she chooses to describe it as *Jessica Rabbit dress*, presumably because it is similar to the dress worn by the famous cartoon character. A similar dress is found on Amazon.com, but it is described as *Strapless Ruched Sweetheart Full Length Long Formal Gown Maxi Dress*. Another example: a user expresses

Algorithm 1: Learning Algorithm.

```
Initialize: (1) Distributions of probabilities (usually uniform distributions)
repeat
    for each word token in each document in the textcollection do
        Estimate the probability of assigning the current word token to eachtopic, conditioned on the topic assignments
        of all other word tokens (the updating step) - see equations 6, 12, 13, 29, 30;
        From this conditional distribution, a topic is sampled and stored as the new topic assignment for this word
        token (the sampling step);
    end
until the equilibrium state is reached;
```



Pinterest

Tara Canning · 7 weeks ago
Dewdrop Mushroom Pink Lamp | I Love Lights

Kim Plummer · 2 years ago
Jessica Rabbit Dress.

amazon.com

White Rabbit England Dewdrop Toadstool Night Light Pink
★★★★★ (4 customer reviews)

Strapless Ruched Sweetheart Full Length Long Formal Gown Maxi Dress
★★★★☆ 59 customer reviews

Fig. 2. Example of items that may be linked between different Web ecosystems (Pinterest and Amazon). We see the difference in language usage between social media posts (left) and online store products (right). On each row, the items are the same (or very similar), but the textual description differs. This difference in language makes it difficult to link the items as referring to the related objects. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interest in a particular kind of lamp (top row Fig. 2). The user chooses to describe it as *Dewdrop Mushroom Pink Lamp*. A similar product can be found in Amazon. However, it is described differently as *White Rabbit England Dewdrop Toadstool Night Light Pink*. We wish to link these items and exploit the similarities between them. For this work, we focus on linking textual sources and leave other modalities (such as images and videos) for future work.

We frame our linking task as an ad-hoc information retrieval task. Concretely, given an information unit (e.g., a single post or pin) from a user, we retrieve (or link) relevant documents where users can find more information related to the original information unit. In the particular example of Pinterest users, a text of a single pin is used as a *query* and online products (either individually or in groups) are used as the *target document collection*. We rank the documents according to their relevance to the query.

Formally, let $\mathcal{D} = \{d_1, d_2, \ldots, d_L\}$ be a target collection of L webshops and Q a textual content of a pin, that is, a query given by the set of m words in the pin/post $Q = \{q_1, q_2, \ldots, q_m\}$. Documents are ranked by the probability $P(Q|d_j)$ that a query Q was generated by a given document model d_j as follows

$$P(Q|d_j) = \prod_{i=1}^{m} P(q_i|d_j). \tag{1}$$

This approach corresponds to the well-known query likelihood model. This probabilistic language modeling framework has been proved effective in ad-hoc retrieval tasks [30,53].

Algorithm 2: Generative story for the unigram model with a Dirichlet prior

```
for each document $d_j$ do
    sample $\theta_j \sim Dirichlet(\alpha_t)$;
    for each word position $i \in d_j$; do
        sample $w_{ji} \sim Multinomial(\theta_j)$;
    end
end
```

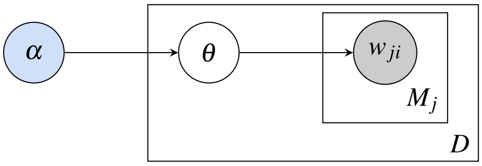


Fig. 3. Plate representation of unigram model with Dirichlet prior. The outer plate denotes D documents. θ denotes the probability distribution of words within the document. The inner plate represents the choice of words (w_{ji}) in the *ith* position of the *jth* document. M_j denotes the number of words in the *jth* document.

5. Modeling cross-idiomatic sources

In light of our task, in this section we compare three existing models: unigram with Dirichlet prior, latent Dirichlet allocation (LDA) and bilingual latent Dirichlet allocation (BiLDA). In addition, we present a new probabilistic topic model, the multi-idiomatic[4] latent Dirichlet allocation (MiLDA) in the next section.

LDA-based models combined with probabilistic language models provide state-of-the-art results for information retrieval tasks [51]. Additionally, the family of LDA-based models produces an elegant probabilistic representation of our main intuition on pairs of aligned documents, i.e., they discuss the same topics using different words.

As we will see in this section, the monolingual LDA concatenates each pair of documents and learns common topics that contain different terms that arise from distinct users. BiLDA keeps separately each document in the pair (no concatenation takes place) and assumes a shared topic distribution; while there are two unique vocabularies (seller and consumer). In this section we perform an in-depth theoretical review of these models and how they relate with the unigram model with Dirichlet prior and among each other.

In general, we can model a textual document from any source by the probability distribution likely to have generated the document, $P(w_t|d)$, where w_t is the *t-th* word in the vocabulary. Each model essentially differs on how $P(w_t|d)$ is represented given the model assumptions.

Given an entire corpus containing documents from different sources (i.e., expressed in multiple idioms), we want to learn a probability distribution $P(w_t|d)$ that is well suited to perform our task. All the models (except the unigram) are trained on a collection of aligned document pairs consisting of product descriptions and reviews.

5.1. Unigram model with Dirichlet prior

The unigram model with a Dirichlet prior is the simplest representation we consider here. It proposes that each word w_{ji} at position i in each document d_j is drawn independently from a single multinomial distribution with parameter θ_j, such that the probability of a word $P(w_t|d_j) = \theta_{jt}$. The parameter θ_j is sampled from a Dirichlet distribution with hyperparameter α. The generative story is presented in Algorithm 2 and the plate representation is shown in Fig. 3. When we observe a collection of documents, our job is to estimate the parameter θ_j. To do so, we compute the posterior distribution given the data $P(\theta_j|d_j)$, as

$$\begin{aligned} P(\theta_j|d_j) &\propto P(d_j|\theta_j)P(\theta_j|\alpha) \\ &\propto \prod_{t=1}^{|V|} \theta_{jt}^{N_{jt}+\alpha_t-1} \\ &= Dir(\theta_j|N_{jt}+\alpha_t), \end{aligned} \tag{2}$$

where $P(\theta_j|\alpha)$ is the Dirichlet prior on θ_j. $P(d_j|\theta_j)$ is the likelihood of the data, $|V|$ is the size of the vocabulary, and N_{jt} is the number of times term t occurs in document d_j.

[4] We use the term cross-idiomatic to denote multiple usages in language between different Web ecosystems. However, one may interpret this as if our models actually employ n-grams (or multiple words). They do not. We leave the use of n-grams in this setup for future work

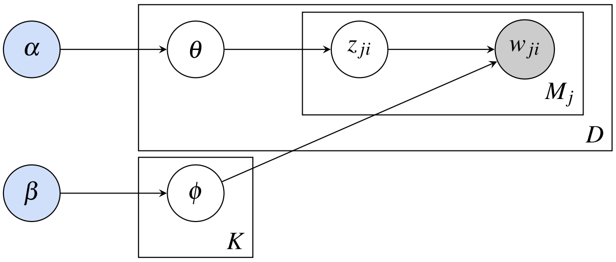


Fig. 4. Plate representation of latent Dirichlet allocation (LDA). The outer large plate denotes D documents, while the inner plate represents the choice of a topic (z_{ji}) for each word (w_{ji}) within a document. The topic distribution of each document is given by θ_j. M_j denotes the number of words in the jth document. The bottom plate represents K word distributions (ϕ) for each topic from a unique language [4].

We estimate θ_j by the expected value of its posterior distribution. Under this model, this is given by

$$E_{Dir}[\theta_{jt}|d_j] = \frac{N_{jt} + \alpha_t}{\sum_{t'=1}^{|V|} N_{jt'} + \alpha_{t'}}. \tag{3}$$

Following the approach in [59], we set the hyperparameter α_t to be proportional to the maximum likelihood estimate of the probability of term t in the entire collection, $\alpha_t = \mu P(w_t|Coll)$, where μ is a constant usually set between (500, 5000).

Eq. (3) allows us to represent each document by the expected value of θ_j. This effectively corresponds to the probability of each word given the document, $P(w_t|d_j)$.

Note that this model fails to capture both the relationship between terms with the same meaning, and the differences between multiple meanings of a term. These are the problems of synonymy and polysemy. It also fails to take into account differences in documents from distinct sources. At learning time, there are no explicit relationships between documents.

5.2. *Latent Dirichlet allocation (LDA)*

LDA models documents as mixtures over latent topics, where each topic is characterized by a distribution over words [4]. One advantage of this representation compared to the unigram model relies on its ability to cope with synonymy and polysemy, since it softly clusters semantically similar co-occurring terms. Additionally, the number of latent topics K is smaller than the original size of the vocabulary $|V|$ and we induce a meaningful yet compact document representation.

Fig. 4 illustrates the plate representation and Algorithm 3 shows the generative story, where K is the number of latent topics in the collection, θ_j is the document-specific topic proportion, $\phi_{z_{ji}}$ is the topic-specific word proportion, z_{ji} is the topic assignment for word w_{ji}, and the subscripts ji denote the ith position in the jth document. α and β are the hyperparameters for the symmetric Dirichlet distributions[5].

Algorithm 3: Generative story for latent Dirichlet Allocation (LDA)

```
Initialize: (1) set the number of topics $K$;
(2) set values for Dirichlet priors $\alpha$ and $\beta$;
sample $K$ times $\phi \sim Dirichlet(\beta)$;
for each document $d_j$ do
    sample $\theta_j \sim Dirichlet(\alpha)$;
    for each word position $i \in d_j$; do
        sample $z_{ji} \sim Multinomial(\theta_j)$;
        sample $w_{ji} \sim Multinomial(\phi_{z_{ji}})$;
    end
end
```

Once again, we model each document by the probability distribution of each vocabulary word w_t, $P(w_t|d_j)$. If we have K topics, we can write this as

$$\begin{aligned} P(w_t|d_j) &= \sum_{k=1}^{K} P(w_t|z_k)P(z_k|d_j) \\ &= \sum_{k=1}^{K} \phi_{tk}\theta_{kj} \quad , \end{aligned} \tag{4}$$

where $P(w_t|z_k) = \phi_{tk}$ and $P(z_k|d_j) = \theta_{kj}$.

[5] In the equations to follow and throughout this article, α and β are set manually. For the sake of clarity, and to avoid overcrowded equations, we may sometimes omit these parameters. However, it should be clear that these parameters are set.

We are interested in estimating the unobserved variables $\mathbf{z}$, ϕ_{tk} and θ_{kj}. The posterior distribution can be written as

$$P(\theta, \phi, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{P(\theta, \phi, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{P(\mathbf{w}|\alpha, \beta)}, \tag{5}$$

where $\mathbf{z}$ is the vector of all topic assignments and $\mathbf{w}$ is the vector of all words.

However, this distribution is intractable, so we use Gibbs sampling as an inference technique for learning the distributions of latent variables (Algorithm 1). Gibbs sampling is a Markov Chain Monte Carlo sampling method. It is easy to implement, has an intuitive interpretation, and is widely used in the literature [19,45]. We infer the topic assignment of a word w_{ji} in document j by the conditional posterior distribution $P(z_{ji} = k|\mathbf{z_{\neg ji}}, \mathbf{w})$, where $\mathbf{z_{\neg ji}}$ denotes all topic assignments in document j, excluding z_{ji}. It has been shown in [21] that

$$P(z_{ji} = k|\mathbf{z_{\neg ji}}, \mathbf{w}) \propto \frac{n_{jk} - 1 + \alpha}{\sum_{k'=1}^{K} n_{jk'} - 1 + K\alpha} \cdot \frac{v_{kt} - 1 + \beta}{\sum_{t'=1}^{|V|} v_{kt'} - 1 + |V|\beta}, \tag{6}$$

where n_{jk} is the number of times topic k was assigned to document j, and v_{kt} is the number of times topic k was assigned to the token t.

Eq. (6) is quite intuitive. The first ratio expresses the expected frequency of topic k in document d_j. The second ratio quantifies the probability that word w_t is sampled from topic k. Consequently, the probability of assigning topic k to a particular word in a particular document is proportional to the number of times that word was already encountered in topic k and how likely topic k is within the document.

