**Jordan Boyd-Graber** and Philip Resnik. **Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation**. *Empirical Methods in Natural Language Processing*, 2010, 11 pages.

```
@inproceedings{Boyd-Graber:Resnik-2010,
Author = {Jordan Boyd-Graber and Philip Resnik},
Url = {docs/jbg-mlslda-2010.pdf},
Booktitle = {Empirical Methods in Natural Language Processing},
Location = {Cambridge, MA},
Year = {2010},
Title = {Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation},
}
```

Links:

- Data [http://cs.colorado.edu/~jbg/downloads/german_movies.tar.gz]

# Holistic Sentiment Analysis Across Languages:
# Multilingual Supervised Latent Dirichlet Allocation

**Jordan Boyd-Graber**
UMD iSchool
and UMIACS
University of Maryland
College Park, MD
jbg@umiacs.umd.edu

**Philip Resnik**
Department of Linguistics
and UMIACS
University of Maryland
College Park, MD
resnik@umd.edu

## Abstract

In this paper, we develop multilingual supervised latent Dirichlet allocation (MLSLDA), a probabilistic generative model that allows insights gleaned from one language's data to inform how the model captures properties of other languages. MLSLDA accomplishes this by jointly modeling two aspects of text: how multilingual concepts are clustered into thematically coherent topics and how topics associated with text connect to an observed regression variable (such as ratings on a sentiment scale). Concepts are represented in a general hierarchical framework that is flexible enough to express semantic ontologies, dictionaries, clustering constraints, and, as a special, degenerate case, conventional topic models. Both the topics and the regression are discovered via posterior inference from corpora. We show MLSLDA can build topics that are consistent across languages, discover sensible bilingual lexical correspondences, and leverage multilingual corpora to better predict sentiment.

Sentiment analysis (Pang and Lee, 2008) offers the promise of automatically discerning how people feel about a product, person, organization, or issue based on what they write online, which is potentially of great value to businesses and other organizations. However, the vast majority of sentiment resources and algorithms are limited to a single language, usually English (Wilson, 2008; Baccianella and Sebastiani, 2010). Since no single language captures a majority of the content online, adopting such a limited approach in an increasingly global community risks missing important details and trends that might only be available when text in multiple languages is taken into account.

Up to this point, multiple languages have been addressed in sentiment analysis primarily by transferring knowledge from a resource-rich language to a less rich language (Banea et al., 2008), or by ignoring differences in languages via translation into English (Denecke, 2008). These approaches are limited to a view of sentiment that takes place through an English-centric lens, and they ignore the potential to share information between languages. Ideally, learning sentiment cues holistically, *across* languages, would result in a richer and more globally consistent picture.

In this paper, we introduce Multilingual Supervised Latent Dirichlet Allocation (MLSLDA), a model for sentiment analysis on a multilingual corpus. MLSLDA discovers a consistent, unified picture of sentiment across multiple languages by learning "topics," probabilistic partitions of the vocabulary that are consistent in terms of both meaning and relevance to observed sentiment. Our approach makes few assumptions about available resources, requiring neither parallel corpora nor machine translation.

The rest of the paper proceeds as follows. In Section 1, we describe the probabilistic tools that we use to create consistent topics bridging across languages and the MLSLDA model. In Section 2, we present the inference process. We discuss our set of semantic bridges between languages in Section 3, and our experiments in Section 4 demonstrate that this approach functions as an effective multilingual topic model, discovers sentiment-biased topics, and uses multilingual corpora to make better sentiment predictions across languages. Sections 5 and 6 discuss related research and discusses future work, respectively.

# 1 Predictions from Multilingual Topics

As its name suggests, MLSLDA is an extension of Latent Dirichlet allocation (LDA) (Blei et al., 2003), a modeling approach that takes a corpus of unannotated documents as input and produces two outputs, a set of "topics" and assignments of documents to topics. Both the topics and the assignments are probabilistic: a topic is represented as a probability distribution over words in the corpus, and each document is assigned a probability distribution over all the topics. Topic models built on the foundations of LDA are appealing for sentiment analysis because the learned topics can cluster together sentiment-bearing words, and because topic distributions are a parsimonious way to represent a document.[1]

LDA has been used to discover latent structure in text (e.g. for discourse segmentation (Purver et al., 2006) and authorship (Rosen-Zvi et al., 2004)). MLSLDA extends the approach by ensuring that this latent structure — the underlying topics — is consistent across languages. We discuss multilingual topic modeling in Section 1.1, and in Section 1.2 we show how this enables supervised regression regardless of a document's language.

## 1.1 Capturing Semantic Correlations

Topic models posit a straightforward generative process that creates an observed corpus. For each document $d$, some distribution $\theta_d$ over unobserved topics is chosen. Then, for each word position in the document, a topic $z$ is selected. Finally, the word for that position is generated by selecting from the topic indexed by $z$. (Recall that in LDA, a "topic" is a distribution over words).

In monolingual topic models, the topic distribution is usually drawn from a Dirichlet distribution. Using Dirichlet distributions makes it easy to specify sparse priors, and it also simplifies posterior inference because Dirichlet distributions are conjugate to multinomial distributions. However, drawing topics from Dirichlet distributions will not suffice if our vocabulary includes multiple languages. If we are working with English, German, and Chinese at the same time, a Dirichlet prior has no way to favor distributions $z$ such that $p(\text{good}|z)$, $p(\text{gut}|z)$, and

---

[1]The latter property has also made LDA popular for information retrieval (Wei and Croft, 2006)).

$p(\text{hǎo}|z)$ all tend to be high at the same time, or low at the same time. More generally, the structure of our model must encourage topics to be consistent across languages, and Dirichlet distributions cannot encode correlations between elements.

One possible solution to this problem is to use the multivariate normal distribution, which can produce correlated multinomials (Blei and Lafferty, 2005), in place of the Dirichlet distribution. This has been done successfully in multilingual settings (Cohen and Smith, 2009). However, such models complicate inference by not being conjugate.

Instead, we appeal to tree-based extensions of the Dirichlet distribution, which has been used to induce correlation in semantic ontologies (Boyd-Graber et al., 2007) and to encode clustering constraints (Andrzejewski et al., 2009). The key idea in this approach is to assume the vocabularies of all languages are organized according to some shared semantic structure that can be represented as a tree. For concreteness in this section, we will use WordNet (Miller, 1990) as the representation of this multilingual semantic bridge, since it is well known, offers convenient and intuitive terminology, and demonstrates the full flexibility of our approach. However, the model we describe generalizes to any tree-structured representation of multilingual knowledge; we discuss some alternatives in Section 3.

WordNet organizes a vocabulary into a rooted, directed acyclic graph of nodes called synsets, short for "synonym sets." A synset is a child of another synset if it satisfies a hyponomy relationship; each child "is a" more specific instantiation of its parent concept (thus, hyponomy is often called an "isa" relationship). For example, a "dog" is a "canine" is an "animal" is a "living thing," etc. As an approximation, it is not unreasonable to assume that WordNet's structure of meaning is language independent, i.e. the concept encoded by a synset can be realized using terms in different languages that share the same meaning. In practice, this organization has been used to create many alignments of international WordNets to the original English WordNet (Ordan and Wintner, 2007; Sagot and Fišer, 2008; Isahara et al., 2008