We note that both θ_j and ϕ_k can be calculated using just the topic assignments z_{ji}. In other words, $\mathbf{z}$ is a sufficient statistic for these distributions. After the burn-in period of the Gibbs sampling, we can compute them as follows

$$\theta_{jk} = \frac{n_{jk} + \alpha}{\sum_{k'=1}^{K} n_{jk'} + K\alpha} \qquad \phi_{kt} = \frac{v_{kt} + \beta}{\sum_{t'=1}^{|V|} v_{kt'} + |V|\beta}. \tag{7}$$

The value θ_{jk} in Eq. (7) corresponds to the predictive distribution of sampling topic k in document j. The value ϕ_{kt} corresponds to the predictive distribution of sampling a new token t from topic k.

After we learn the word-topic distribution ϕ from a training collection, we can infer the model on unseen documents. The per-document topic distribution θ can be estimated by sampling topic assignments for each word position in every new document. This sampling is very similar to Eq. (6), except that the distribution ϕ is known so the counts v_{kt} are fixed. Iteratively sampling the topics for unseen documents can then be done by

$$P(z_{ji} = k|\mathbf{z_{\neg ji}}, \mathbf{w}) \propto \phi_{kt} \cdot \frac{n_{jk} - 1 + \alpha}{\sum_{k'=1}^{K} n_{jk'} - 1 + K\alpha}. \tag{8}$$

While the LDA model provides a richer representation compared to the unigram model in Section 5.1, it does not allow to explicitly model the different language usages that arise naturally on aligned documents from varied Web sources. It merely concatenates the aligned documents and treats all the words as coming from one unique vocabulary. We turn now to study the bilingual latent Dirichlet allocation (BiLDA), where we have distinct word-topic distributions that may serve to model different idiomatic usages of the same language.

5.3. Bilingual latent Dirichlet allocation (BiLDA)

Unlike LDA, the bilingual latent Dirichlet allocation explicitly takes into account different languages. It considers that two documents, while written in different languages, may refer to the same concepts. An example where this model is a good fit [13,14], consists of pairs of Wikipedia articles written in different languages, where the discussed topics are the same, yet the words are completely different. BiLDA proposes that a pair of aligned documents shares the same topic distribution θ; and there exist two distinct word-topic distributions, ϕ and ψ, which are unique to the two different languages. We denote a pair of different languages as source (S) and target (T). For each pair of aligned documents, topics are sampled from θ. Depending on the language the document is written in, each word is sampled from two distinct distributions, ϕ or ψ. The former corresponds to the S (source) language and the latter to the T (target) language. Fig. 5 shows the plate representation and the generative story is summarized in Algorithm 4.

Similar to LDA, we model each document by the probability distribution of each vocabulary word, except that this time, it is possible to make an explicit distinction between the languages, which allows for a richer representation of aligned documents. Specifically, a document in the source language is modeled by the probability of a source-language word, given the document $P(w_t^S|d_j^S)$. A similar statement is true for a document in the target language. Formally this is written as

$$P(w_t^S|d_j^S) = \sum_{k=1}^{K} \phi_{tk}\theta_{kj} \qquad P(w_t^T|d_j^T) = \sum_{k=1}^{K} \psi_{tk}\theta_{kj}, \tag{9}$$

where $P(w_t^S|z_k^S) = \phi_{tk}$, $P(w_t^T|z_k^T) = \psi_{tk}$, and $P(z_k^S|d_j^S) = P(z_k^T|d_j^T) = \theta_{kj}$. Eq. (9) makes clear the assumption that a pair of aligned documents share the same topic distribution θ, while the word distributions for a given topic are unique for each language.

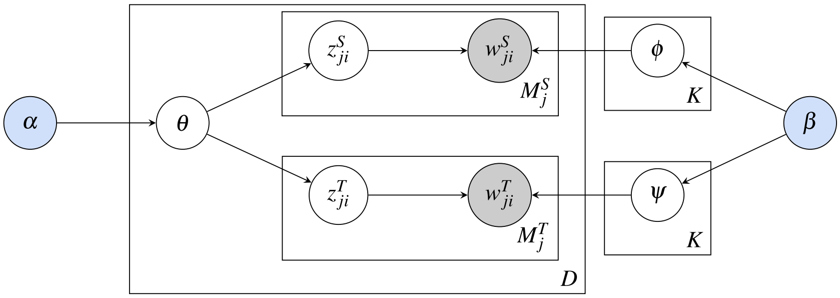


Fig. 5. Plate representation of bilingual latent Dirichlet allocation (BiLDA). The large outer plate (left) denotes D document pairs. Document pairs contain the same topic distribution (θ_j). Inner plates represent the choice of a topic (z_{ji}) for each word (w_{ji}) within source S (top inner plate) and target T (bottom inner plate) documents. M_j denotes the number of words in the *jth* document. The rightmost plates denote the choice of K word distributions for each topic, corresponding to unique words in the source (ϕ) and unique words in the target language (ψ) [14].

Algorithm 4: Generative story for BiLDA

```
**Initialize:** (1) set the number of topics $K$;
(2) set values for Dirichlet priors $\alpha$ and $\beta$;
sample $K$ times $\phi \sim Dirichlet(\beta)$;
sample $K$ times $\psi \sim Dirichlet(\beta)$;
**for** *each document pair* $d_j = \{d_j^S, d_j^T\}$ **do**
    sample $\theta_j \sim Dirichlet(\alpha)$;
    **for** *each word position* $i \in d_j^S$; **do**
        sample $z_{ji}^S \sim Multinomial(\theta_j)$;
        sample $w_{ji}^S \sim Multinomial(\phi_{z_{ji}^S})$;
    **end**
    **for** *each word position* $i \in d_j^T$; **do**
        sample $z_{ji}^T \sim Multinomial(\theta_j)$;
        sample $w_{ji}^T \sim Multinomial(\psi_{z_{ji}^T})$;
    **end**
**end**
```

We are interested in estimating the unobserved variables θ, ϕ, ψ, $\mathbf{z}^S$ and $\mathbf{z^T}$. The posterior distribution for the variables of interest can be written as

$$P(\theta, \phi, \psi, \mathbf{z}^S, \mathbf{z}^T | \mathbf{w}^S, \mathbf{w}^T \alpha, \beta) = \frac{P(\theta, \phi, \psi, \mathbf{z}^S, \mathbf{z}^T, \mathbf{w}^S, \mathbf{w}^T | \alpha, \beta)}{P(\mathbf{w}^S, \mathbf{w}^T | \alpha, \beta)}. \tag{10}$$

This distribution is intractable and once again we use Gibbs sampling as an inference technique. For each language, source (S) and target (T), we compute the probability that the current (document j, position i) topic assignment is k, given all the other topic assignments and words. For the source language this corresponds to

$$P(z_{ji}^S = k | \mathbf{z}_{\neg ji}^S, \mathbf{z}^T, \mathbf{w}^S, \mathbf{w}^T). \tag{11}$$

An analogous expression can be written for the topic assignments in the target language. The Gibbs sampling expressions used for training the model (Algorithm 1) for BiLDA [14] are

$$P(z_{ji}^S = k | \mathbf{z}_{\neg ji}^S, \mathbf{z}^T, \mathbf{w}^S, \mathbf{w}^T) \propto \frac{{n_{jk}}^S - 1 + n_{jk}^T + \alpha}{\sum_{k'=1}^K {n_{jk'}}^S - 1 + n_{jk'}^T + K\alpha} \cdot \frac{{\upsilon_{kt}}^S - 1 + \beta}{\sum_{t'=1}^{|V^S|} {\upsilon_{kt'}}^S - 1 + |V^S|\beta}. \tag{12}$$

$$P(z_{ji}^T = k | \mathbf{z}_{\neg ji}^T, \mathbf{z}^S, \mathbf{w}^S, \mathbf{w}^T) \propto \frac{{n_{jk}}^T - 1 + n_{jk}^S + \alpha}{\sum_{k'=1}^K n_{jk'}^T - 1 + {n_{jk'}}^S + K\alpha} \cdot \frac{{\upsilon_{kt}}^T - 1 + \beta}{\sum_{t'=1}^{|V^T|} {\upsilon_{kt'}}^T - 1 + |V^T|\beta}. \tag{13}$$

Similarly to the notation for LDA (Section 5.2), n_{jk} is the number of times topic k was assigned to document j, and υ_{kt} is the number of times topic k was assigned to the token t. We use the superscripts S and T to differentiate between the two languages. $|V^S|$ and $|V^T|$ represent the vocabulary size of each language.

After the burn-in period of the Gibbs sampling, the topic proportion for the jth document pair is given by

$$\theta_{jk} = \frac{n_{jk}^T + n_{jk}^S + \alpha}{\sum_{k'=1}^K n_{jk'}^T + n_{jk'}^S + K\alpha}. \tag{14}$$

The word distributions for the source and target language are respectively

$$\phi_{kt} = \frac{v^S_{kt} + \beta}{\sum_{t'=1}^{|V^S|} v^S_{kt'} + |V^S|\beta} \qquad \psi_{kt} = \frac{v^T_{kt} + \beta}{\sum_{t'=1}^{|V^T|} v^T_{kt'} + |V^T|\beta}. \tag{15}$$

We may infer the per-document topic proportions for new unseen documents by iteratively sampling topic assignments for each word position. This is done one language at a time and the sampling equation is completely analogous to Eq. (8), except that we differentiate between each unique word-topic distribution for each language ϕ and ψ. In other words, if we wish to infer the topic proportions on an unseen document written in the source language, we follow

$$P(z_{ji} = k|\mathbf{z}_{\neg\mathbf{ji}}, \mathbf{w}) \propto \phi_{kt} \cdot \frac{n_{jk} - 1 + \alpha}{\sum_{k'=1}^{K} n_{jk'} - 1 + K\alpha}, \tag{16}$$

where ϕ_{kt} is fixed. Whereas if the unseen document is written in the target language, the sampling equation is analogous to Eq. (16), except that we use ψ_{kt}, instead of ϕ_{kt}.

From Eq. (14) we can see that document pairs are linked by the count variables $n^S_{j\cdot}$ and $n^T_{j\cdot}$, since both z^S_{ji} and z^T_{ji} are drawn from the same distribution θ. Additionally, Eq. (15) explicitly shows that words from different languages are sampled from distinct word-topic distributions. This is because BiLDA assumes that a pair of aligned documents are written in completely distinct languages. For our task, this assumption does not hold (i.e., a portion of the vocabulary is shared between pairs of aligned documents from different Web sources). Thus, we will study how to relax this constraint in the next section where we propose a new model capable to cope with not only the differences in language but also the similarities.

6. The multi-idiomatic topic model (MiLDA)

In Section 5.2 we saw that the LDA model uses a compact representation of a document by clustering words that are semantically similar. This presents an advantage over the unigram model, where no semantic clustering takes place.

We also saw that BiLDA assumes that each language has its own distinct set of words. However, this assumption does not hold here. While we deal with different idioms on each document pair, there is a large portion of shared words. We will model this explicitly.

Given the setup, we propose a new model, *multi-idiomatic LDA* (MiLDA). It 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 described in Section 5.3; (2) A portion of words in the colloquial idiom differs from those in the formal idiom and vice versa. Meanwhile, a portion of words is shared between the two language idioms, and each document may be observed as a set of shared and non-shared words combined together. This is conceptually different from BiLDA, 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 over shared words).

Fig. 6 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.

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. Algorithm 5 shows the full generative story of the MiLDA model.

We represent each document by the probability distribution of each vocabulary word. For a source document, a vocabulary word may come from the unique source language distribution or from the shared one. This is analogous for a target document. Formally this is written as[6]

$$P(w_t|d^S_j, s_t) = (1 - s_t)\sum_{k=1}^{K} \phi_{tk}\theta_{kj} + s_t \sum_{k=1}^{K} \chi_{tk}\theta_{kj} \tag{17}$$

$$P(w_t|d^T_j, s_t) = (1 - s_t)\sum_{k=1}^{K} \psi_{tk}\theta_{kj} + s_t \sum_{k=1}^{K} \chi_{tk}\theta_{kj} \tag{18}$$

[6] A note on notation: we use lowercase s to denote the indicator variable and uppercase S as a superscript to differentiate the source language

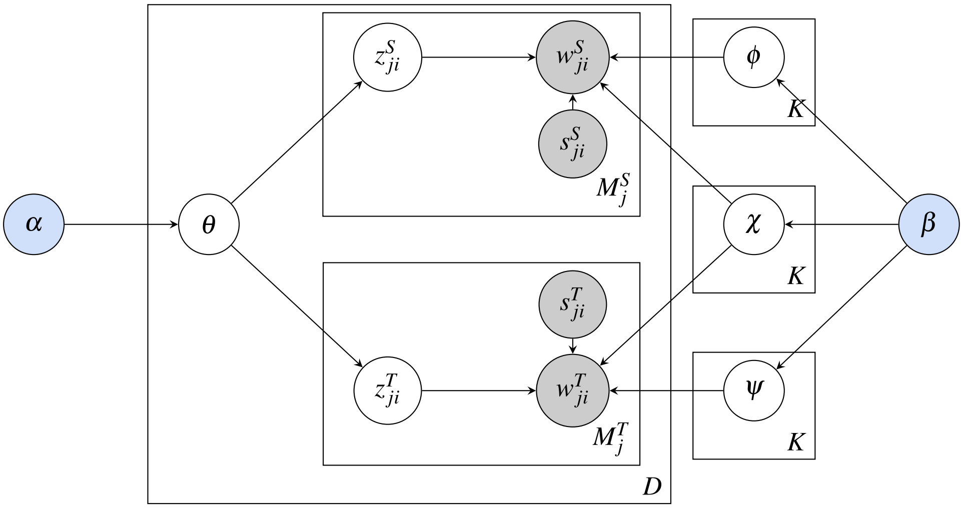


Fig. 6. Plate representation of the multi-idiomatic LDA (MiLDA) model. The large outer plate (left) denotes *D* document pairs. Document pairs contain the same topic distribution (θ_j). Inner plates represent the choice of a topic (z_{ji}) for each word (w_{ji}), and the sharedness indicator (s_{ji}) within source *S* (top inner plate) and target *T* (bottom inner plate) documents. M_j denotes the number of words in the *jth* document. The rightmost plates denote the choice of *K* word distributions for each topic, corresponding to the source (ϕ), target (ψ) and shared (χ) language.

Algorithm 5: Generative story for MiLDA

```
Initialize: (1) set the number of topics $K$;
(2) set values for Dirichlet priors $\alpha$ and $\beta$;
(3) set values for $s_{ij}^S$ and $s_{ij}^T$.
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 $d_j = \{d_j^S, d_j^T\}$ do
    sample $\theta_j \sim Dirichlet(\alpha)$;
    for each word position $i \in d_j^S$; do
        sample $z_{ji}^S \sim Multinomial(\theta_j)$;
        if $s_{ji}^S = 1$; then
            sample $w_{ji}^S \sim Multinomial(\chi_{z_{ji}^S})$;
        else
            sample $w_{ji}^S \sim Multinomial(\phi_{z_{ji}^S})$;
        end
    end
    for each word position $i \in d_j^T$; do
        sample $z_{ji}^T \sim Multinomial(\theta_j)$;
        if $s_{ji}^T = 1$; then
            sample $w_{ji}^T \sim Multinomial(\chi_{z_{ji}^T})$;
        else
            sample $w_{ji}^T \sim Multinomial(\psi_{z_{ji}^T})$;
        end
    end
end
```

When the word is unique to one of the languages ($s_t = 0$), the probability of a term in a document is exclusively governed by the unique word-topic distribution (ϕ for the source document or ψ for a target document). When the word is shared ($s_t = 1$), the probability of a term in a document is governed by the shared word-topic distribution (χ). We see how this representation allows us to cross-link pairs of documents, not only through the shared topic distribution, but also by selectively drawing words from a shared word-topic distribution. The topic distribution θ is the same for both documents, as stated in our assumptions.

This document representation allows us to exploit both the differences and similarities between two idiomatic usages of the same language.

6.1. Learning the multi-idiomatic topic model from data

We want to estimate the unobserved variables θ, ϕ, ψ, χ, $\mathbf{z}^S$ and $\mathbf{z^T}$. The posterior distribution for these variables of interest can be written as

$$P(\theta, \phi, \psi, \chi, \mathbf{z}|\mathbf{w}, \mathbf{s}, \alpha, \beta) = \frac{P(\theta, \phi, \psi, \chi, \mathbf{z}, \mathbf{w}, \mathbf{s}|\alpha, \beta)}{P(\mathbf{w}, \mathbf{s}|\alpha, \beta)}. \tag{19}$$

To simplify the notation, in this section we group source and target variables to represent $\mathbf{z} = \{\mathbf{z}^S, \mathbf{z}^T\}$, $\mathbf{w} = \{\mathbf{w}^S, \mathbf{w}^T\}$ and $\mathbf{s} = \{\mathbf{s}^S, \mathbf{s}^T\}$. We will make explicit distinctions when necessary.

The distribution in Eq. (19) is intractable and once again we use Gibbs sampling as an inference technique. For each document, we want to compute the probability that the current topic assignment is k, given all the other topic assignments, words and indicators. For a document in the source language this corresponds to

$$P(z^S_{ji} = k|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^S, \mathbf{w}^T, \mathbf{s}^S, \mathbf{s}^T). \tag{20}$$

An analogous expression can be written for a target document.

In the following paragraphs we formally derive the equations to perform Gibbs sampling for MiLDA, which result in Eqs. (29) and (30).

Applying Bayes' rule to Eq. (20),

$$\begin{aligned} &P(z^S_{ji} = k|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^S, \mathbf{w}^T, \mathbf{s}^S, \mathbf{s}^T) \\ &\quad \propto \quad P(w^S_{ji}|z^S_{ji} = k, \mathbf{z}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^S_{\neg ji}, \mathbf{w}^T, s^S_{ji})P(z^S_{ji} = k|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T). \end{aligned} \tag{21}$$

We integrate over the parameter values that arise in each of the terms on the right hand side of the Eq. (21). The first term corresponds to the posterior predictive distribution of the word w^S_{ji} in the source document. It can be decomposed in two cases, depending on the indicator s^S_{ji}. When $s^S_{ji} = 0$,

$$P(w^S_{ji}|z^S_{ji} = k, s^S_{ji} = 0, \mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji}) = \int_{\phi_k} P(w^S_{ji}|z^S_{ji} = k, \phi_k)P(\phi_k|\mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji})d\phi_k \tag{22}$$

From Bayes' rule we can obtain the rightmost term

$$P(\phi_k|\mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji}) \propto P(\mathbf{w}^S_{\neg ji}|\mathbf{z}^S_{\neg ji}, \phi_k)P(\phi_k) \tag{23}$$

Since $P(\phi_k)$ is distributed as $Dirichlet(\beta)$ and is conjugate to the multinomial $P(w^S_{ji}|z^S_{ji} = k, \phi_k)$, the posterior distribution $P(\phi_k|\mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji})$ is $Dirichlet(v^{-iS}_{kt} + \beta)$, where v^{-iS}_{kt} is the number of times that term t (unique to the source language) was assigned to topic k, not including the current word. Additionally, since the first term in the right-hand side of Eq. (22) evaluates to ϕ_k, the integral evaluates to the expected value of the Dirichlet distribution

$$\begin{aligned} P(w^S_{ji}|z^S_{ji} = k, s^S_{ji} = 0, \mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji}) &= \int_{\phi_k} \phi_k Dir(\phi_k|v^{-iS}_{kt} + \beta_t)d\phi_k \\ &= E[\phi_k|v^{-iS}_{kt} + \beta_t] \\ &= \frac{v^{-iS}_{kt} + \beta_t}{\sum_{t'=1}^{|V^S|} v^{-i\,S}_{kt'} + \beta_{t'}} \end{aligned} \tag{24}$$

A similar derivation can be obtained when the word is shared ($s^S_{ji} = 1$), except that in this case we need to keep track of the topic assignments on both source and target documents,

$$\begin{aligned} &P(w^S_{ji}|z^S_{ji} = k, s^S_{ji} = 1, \mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^T) \\ &\quad = \int_{\chi_k} P(w^S_{ji}|z^S_{ji} = k, \chi_k)P(\chi_k|\mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^T)d\chi_k. \end{aligned} \tag{25}$$

The rightmost term in Eq. (25) can be computed with Bayes' rule,

$$P(\chi_k|\mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^T) \propto P(\mathbf{w}^S_{\neg ji}|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^T, \chi_k)P(\chi_k). \tag{26}$$

It corresponds to the posterior distribution of the parameter χ_k, given a Dirichlet prior $P(\chi_k)$, and a multinomial likelihood $P(\mathbf{w}^S_{\neg ji}|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^T, \chi_k)$. Thus, the posterior distribution is $Dir(\phi_k|v^{-iC}_{kt} + \beta_t)$, where v^{-iC}_{kt} is the number of times that a shared term t (shared between the two languages) was assigned to topic k, not including the current word. We use the superscript C to differentiate the shared (common) words from the non-shared ones in the source and target languages. $|V^C|$ denotes the size of the shared vocabulary. Finally, we can evaluate the integral in Eq. (25) by the expected value of the posterior Dirichlet distribution on the parameter χ_k,

$$\begin{aligned} P(w^S_{ji}|z^S_{ji} = k, s^S_{ji} = 1, \mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^T) &= \int_{\chi_k} \chi_k Dir(\chi_k|v^{-iC}_{kt} + \beta_t) d\chi_k \\ &= E[\chi_k|v^{-iC}_{kt} + \beta_t] \\ &= \frac{v^{-iC}_{kt} + \beta_t}{\sum_{t'=1}^{|V^C|} v^{-i\,C}_{kt'} + \beta_{t'}}. \end{aligned} \tag{27}$$

We turn now our attention to the rightmost term in Eq. (21). $P(z^S_{ji} = k|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T)$ is the posterior predictive distribution for the topic assignment z_{ji} given all other topic assignments. Marginalizing over the parameter θ_j, we obtain

$$\begin{aligned} P(z^S_{ji} = k|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T) &= \int_{\theta_j} P(z^S_{ji} = k|\theta_j) P(\theta_j|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T) d\theta_j \\ &= \int_{\theta_j} \theta_j Dir(\theta_j|n^{-iS}_{jk} + n^T_{jk} + \alpha_k) d\theta_j \\ &= E_{Dir}[\theta_j|n^{-iS}_{jk} + n^T_{jk} + \alpha_k] \\ &= \frac{n^{-iS}_{jk} + n^T_{jk} + \alpha_k}{\sum_{k'=1}^{K} n^{-i\,S}_{jk'} + n^T_{jk'} + \alpha_{k'}}. \end{aligned} \tag{28}$$

$P(z^S_{ji} = k|\theta_j)$ evaluates to θ_j. $P(\theta_j|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T)$ is the posterior distribution of the parameter θ_j, with a Dirichlet prior $P(\theta_j)$, and a multinomial likelihood $P(\mathbf{z}^S_{\neg ji}, \mathbf{z}^T|\theta_j)$. So $P(\theta_j|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T) = Dir(\theta_j|n^{-iS}_{jk} + n^T_{jk} + \alpha_k)$. Eq. (28) evaluates to the expected value of this Dirichlet distribution.

We put together the results from Eqs. (24), (27) and (28), and obtain an expression for the Gibbs sampler used for training the model (Algorihtm 1),

$$\begin{aligned} &P(z^S_{ji} = k|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T, \mathbf{w}^S, \mathbf{w}^T, \mathbf{s}^S, \mathbf{s}^T) \\ &\quad\propto \frac{n^{-iS}_{jk} + n^T_{jk} + \alpha_k}{\sum_{k'=1}^{K} n^{-i\,S}_{jk'} + n^T_{jk'} + \alpha_{k'}} \left[(1 - s^S_{ji}) \frac{v^{-iS}_{kt} + \beta_t}{\sum_{'t=1}^{|V^S|} v^{-i\,S}_{kt'} + \beta_{t'}} + s^S_{ji} \frac{v^{-iC}_{kt} + \beta_t}{\sum_{t'=1}^{|V^C|} v^{-i\,C}_{kt'} + \beta_{t'}} \right]. \end{aligned} \tag{29}$$

In Eq. (29), the first ratio expresses the probability of topic k in the document pair d_j. The last two ratios express the probability that the word is generated by topic k, considering two different vocabularies (source S and shared C). If the word is not shared ($s^S_{ji} = 0$), we only consider the source word-topic distribution, whereas if the word is shared ($s^S_{ji} = 1$), we consider the shared word-topic distribution.

Similar equations can be derived for the target documents,

$$\begin{aligned} &P(z^T_{ji} = k|\mathbf{z}^T_{\neg ji}, \mathbf{z}^S, \mathbf{w}^S, \mathbf{w}^T, \mathbf{s}^S, \mathbf{s}^T) \\ &\quad\propto \frac{n^{-iT}_{jk} + n^S_{jk} + \alpha_k}{\sum_{k'=1}^{K} n^{-i\,T}_{jk'} + n^S_{jk'} + \alpha_{k'}} \left[(1 - s^T_{ji}) \frac{v^{-iT}_{kt} + \beta_t}{\sum_{'t=1}^{|V^T|} v^{-i\,T}_{kt'} + \beta_{t'}} + s^T_{ji} \frac{v^{-iC}_{kt} + \beta_t}{\sum_{t'=1}^{|V^C|} v^{-i\,C}_{kt'} + \beta_{t'}} \right]. \end{aligned} \tag{30}$$

Note from Eqs. (29) and (30) that the counts from shared words (v^C_{kt}) influence the topic assignments on both source and target documents. This shared influence allows us to effectively model the similarities in the language; while the counts of non-shared words model the differences.

After the burn-in period of Gibbs sampling, we obtain estimates of the underlying distributions. The estimated topic proportions are given by

$$\theta_{jk} = \frac{n^T_{jk} + n^S_{jk} + \alpha_k}{\sum_{k'=1}^{K} n^T_{jk'} + n^S_{jk'} + \alpha_{k'}}. \tag{31}$$

While the word-topic distributions for source ϕ, target ψ and shared χ vocabularies are given respectively by

$$\phi_{kt} = \frac{v^S_{kt} + \beta_t}{\sum_{t'=1}^{|V^S|} v^S_{kt'} + \beta'_t} \qquad \psi_{kt} = \frac{v^T_{kt} + \beta_t}{\sum_{t'=1}^{|V^T|} v^T_{kt'} + \beta_{t'}} \qquad \chi_{kt} = \frac{v^C_{kt} + \beta_t}{\sum_{t'=1}^{|V^C|} v^C_{kt'} + \beta_{t'}}. \tag{32}$$

Unlike BiLDA, the source and target vocabularies, V^S and V^T, respectively, contain only source or language words, which are non-shared. All the shared words lie in V^C. In our experimental sections, we set the same hyperparameter β_t for all vocabulary distributions and we leave to future work exploring asymmetric Dirichlet priors.

Eqs. (31) and (32) are substituted in Eqs. (17) and (18) and we obtain a cross-idiom representation of pairs of aligned documents. Under this representation, we shall see in Section 10.1, that observing the most likely words in each word-topic distribution allows us to discover new word associations that other models fail to capture.

6.2. Inferring the model on unseen documents

Inferring the per-document topic proportions for unseen documents for this model is analogous to the procedure presented in Sections 5.2 and 5.3. For every new document, we iteratively sample topic assignments having the word-topic distributions ϕ, ψ, χ fixed. For a source document, this is given by

$$P(z_{ji}=k|\mathbf{z_{\neg ji}},\mathbf{w},s_{ji}) \propto \left[(1-s_{ji})\phi_{kt}+s_{ji}\chi_{kt}\right]\cdot\frac{n_{jk}^{-i}+\alpha}{\sum_{k'=1}^{K}n_{jk'}^{-i}+K\alpha}. \tag{33}$$

For a target document, we use the same equation except that we substitute ϕ_{kt} by ψ_{kt}.

6.3. Remarks on the modelling approach

MiLDA aims to model both the similarities and differences between two modes of expression (e.g., formal vs. informal). To address this, it considers three sets of word-topic distributions. From Eq. (32), ϕ contains terms that are unique to an informal ecosystem, ψ contains terms unique to a formal ecosystem and χ contains all shared terms. The shared terms aim to model the similarities in the vocabulary, while the unique ones aim to model the differences. In contrast, the LDA model considers that all words are shared, i.e., there is only one common vocabulary and it disregards that some terms appear only in one of the settings (formal or informal). While the BiLDA model considers that all words are non-shared, i.e., there are two distinct vocabularies and it disregards that some terms are shared between the two ecosystems. The MiLDA provides a sort of middle ground between these two models, where a portion of the vocabulary is shared and a portion is not shared.

We have chosen to partition the vocabulary in a simple way, i.e., by considering a fully observed binary indicator variable that reveals whether or not the term is shared or unique from the training collection. This could be taken into account in other ways. For example, we may utilize other heuristics to partition the vocabulary. For example, one may impose a minimum threshold for the number of times that a term should appear in both modes of expressions to be considered as shared. Another technique would be to treat the sharedness indicator as a continuous latent variable, and to learn its posterior distribution based on the data. This way, words may be softly assigned to each of the sets, allowing for greater flexibility of the model.

In this article we do not aim to explore all different ways in which one may partition the vocabulary. Our contribution is to propose and demonstrate that partitioning the vocabulary is beneficial when addressing multiple expressions of the same language. We leave the investigation of such approaches for future work.

7. Retrieval model

As mentioned in Section 4, we define the task of linking Web sources of consumers and sellers as an ad-hoc information retrieval task. Given a query from a particular source, we rank the documents from a target collection according to the relevance to the query. In the example of linking pins to Amazon's webshops, the pins represent the queries which come from the Pinterest (source) ecosystem; whereas the webshops represent the documents which live in the Amazon (target) ecosystem.

As a retrieval model, we follow the state-of-the-art approach in [51]. Thus, we combine the probabilistic representation from the unigram model with a probabilistic topic representation. We adopt a linear combination of the two models, as follows

$$P_{uni+tr}(q_i|d_j)=\lambda P_{uni}(q_i|d_j)+(1-\lambda)P_{tr}(q_i|d_j), \tag{34}$$

where P_{uni} is the document model corresponding to the simple unigram with Dirichlet prior representation given by Eq. (3). P_{tr} is the document representation given by a representation model and it evaluates different expressions depending on the model considered: LDA (Eq. (4)), BiLDA (Eq. (9)) and MiLDA (Eqs. (17) and (18)).

The interpolation parameter λ weighs the importance of each term: $\lambda = 0$ reduces the model to the simple unigram model, while $\lambda = 1$ assumes a full topical document representation. In our experimental results, we shall see the importance of combining these two representations. We also study the influence of the parameter λ and the final linking quality for different numbers of topics K.

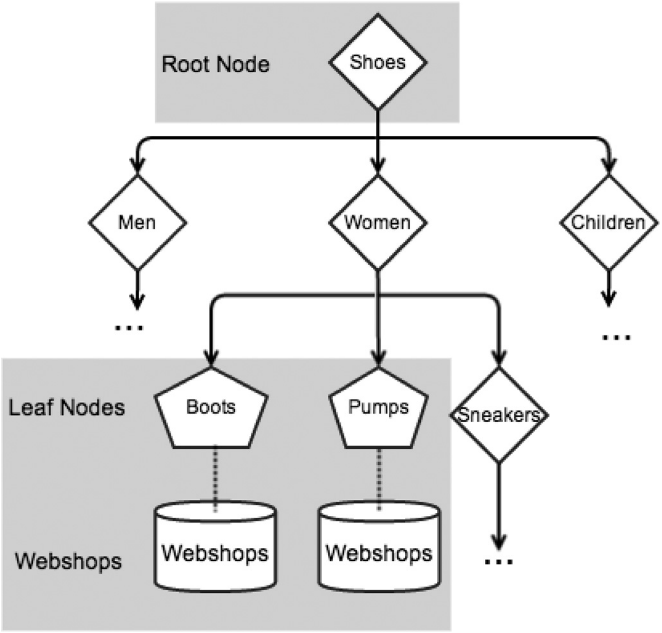


Fig. 7. Example of Amazon's hierarchical product classification structure. Products in leaf nodes are randomly grouped into webshops. In our dataset, each webshop contains approximately 20 products.

8. Datasets

We wish to assess how well each model is able to represent the cross-idiomatic nature of different sources on the Web. We searched for benchmark datasets that could provide aligned documents, i.e., documents that discuss similar topics using different expressions of language depending on context, complexity, formality, etc. While the Web is full of such documents, to our knowledge, there is not a benchmark dataset that we could use. Therefore, we collected three datasets from different Web sources: social media posts from Pinterest.com, product descriptions from Amazon.com and product reviews from Amazon.com. Items described by Pinterest users may be linked to Amazon products. We use Amazon product reviews aligned with the product descriptions to learn how different users may freely -and often informally- refer to the same products. Here we describe each dataset[7].

Dataset I: Social media - Pinterest.com Pinterest.com is a social media site that allows users to *pin* (or post) images on virtual boards. For each pin, a user often writes a description and expresses her opinion about the item. A board is a collection of pins often related to a given category. Board categories include *fashion, travel, cars, food, film, humor, home design, sports, and art*. Pins often present items or activities users are interested in. Examples of two pins are shown on the right column of Fig. 2.

We implemented a crawler to find Pinterest users, their boards and pins. A user page contains a set of boards. Our crawler performed a depth-first search starting from a popular (many followers) Pinterest user. Our initial dataset consists of over one million pins, corresponding to over 18,000 boards and 650 users. The number of pins in a board varies from a couple to several thousands. The average number of pins per user is 2476, while the average number of pins per board is 55.6.

Dataset II: Product descriptions - Amazon.com Amazon contains a large collection of products. It organizes its items in a hierarchy of browse nodes (or categories). Each node is a collection of related items, i.e., products that belong to the same category. We focused on a set of top categories: *Apparel, Beauty, Books, Electronics, Groceries, Jewelry, Kitchen, Music, Shoes, Sporting Goods* and *Watches*. Each of these contain subcategories. For example, the category *Shoes* is further categorized into *Men's*, *Women's* and *Children's*. These can be further categorized into *Boots*, *Pumps*, *Sneakers*, and so on. A schematic representation is shown in Fig. 7.

We implemented an XML parser to traverse Amazon's node hierarchy. We started by querying the top categories and gathered information from related child nodes. We downloaded the textual description and images from 19, 955 products, corresponding to 235 leaf categories.

Dataset III: Product reviews - Amazon.com We crawled Amazon's reviews and obtained the top 10 most helpful reviews for a subset of products in Dataset II. The helpfulness of each review is indicated by Amazon users who rate the reviews

[7] All datasets will be available online upon acceptance of this manuscript.

on the site. We concatenated the reviews corresponding to the same product in one document and aligned it to its corresponding product description. Not all products have reviews, so we only considered the ones that did. This comprises a set of 15,566, which is a subset of the 19, 955 products in Dataset II.

9. Experiments

Our experiments demonstrate the task of linking cross-idiomatic Web sources, as described in Section 4. We frame this as a retrieval task in which, given an item from one Web source, we obtain a list of relevant items from a different Web source, where the idiomatic usage of the language differs. One realisation of this consists of linking users' pins posted on a social media site (e.g., Pinterest.com) to a set of relevant products or webshops (e.g., Amazon.com). The idea is that when a user expresses interest in an item through social media, we can automatically identify the item and link it to further information. Specifically, we use pins from Pinterest.com as queries and collections of products from Amazon.com as target documents.

This section describes the training and target collections, queries, ground truth and the model parameters used for our experiments.

Our overall experimental approach is as follows: We first use the training collection to learn the models. Then we infer the models in the target collection and perform the linking task between social media items and online shops. Finally, we evaluate what we have learned in the representations and how well we are able to find related items between the different Web environments. Our experiments characterize the behaviour of the models for the proposed task. In particular, we explore what happens when we vary the number of topics and the interpolation parameter λ in Eq. (34). These parameters can be set using cross-validation.

9.1. Training collection

The training collection is the set of documents from which we learn the word distributions for each topic. Since we are interested in learning multi-idiomatic usages of the same language, we use the natural alignment between Dataset II (product descriptions) and Dataset III (product reviews) to train our models. These collections align documents written in the seller and consumer languages respectively. On the description side, products are described in a rather formal way. On the reviews side, the same products may be described using colloquialisms and subtle references. So our training collection consists of 15, 566 aligned document pairs, where each pair consists of a product description and a collection of the top 10 most helpful users' reviews for that particular product (products without reviews are not considered for training). This setup is different from the one described in [61] since they use the full target collection for training and do not consider users' reviews.

The vocabulary size (number of unique words) of the full training collection is 102,077. Out of these, 8% (8626) are found only in the product descriptions and not in reviews; 65% (65,898) are found only in the reviews (not in products); and 27% (27,553) are shared between the two. We see that users tend to employ many words already mentioned in the descriptions, but they also add a large set of new words not previously used to describe the items.

As a comparison, Dataset I contains 164,590 unique words. Out of these, 15% (25,248) are shared with the product descriptions in the training collection. Whereas 29.3% (48,180) are shared with the reviews. While the Amazon reviews may not fully span over the full vocabulary in Pinterest, they have a larger overlap than the product descriptions alone. This suggests that the reviews have more in common with the language in social media than the product descriptions from Amazon.

9.2. Target collection

We consider two setups to define the target collection: Webshop and Product Setups. In the *Webshop Setup*, we grouped similar products, from the full collection (approx. 20,000 products) in Dataset II, into what we call a *webshop document*. A webshop document (or target document) is a collection of products that belong to the same subcategory, according to Amazon's classification scheme. We take all the products that belong to the same leaf node (or leaf subcategory) and randomly separate them into subgroups of at most 20 products each. Fig. 7 illustrates this. We create at most five webshop documents for each subcategory. Thus our target collection in the first setup consists of 1171 documents (or webshops). The idea here is to simulate an online retail business that has a set of webshops containing related products. The number of webshops allows us to annotate every single document as relevant or not for every single query and for each annotator. For example, for 100 queries, each annotator performs over 100,000 relevance annotations. These annotations allow us to compute precision and recall at every position in the ranked sequence of retrieved documents, which in turn allows us to compute mean average precision.

The *Product Setup* simply consists of the full collection of individual products (without grouping). This setup considers almost 20,000 documents. The relevance of the top 5 retrieved documents is indicated from judges on a crowd-sourcing site. We present the full details of the annotation process in Section 9.3

The vocabulary size of the target collection is 62,591 and approximately 50% of these are shared with the training collection.

Table 1
Query length (number of words) statistics.

Minimum	Maximum	Mode	Average
1	52	4	6.86

Table 2
Query examples. Each query is manually annotated with relevant Amazon's categories. Each category contains approx. 100 products divided into 5 webshops (approx. 20 products each).

Sample Pins (used as queries)	Relevant Amazon Category
Be daring, go all out in red! Modern Jessica Rabbit	Apparel/Women/Dresses
Pandora New Design Fashion Lively Ladies Bracelet	Jewelry/Bracelets
Best "going home" outfit	Apparel/Baby
Sled riding!	Sporting Goods/Snow Sports
David Bromstad Kitchen	Kitchen/Furniture/ Kitchen Furniture
blue suede shoes	Shoes/Women/Flats
Hue Layered Net Tights	Apparel/Women/ Leggings
Mens Covington Cargo Shorts size 34 NWT	Apparel/Men/Shorts
Dark on the bottom	Beauty/Makeup/Eyes
Rebecca Minkoff 'ILY' Leather Tote	Shoes/Handbags
Fashion, Make up, Mouth, Red	Beauty/Makeup/Lips
TIFFANY & CO. Diamond Platinum Pink Spinel 'Blue Book' Ring	Jewelry/Rings
Paint first coat then before second coat sets press lines with a ruler diagonally quilted nails	Beauty/Makeup/ Nails/Nail Art
Baby Hat Brown Wig Hat Winter Cap Christmas Gift Ideas by YumBaby, $29.95	Apparel/Baby/
Chair Pose From Three Minute Egg Yoga Pose Weekly	Apparel/Women/Active
Luna Sofa	Kitchen/Furniture/ Living Room Furniture
high-heels-2	Shoes/Women/Pumps
Lace-up Fur Ankle Boots High Heels	Shoes/Women/Pumps
Stripe nail with blue points	Beauty/Makeup/Nails/ Nail Art
Tips To Stay Fit and Healthy	Books/Health, Fitness & Dieting
A Ten Step Guide to Nailing Office Style	Apparel/Men/Suits
beautiful white wedding dress & wedding bouquet - pink rose & little white flowers	Apparel/Women/Dresses/Wedding Dresses
Gifts Under 50: Coach Stone Stud Earrings	Jewelry/Earrings
Stack it up	Watches/Women
browning	Beauty/Makeup/Nails/Nail Art

Although in this work we use Amazon data to simulate a target set of webshops to be retrieved , the proposed linking framework may be extended to any other Web sources for which a textual representation is provided.

9.3. Queries and annotations

We used the text from 100 pins (randomly selected from our collection of one million pins) as queries to link to webshops. In the previous preliminary study [48], a subset of 50 pins were used. Table 2 shows examples of 25 queries. We see that the language usage may differ significantly from that of the product descriptions.

For example, on the first query, "Be daring, go all out in red! Modern Jessica Rabbit", the user encourages the reader to wear a sexy red dress like the one worn by the famous cartoon character Jessica Rabbit. We know this by looking at the image. Another example, "Best going home outfit", refers to the outfit that a baby would wear as it goes home for the first time after it is born. The pin "Dark on the bottom" refers to a dark eye shadow that can be used on the bottom of the eye. These are all examples of how people may refer to certain objects but the words to describe them are full of subtle, implicit references and informal language. Table 1 shows basic statistics regarding the length (number of words) of the query set.

For the Webshop Setup, two annotators manually evaluated each query with the relevant Amazon webshops. The annotators were presented both the text and image of the pin. Additionally, the annotators were presented with all the webshops, containing both the textual and visual information from all the products[8]. For each query, the annotator indicated the path to all the relevant webshops. Table 2 shows these annotations for the query examples. For each Amazon category, expressed as a path on the table, there are at most five webshop. We assume that a human is able to provide correct links between the pin and the relevant webshop documents. To measure the inter-annotator agreement, we computed the Cohen's Kappa coefficient and obtained 0.732. In total, both annotators completed over 200,000 relevance annotations.

For the Product Setup, we first performed the retrieval task for each model condition and kept the top 5 products for each query. We then uploaded the queries and retrieved top products to the Crowdflower platform[9]. This is a crowdsourcing

[8] We show the pin and product images to the annotator to facilitate identifying the relevant webshops since the text alone might not be suffciently descriptive. However, let us emphasize that we only use the text in the linking task. Using the image as part of the query is left for future work.

[9] www.crowdflower.com

platform similar to Amazon's Mechanical Turk. The relevance of each pin-product pair for each model condition is evaluated by at least 3 judges in a scale from 0 to 3, where 0 denotes irrelevant, 1 denotes relevant with many differences, 2 denotes relevant with a few differences and 3 denotes a perfect match. Based on the judges input, Crowdflower returned the relevance value with the highest confidence for each pin-product pair. It also reported an average inter-annotator agreement of 0.647. We considered scores between 2 and 3 as relevant, whereas 0 and 1 were considered irrelevant.

9.4. Model parameters

We use all the models described in Sections 5 and 6. All topic models have been trained with the same number of topics (K=100, 200, 500, 800, 1000, 1200 and 1500) with the same number of iterations (1 000) 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 [45,51]. All the topic models reached the Gibbs sampling burn-in period and successfully converged. 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. We use a linear interpolation between the topic representation and the unigram model, as in Eq. (34). We test $\lambda \in [0, 1]$ with step size 0.1. When $\lambda = 0$, we only have the influence of the topic representation; whereas $\lambda = 1$ corresponds to a full unigram representation with no contribution of the topics.

9.5. Evaluation

As it is common practice, we qualitatively study the output of the different topic models by evaluating the top (most probable) words for a set of topics. Additionally, with regard to the retrieval task, we compute the Mean Average Precision (MAP) for the set $Q = \{Q_1, Q_2, \ldots, Q_s\}$, where s is the number of queries. Let $\{D_1, \ldots, D_{c_j}\}$ be the set of c_j relevant documents for an information need $Q_j \in Q$. Let R_{jk} be the set of ranked retrieved results ordered from the highest scored document until the relevant document D_k is reached. The MAP score for the set Q is given as

$$MAP = \frac{1}{s}\sum_{j=1}^{s}\frac{1}{c_j}\sum_{k=1}^{c_j} Precision\, R_{jk}, \tag{35}$$

where precision is the fraction of the documents retrieved that are relevant to the query. When a relevant document is not retrieved, the precision value in the above equation is zero.

10. Experimental results

10.1. Qualitative comparison

Comparing the top words, in a qualitative way, of a few example topics is a standard practice to provide an intuition of what topic models represent [21,26,36,44,45,49]. In this section we present the top most likely words of six example topics. They have been selected from a set where there exist overlapped words between the topics generated by different models.Tables 3–5 show these examples for LDA, BiLDA and MiLDA, respectively, where we indicate in bold common words between models.

In these examples, MiLDA assigns higher probability —within the topic's word distribution— to words discovered from different content producers. For example, in addition to the words lens, focus and canon, which have high likelihood in both LDA and BiLDA, the MiLDA brings to the top of its lists interesting words like *bokeh*, *tamron* and *vigneting*. 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". Tamron is a lens manufacturer and vignetting refers to a reduction of an image's brightness or saturation at the periphery compared to the image center.

Table 3
An illustration of top 10 words of six topics generated by LDA for $K = 500$. We show in bold words that also appear as top entries for the other models. LDA yields word-topic distributions, ϕ in Eq. (7), in one distinct vocabulary.

camera	**coffee**	**tan**	**time**	**pillow**	**tea**
lens	**maker**	**color**	**watch**	**neck**	one
canon	**cup**	**skin**	**casio**	**pillows**	**teas**
nikon	**espresso**	**legs**	**shock**	**back**	leaves
video	filter	**self**	digital	**head**	**use**
zoom	brew	**natural**	features	**comfortable**	**cup**
image	**grind**	**glow**	compass	**support**	**loose**
shooting	**grounds**	**tanning**	light	**sleep**	**good**
flash	carafe	look	timex	shape	make
shoot	**water**	**dark**	**resistant**	side	**hot**

Table 4
An illustration of top 10 words of six topics generated by BiLDA for $K = 500$. BiLDA produces word-topic distributions, ϕ and ψ in Eq. (15), in two distinct vocabularies, corresponding to the consumers' vocabulary and sellers' vocabulary.

Sellers	Consumers	Sellers	Consumers	Sellers	Consumers
canon	**camera**	**coffee**	**coffee**	**tan**	**tan**
camera	**canon**	kitchen	**maker**	oz	**color**
nikon	**nikon**	**cup**	**cup**	beauty	**skin**
image	**shoot**	**espresso**	machine	belloccio	**legs**
lens	cameras	**brew**	**use**	glow	**self**
slr	photos	**water**	**espresso**	**body**	n**atural**
cmos	**flash**	brewing	**water**	airbrush	**dark**
zoom	**shooting**	**aeropress**	**grind**	**tanning**	**tanning**
shooting	iso	dcc	one	**self**	**glow**
digital	one	cuisinart	**grounds**	bronze	use
sellers	consumers	sellers	consumers	sellers	consumers
watch	**watch**	**pillow**	**pillow**	**cup**	**tea**
case	watches	**sleep**	**pillows**	cups	**cup**
dial	**time**	**neck**	**neck**	**tea**	cups
resistant	face	innovations	one	oz	**use**
chronograph	wear	**pillows**	**sleep**	kitchen	**teas**
water	**one**	standard	**head**	clear	**hot**
stainless	looks	**back**	**comfortable**	**iced**	water
quartz	**band**	throughout	side	lid	**loose**
steel	**wrist**	cover	**back**	**teas**	using
feet	big	**support**	**night**	drink	great

Table 5
An illustration of top 10 words of six topics generated by MiLDA for $K = 500$. MiLDA outputs word-topic distributions, ϕ, ψ and χ in Eq. (32), in three vocabularies associated with consumers, sellers and shared words.

k = 204			k = 497		
consumers	shared	sellers	consumers	shared	sellers
bokeh	**lens**	chdoh	illy	**espresso**	decalcification
tamron	gopro	florine	tierra	**machine**	joergen
primes	focus	chdhn	robusta	press	normandy
apertures	**canon**	ahdwh	gaggia	**coffee**	actualpress
xti	light	chdhe	robusto	**grind**	aeropressing
pentax	**lenses**	astigmatism	actuality	**beans**	byam
vignetting	quality	ptimized	preground	grinder	tassen
cineform	sharp	tvpl	brikkas	**aeropress**	maile
rendition	great	spectator	tamp	use	agglomerated
shaky	**zoom**	refractive	coarseness	french	seatspring
k = 376			k = 5		
consumers	shared	sellers	consumers	shared	sellers
tanners	**tan**	perfekt	tiesto	**watch**	gwx
rebirthing	**skin**	dcollet	compasses	michael	topring
comatose	lotion	vanillyl	observer	kors	opaline
patchy	**self**	hedera	tides	**time**	luminiscent
jergens	**tanning**	asiatica	baro	**casio**	logoed
pasty	use	biloba	dub	**shock**	inhg
streaked	beauty	laminaria	gulfman	**watches**	xlander
palms	**body**	setaf	dealbreaker	compass	alti
consistancy	**dark**	centella	unreadable	**wrist**	sodw
bothers	natural	ginkgo	timezones	**band**	sliconestainless
k = 15			k = 16		
consumers	shared	sellers	consumers	shared	sellers
fluffed	**pillow**	spasilk	earl	**tea**	mrtea
brest	**pillows**	cabeau	yorkshire	**teas**	regency
excruciating	**neck**	higear	lipton	**leaves**	assam
adequately	**sleep**	waterbase	mississippi	**cup**	ard
cervical	**back**	isocool	finum	**use**	smreya
leaked	**head**	nojo	teabag	**loose**	rosemaryand
husks	**comfortable**	ultrafresh	numi	**hot**	feflectance
annoys	foam	zipcover	marmalade	one	occation
poked	**support**	sqush	pekoe	**good**	drippers
hehe	**night**	sofloft	roommate	**iced**	casca

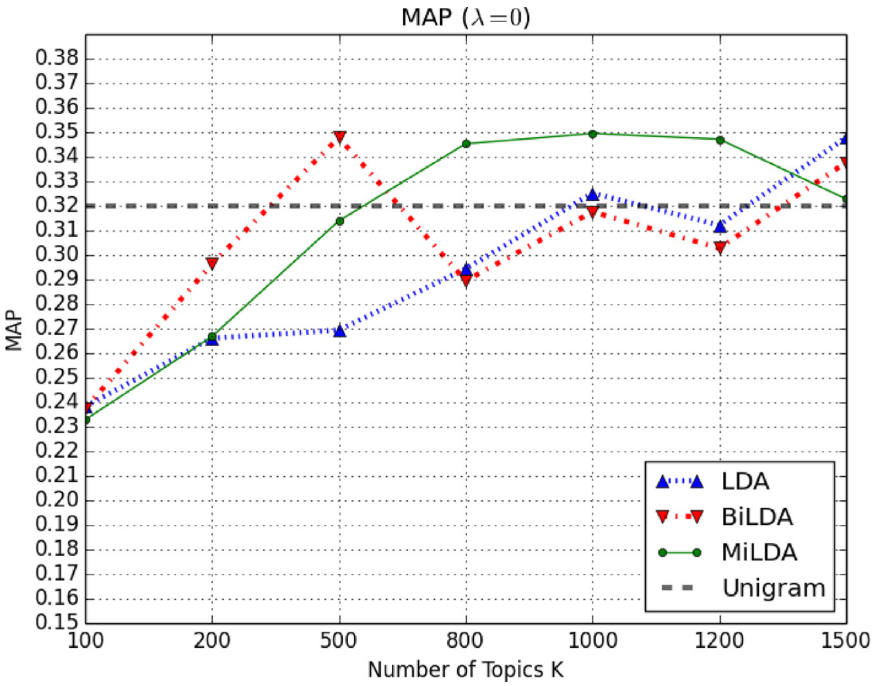


Fig. 8. MAP scores on Webshop Setup for LDA, BiLDA and MiLDA for $\lambda = 0$ with respect to number of topics.

Another example, in the *coffee* topic, the shared word distribution shows typical words, such as *espresso*, *press* or *beans*. On the consumers' side, the MiLDA brings up words like *illy*, *tierra*, *robusta* and *gaggia*, which are all brands related to coffee or coffee machines. This might be also useful for merchants interested in learning about leading or most-talked-about brands. *Tamp* is a small handheld device used to compress ground coffee to prepare espresso.

The topic about *tanning* shows some aspects and attributes particularly related to this type of product, such as *patchy*, *pasty*, *consistency* and *streaked*. This may be particularly helpful for opinion mining applications, where one is often interested, not only in sentiment, but also in the features that are important for consumers. A person might type a query to find out which of these products is not patchy or pasty. The topic about *watches* shows the brand *tiesto* and other nouns relevant to this product. For example, the word *tides* in this context refers to the capability of the watch of measuring ocean tides, which is useful for people who surf and would like to catch good waves. *Baro* refers to measuring the barometric pressure, etc. For the *pillows* topic, we see also related concepts, such as *fluffed, excruciating, annoys, leaked, cervical*, etc.

These examples illustrate how our proposed model can emphasize concepts within the colloquial expression of language and link them to broader topics. Words that were not mentioned in the seller's description show up on consumers' reviews.

The word distribution from the product vocabulary in Table 5 does not add much more interesting information to what is already contained within the shared and users vocabulary. In fact, for some topics, there are less than 10 words with some probability mass. This is because, most words in the product description tend to also be employed by the users in the reviews. We include the product word-topic top most likely words for completeness.

10.2. Quantitative evaluation

Cross-idiomatic linking of pins and webshops. We present the results for the Webshop Setup, where the target collection consists of 1171 webshops, as described in Section 9.2.

In this setup, we study what happens when only the topical knowledge is utilized in the retrieval process. This corresponds to setting λ equal to 0 in Eq. (34). Fig. 8 shows MAP scores with respect to the number of topics for this condition. On the same figure, we plot the constant MAP $= 0.3199$ attained by the unigram model ($\lambda = 1$).

Each topic model is able to outperform the unigram model for at least one value of K, as shown in Fig. 8. This finding strengthens our intuition that topical knowledge, regardless of the chosen model, is extremely useful for bridging the gap between users' informal ecosystems to the more formal environment of e-retailing.

We explore further the performance and qualify each model. The score dynamics are not uniformly distributed between and within topic models. Specifically, BiLDA's performance increases more rapidly than LDA and MiLDA, and it peaks to MAP $= 0.3479$ at $K = 500$. However, it exhibits a slightly erratic behaviour as K changes. The LDA performance increases slower than BiLDA and MiLDA. It then climbs up and peaks to MAP $= 0.3478$ at $K = 1500$. MiLDA's performance increases slightly slower than BiLDA and faster than LDA. Interestingly, unlike the other two models, MiLDA can sustain a MAP score

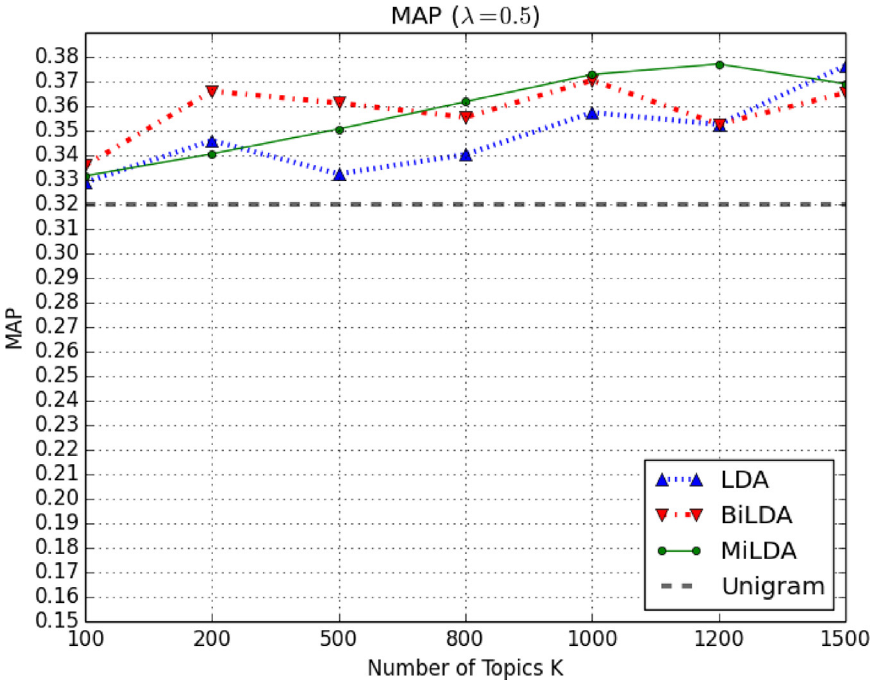


Fig. 9. MAP scores on Webshop Setup for LDA, BiLDA and MiLDA for $\lambda = 0.5$ with respect to number of topics.

Table 6
p-values from the post-hoc Finner test comparing the average precision of MiLDA against other models for $K \geq 500$ and $\lambda \leq 0.5$.

	Unigram	LDA	BiLDA
MiLDA	p = 0.0427	0.0160	0.0018

above 0.34 for different values of K (i.e., K = 800, 1000, 1200). The highest value is MAP = 0.3495 at K = 1000. This sustained behaviour may indicate a better fit to the data compared to the other models. This is because it may be hard to set the number of topics to a precise optimal value. A model that exhibits comparable results across different K values is naturally preferred.

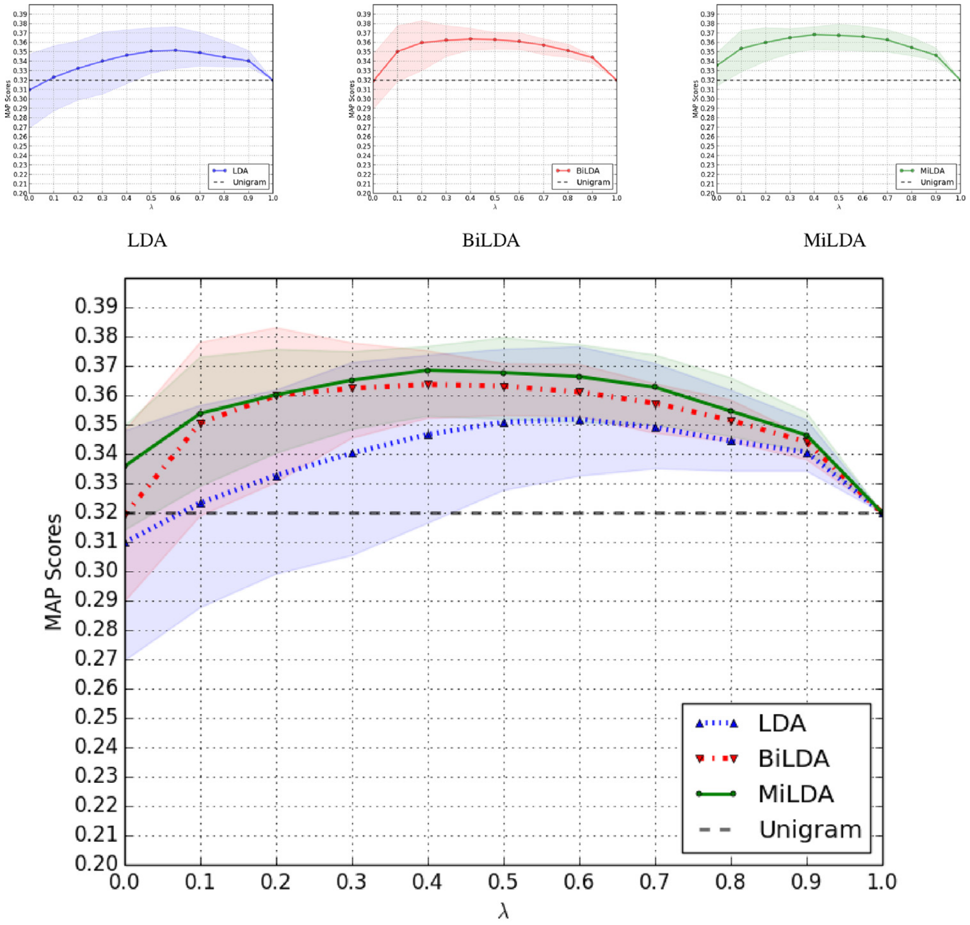
In another evaluation test, we have varied the parameter λ which controls the degree of contribution of the topical part and the unigram part in Eq. (34). Fig. 9 presents the results for $\lambda = 0.5$. This value sets an equal contribution in the interpolation between the topical and unigram representations. Combining both models yields higher scores than each model individually, creating a positive synergy between them. All data points are shifted up under this condition. The maximum MAP score of 0.3796 ($K = 1200$) is obtained by the MiLDA.

We further illustrate how performance varies for different values of λ in Fig. 10. For each value of λ, we average the MAP scores across different numbers of topics ($K \geq 500$ to allow sufficient topic expressiveness) and plot a solid line to indicate this average. The shaded regions indicate the areas spanned by the maximum and minimum attained. They provide a measure of the spread of scores.

MiLDA achieves the highest average scores, closely followed by BiLDA, LDA and lastly the unigram model. For $\lambda < 0.4$, the spread for MiLDA is smaller than LDA and BiLDA. This indicates that MiLDA achieves higher scores than the other two topic models more consistently regardless of the number of topics. As λ increases and the unigram contribution becomes more important, the spread from all topic models reduces until they converge to a single point that corresponds to the full unigram constant (MAP = 0.3199).

We performed a Friedman test to compare the average precision of all models, and obtained $p - value = 0.0110$. We then performed post-hoc pairwise comparisons using the Finner procedure [16,17]. Table 6 presents the p-values. The MiLDA performance is significantly higher than the other models.

Cross-idiomatic linking of pins and products. We now present results for the Product Setup, where the target collection consisted of almost 20,000 individual products, without being grouped into webshops. Fig. 11 shows Precision at 5, i.e., the fraction of products (from the top 5) that were considered relevant for each query by the Crowdflower judges for $\lambda = 0.5$. The family of LDA models outperforms the unigram model. They were all able to achieve the highest value of $P@5 = 0.47$. However, as K varies, MiLDA is the one with the least score variation. Its minimum score is 0.43, while the minimum score



LDA

BiLDA

MiLDA

Superposition of MAP scores for LDA, BiLDA and MiLDA

Fig. 10. MAP scores on Webshop Setup with respect to λ for LDA (left), BiLDA (middle) and MiLDA (right) averaged over different values of $K \geq 500$ (solid lines). Shaded regions represent the maximum and minimum values spanned by the models' scores at particular λ values. The bottom figure supperposes the figures on the top row for comparison purposes. It also shows the results of the unigram model with a dashed line.

for LDA and BiLDA is 0.37 and 0.39, respectively. Just as in the Webshop Setup, BiLDA exhibits a slight erratic behaviour, whereas LDA slowly climbs up in performance as the number of topics increases. The unigram model achieved a lower score: $P@5 = 0.36$. This study validates our findings in the Webshop Setup, where MiLDA yielded an overall better performance.

Perplexity. After training on product-reviews pairs, we also computed perplexity scores for a held-out collection of products to evaluate the ability of the models to generalize on unseen data, as often found in the literature [4,6,50]. We found that for any number of topics, the perplexity score of the MiLDA model is always lower (better) than its counterparts. For the best performing conditions, perplexity scores for LDA, BiLDA and MiLDA are respectively 185.39, 249.645 and 125.484.

11. Further discussion

We have provided a rich representation of cross-idiomatic data that allows us to bridge items from different Web environments. However, it is important to note that there is no ideal representation for all queries. The ability to find the relevant documents depends not only on the chosen document representation, but also on the query itself. While some queries are well captured for all models, some queries are better suited for certain models.

Table 7 presents randomly selected queries with various degrees of performance under different models. For queries easily retrieved under all models, learning from the consumers' vocabulary does not make a difference. On the other hand, queries where the LDA-family perform better than the unigram, illustrate the benefits of exploiting the consumer-generated content. Other examples throughout show that LDA-like models have varying degrees of success depending on the query. The LDA family learns from the same data, but the particular modelling paradigm affects each query individually. Additionally, some queries are not well suited for any of the models. Thus, one limitation of this study is that there is no one model that fits all queries perfectly.

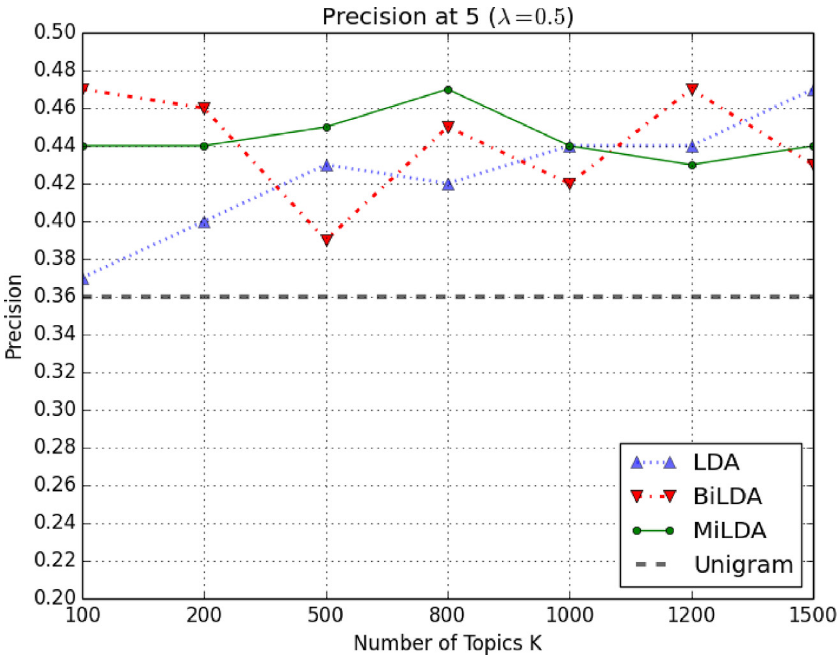


Fig. 11. Precision at 5 (*P*@5) on the Product Setup for LDA, BiLDA, MiLDA, and Unigram, as judged by the Crowdflower annotators.

Table 7
Example queries with average precision under each model.

Example queries easily retrieved by all models	Unigram	LDA	BiLDA	MiLDA
Nikon	1.0000	1.0000	1.0000	1.0000
14kt white gold engagement ring mounting total diamond weight 22 carats msr10150 99 piercing green gem turtle belly ring	0.8714	0.9440	0.8922	0.9132
Jewelry diamond wedding	1.0000	1.0000	0.9959	1.0000
Gold silver earrings 25	0.9000	0.9208	0.8984	0.9156
Rebecca minkoff ily leather tote	1.0000	1.0000	1.0000	1.0000
Example queries where all models have difficulty				
Fashion make mouth red	0.0309	0.0219	0.0459	0.0494
Browning	0.0050	0.0046	0.0041	0.0047
Ten step guide nailing office style	0.0082	0.0037	0.0051	0.0066
Monogrammed toes	0.0273	0.0143	0.0150	0.0156
Christmas cranberry mojitos	0.0157	0.0275	0.0060	0.0250
Example queries where all LDA-like models perform better than unigram				
Gifts 50 coach stone stud earrings	0.5403	0.8231	0.8355	0.9848
Colorful fruit color nail	0.2383	0.4248	0.6476	0.4981
Sexy lace push adjustable bra wendybox	0.6321	0.8401	0.9114	0.9604
Plain low waist thickened slim leisure jeans	0.1565	0.2524	0.2527	0.4106
Lace dress	0.4525	0.7296	0.5982	0.7669
Other Examples				
Retro swimsuit	0.5687	0.2777	0.4492	0.7162
Purity ring 25	0.5671	0.6867	0.6299	0.6855
Prep freeze slow cooker recipes melissafallistest	0.6644	0.7197	0.4992	0.6996
Sled riding	0.2644	0.2140	0.2396	0.5296
Diy hair	0.3741	0.2913	0.7298	0.2647

Another limitation is that performance on the task depends highly on the training data. This can be seen as a drawback for any of the models, as we can only learn from words that we have seen before. However, this is also the case for people. The average human would rarely be able to discover the meaning of a word that she has never seen before in any context.

Additionally, our model has partitioned the vocabulary in a simple way, but more nuanced models that softly assign words to vocabularies may be beneficial.

Regarding training time, we did not experience major differences between the three topic models considered. For a fixed number of topics K, all models take roughly the same time to train. For the best performing conditions, on an Intel(R) Xeon(R) CPU at 2.90GHz the average training time of LDA is 4.19×10^6 seconds (4.9 days), that of BiLDA is 3.98×10^6 seconds (4.6 days) and MiLDA's is 4.22×10^6 seconds (4.9 days). MiLDA has the added complexity that it first has to go through the two vocabularies and label the shared words. However, this is a negligible cost compared to the actual Gibbs sampling training time. The average time to perform a query is 0.3122, 0.3593, and 0.4063 seconds, for LDA, BiLDA and MiLDA, respectively.

In this work, the corpus focuses on e-commerce only. However, this limits they type of concepts that we can learn. Extending the current collection to include different domains of knowledge, e.g., political issues discussed by people of different views (both formally and informally), history events discussed by different countries, even philosophical views, and others may be highly beneficial to broaden the scope of this work.

This way, instead of only using Pinterest data as queries, we could have a large variety of sources (e.g., tweets are more likely to contain political or philosophical statements) that could be linked to other content on the Web, instead of just the e-commerce application we have presented here. We hope that this work sparks the interest of the community to further study the different usages in language across the Web and to create links between content that may not be obviously linked.

12. Conclusions and future work

We set out to create links between different Web sources. This is an important task because we automatically increase the connectivity between items and potentially enable easier navigation accross related content.

To perform the linking task, we proposed an information retrieval framework, where a Web element, from a source environment, is considered a query, whereas other Web elements, from a target environment, are documents to be ranked according to the relevance to the query.

We made an extensive theoretical and didactic comparison between the unigram model, and three different topic models: latent Dirichlet allocation (LDA), bilingual latent Dirichlet allocation (BiLDA) and a novel topic model proposed in this work, the multi-idiomatic latent Dirichlet allocation (MiLDA).

The LDA model is able to reduce the dimensionality of the document representation by clustering words into topics. However it does not explicitly consider the differences in the vocabulary usage or the cross-idiomatic structure that naturally occurs between different Web environments (e.g., between Pinterest and Amazon). Words are drawn from a monolingual vocabulary distribution.

The BiLDA model preserves the dimensionality reduction advantage of the LDA, and it explicitly models the differences between the languages in different environments. Words are either drawn from the source vocabulary or the target vocabulary.

MiLDA also preserves the advantages of LDA and BiLDA but it softens the constraint of two completely different vocabularies assumed in BiLDA. Thus, in addition to having the source and target vocabularies to model language differences, it also has a shared vocabulary to model the similarities between the environments. Therefore, we can draw words not only from the source and target but also from a shared vocabulary. This representation seems theoretically more adequate for the type of data we are modeling. Since we are dealing with different idiomatic expressions of the same language, it seems intuitive to model both differences and similarities in the choice of words.

We carried out an example of the proposed task by linking pins from Pinterest to Amazon products. Single pins were used as queries and sets of Amazon products formed the target collection. We used 100 pins randomly selected from a large pin collection, and almost 20,000 Amazon products. We showed that this task is a cross-idiomatic one, since people from different environments tend to express ideas with different language usages and expressions.

We performed an empirical comparison between the models. Our experiments suggested that MiLDA is better suited for the task, as its overall performance is significantly higher than the other models.

When solely considering the topic representation in the retrieval model, MiLDA is the only one able to sustain its highest MAP score for different values of K; while the other models tend to exhibit larger variations. Furthermore, for any fixed λ, MiLDA consistently obtained the highest MAP scores when averaging over different numbers of topics. It also obtained the smallest spread between the maximum and minimum MAP scores for $\lambda < 0.4$. This suggests robustness regardless of the choice of K.

Our results for the MiLDA model have shown that, for this task, it is useful to partition the vocabulary into shared and non-shared words. In future work, we may consider other techniques to partition the vocabulary. Additionally, it would be worthwhile to explore the use of n-grams on the modelling approaches that we have presented here.

In the future, we may consider extending this study to different training collections of aligned documents, such as news articles paired up with users' comments, scientific articles paired with discussions by laymen or historical events paired with public discussions. These might be interesting scenarios for our framework. The presented corpus can also be used in the comparison of sentiment between seller vs consumer language, as sellers are unlikely to use negative language.

Further optimizations can be made to industrial applications with the proposed methods For example, a convex combination of the three studied models, i.e., LDA, BiLDA and MiLDA, and the unigram model could further improve the retrieval results, for certain applications and where one can rely on a representative development set with ground truth annotations to train the respective weights of each model.

Our proposed task opens possibilities for new applications. Perhaps we can think of the Web, not only in terms of explicitly linked items, but also in terms of implicitly connected ones. Increasing the connectivity between related content in a meaningful way, as we have proposed here, increases the probability of meeting users' information needs.

We foresee that considering the multi-idiomatic nature of language expressions across users, may be extremely useful in opinion mining, sentiment analysis, query expansion, e-commerce applications, social media analysis, dialect mining, cross-lingual information retrieval for closely related languages, and cross-language document retrieval in historic archives [29].

Besides the pure textual representations which were discussed in detail in this article, a natural step in future research is to move towards cross-modal and multi-modal retrieval and linking models. For instance, as mentioned before, users' pins organically come in two different modalities, that is, pictures and photos (vision) accompanied by textual comments provided by the users. These two pieces of information are often complementary, and contain the complete representation of the user's information need only when they are observed together. Therefore, one future research path leads to further investigating and expanding preliminary work in multi-modal information retrieval [10,40,42], and applying the multi-modal models to link content between heterogeneous data and ecosystems.

Acknowledgements

This project is part of the SBO Program of the IWT (IWT-SBO-Nr. 110067).

References

[1] S. Aggarwal, H. van Oostendorp, B. Indurkhya, Automating Web-navigation support using a cognitive model, in: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS), 19, 2014.

[2] Y. Bao, N. Collier, A. Datta, A partially supervised cross-collection topic model for cross-domain text classification, in: Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management, in: CIKM '13, ACM, New York, NY, USA, 2013, pp. 239–248, doi:10.1145/2505515.2505556.

[3] A. Bellogín, J. Wang, P. Castells, Text retrieval methods for item ranking in collaborative filtering, in: Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR), 2011, pp. 301–306.

[4] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[5] R.C. Bunescu, M. Pasca, Using encyclopedic knowledge for named entity disambiguation, in: Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2006, pp. 9–16.

[6] W. Buntine, Estimating likelihoods for topic models, in: Proceedings of the 1st Asian Conference on Machine Learning: Advances in Machine Learning (ACML), 2009, pp. 51–64.

[7] S. Busemann, W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schäfer, H. Uszkoreit, F. Xu, Integrating information extraction and automatic hyperlinking, in: Proceedings of 41st Annual Meeting of the Association for Computational Linguistics (ACL), 2003, pp. 117–120.

[8] J.A. Camacho-Guerrero, A.A. Carvalho, M.G.C. Pimentel, E.V. Munson, A.A. Macedo, Clustering as an approach to support the automatic definition of semantic hyperlinks, in: Proceedings of the 18th Conference on Hypertext and Hypermedia (HT), 2007, pp. 81–84.

[9] C. Carpineto, G. Romano, A survey of automatic query expansion in information retrieval, ACM Comput. Surv. 44 (1) (2012) 1:1–1:50, doi:10.1145/2071389.2071390.

[10] S. Clinchant, J. Ah-Pine, G. Csurka, Semantic combination of textual and visual information in multimedia retrieval, in: Proceedings of the 1st International Conference on Multimedia Retrieval (ICMR), 2011, pp. 44:1–44:8.

[11] A. Costa, F. Roda, Recommender systems by means of information retrieval, in: Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS), 57, 2011.

[12] S. Cucerzan, Large-scale named entity disambiguation based on Wikipedia data, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007, pp. 708–716.

[13] W. De Smet, M.-F. Moens, Cross-language linking of news stories on the Web using interlingual topic modeling, in: Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining (SWSM), 2009, pp. 57–64.

[14] W. De Smet, J. Tang, M.-F. Moens, Knowledge transfer across multilingual corpora via latent topics, in: Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2011, pp. 549–560.

[15] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sc. 41 (6) (1990) 391–407.

[16] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[17] J. Derrac, S. GarcÃa, D. Molina, F. Herrera, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms., Swarm Evol. Comput. 1 (1) (2011) 3–18.

[18] P. Ferragina, U. Scaiella, TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities), in: Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM), 2010, pp. 1625–1628.

[19] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 6 (6) (1984) 721–741.

[20] E. Gilbert, S. Bakhshi, S. Chang, L.G. Terveen, "I need to try this"?: A statistical overview of Pinterest, in: Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems (CHI), 2013, pp. 2427–2436.

[21] T.L. Griffiths, M. Steyvers, Finding scientific topics, in: Proceedings of the National Academy of Sciences of the United States of America (PNAS), 2004, pp. 5228–5235.

[22] B. Hachey, W. Radford, J. Nothman, M. Honnibal, J.R. Curran, Evaluating entity linking with Wikipedia, Artif. Intell. 194 (2013) 130–150.

[23] X. Han, L. Sun, J. Zhao, Collective entity linking in Web text: a graph-based method, in: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2011, pp. 765–774.

[24] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '99, ACM, New York, NY, USA, 1999, pp. 50–57, doi:10.1145/312624.312649.

[25] L. Hong, B.D. Davison, Empirical study of topic modeling in Twitter, in: Proceedings of the 1st Workshop on Social Media Analytics (SOMA), 2010, pp. 80–88.

[26] Y. Hu, J. Boyd-Graber, B. Satinoff, Interactive topic modeling, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, in: HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 248–257.

[27] K.Y. Kamath, A.-M. Popescu, J. Caverlee, Board recommendation in Pinterest, in: Workshops of the 21st Conference on User Modeling, Adaptation, and Personalization (UMAP Workshops), 2013.

[28] K. Kireyev, L. Palen, K. Anderson, Applications of topics models to analysis of disaster-related Twitter data, in: Proceedings of the 2009 NIPS Workshop on Applications for Topic Models: Text and Beyond, 2009.

[29] M. Koolen, F. Adriaans, J. Kamps, M. de Rijke, A cross-language approach to historic document retrieval, in: M. Lalmas, A. MacFarlane, S. Rger, A. Tombros, T. Tsikrika, A. Yavlinsky (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science, 3936, Springer Berlin Heidelberg, 2006, pp. 407–419.
[30] V. Lavrenko, M. Choquette, W.B. Croft, Cross-lingual relevance models, in: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2002, pp. 175–182.
[31] G. Linden, B. Smith, J. York, Amazon.com recommendations: Item-to-item collaborative filtering, IEEE Internet Comput. 7 (1) (2003) 76–80.
[32] A. Mehnaz, J. Warren, M. Orr, Semlink - dynamic generation of hyperlinks to enhance patient readability of discharge summaries, in: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS), 2013, pp. 35–40.
[33] R. Mehrotra, S. Sanner, W. Buntine, L. Xie, Improving LDA topic models for microblogs via tweet pooling and automatic labeling, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2013, pp. 889–892.
[34] R. Mihalcea, A. Csomai, Wikify!: Linking documents to encyclopedic knowledge, in: Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM), 2007, pp. 233–242.
[35] D.N. Milne, I.H. Witten, Learning to link with Wikipedia, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM), 2008, pp. 509–518.
[36] D. Mimno, H. Wallach, J. Naradowsky, D.A. Smith, A. McCallum, Polylingual topic models, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2009, pp. 880–889.
[37] X. Ni, J.-T. Sun, J. Hu, Z. Chen, Mining multilingual topics from Wikipedia, in: Proceedings of the 18th International World Wide Web Conference (WWW), 2009, pp. 1155–1156.
[38] J. Nielsen, Multimedia and Hypertext: The Internet and Beyond, Academic Press Professional, Inc., 1995.
[39] M.J. Paul, R. Girju, Cross-cultural analysis of blogs and forums with mixed-collection topic models, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2009, pp. 1408–1417.
[40] J.C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G.R.G. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 521–535.
[41] A.-M. Popescu, Pinteresting: Towards a better understanding of user interests, in: Proceedings of the 2012 Workshop on Data-driven User Behavioral Modelling and Mining from Social Media (DUBMMSM), 2012, pp. 11–12.
[42] N. Rasiwasia, J.C. Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of the 18th International Conference on Multimedia (MM), 2010, pp. 251–260.
[43] L.-A. Ratinov, D. Roth, D. Downey, M. Anderson, Local and global algorithms for disambiguation to Wikipedia, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), 2011, pp. 1375–1384.
[44] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, in: UAI '04, AUAI Press, Arlington, Virginia, United States, 2004, pp. 487–494.
[45] M. Steyvers, T. Griffiths, Probabilistic topic models, Handbook of Latent Semantic Analysis 427 (7) (2007) 424–440.
[46] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, X. He, Interpreting the public sentiment variations on Twitter, IEEE Trans. Knowl. Data Eng. 26 (5) (2014) 1158–1170.
[47] I. Vulić, W.D. Smet, J. Tang, M. Moens, Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications, Inf. Process. Manage. 51 (1) (2015) 111–147.
[48] I. Vulić, S. Zoghbi, M.-F. Moens, Learning to bridge colloquial and formal language applied to linking and search of e-commerce data, in: Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2014, pp. 1195–1198.
[49] H.M. Wallach, Topic modeling: Beyond bag-of-words, in: Proceedings of the 23rd International Conference on Machine Learning, in: ICML '06, ACM, New York, NY, USA, 2006, pp. 977–984, doi:10.1145/1143844.1143967.
[50] H.M. Wallach, I. Murray, R. Salakhutdinov, D. Mimno, Evaluation methods for topic models, in: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), 2009, pp. 1105–1112.
[51] X. Wei, W.B. Croft, LDA-based document models for ad-hoc retrieval, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2006, pp. 178–185.
[52] B. Xu, Y.H. Yu, Simplified recommendation algorithm based on content and clustering in e-commerce, Adv. Mater. Res. 403–408 (2011) 2498–2501.
[53] J. Xu, R. Weischedel, C. Nguyen, Evaluating a probabilistic model for cross-lingual information retrieval, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '01, ACM, New York, NY, USA, 2001, pp. 105–110, doi:10.1145/383952.383968.
[54] G.-R. Xue, W. Dai, Q. Yang, Y. Yu, Topic-bridged plsa for cross-domain text classification, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '08, ACM, New York, NY, USA, 2008, pp. 627–634, doi:10.1145/1390334.1390441.
[55] P. Yang, W. Gao, Q. Tan, K. Wong, A link-bridged topic model for cross-domain document classification, Inf. Process. Manage. 49 (6) (2013) 1181–1193, doi:10.1016/j.ipm.2013.05.002.
[56] S.-H. Yang, S.P. Crain, H. Zha, Bridging the language gap: Topic adaptation for documents with different technicality., in: G.J. Gordon, D.B. Dunson, M. DudÃk (Eds.), AISTATS, JMLR Proceedings, 15, JMLR.org, 2011, pp. 823–831.
[57] J. Yu, S. Mohan, D. Putthividhya, W.-K. Wong, Latent Dirichlet allocation based diversified retrieval for e-commerce search, in: Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM), 2014, pp. 463–472.
[58] M. Zarro, C. Hall, Pinterest: Social collecting for #Linking #Using #Sharing, in: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), 2012, pp. 417–418.
[59] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to information retrieval, ACM Trans. Inf. Syst. 22 (2) (2004) 179–214.
[60] L. Zhao, Modeling and Solving Term Mismatch for Full-text Retrieval., Ph.D. thesis, Carnegie Mellon University, 2012.
[61] S. Zoghbi, I. Vulić, M.-F. Moens, Are words enough?: A study on text-based representations and retrieval models for linking pins to online shops, in: Proceedings of the 2013 CIKM UnstructureNLP Workshop, 2013, pp. 45–52.
[62] S. Zoghbi, I. Vulić, M.-F. Moens, I pinned it. Where can i buy one like it?: Automatically linking Pinterest pins to online webshops, in: Proceedings of the 2013 Workshop on Data-driven User Behavioral Modelling and Mining from Social Media, in: DUBMOD '13, ACM, New York, NY, USA, 2013, pp. 9–12, doi:10.1145/2513577.2513581.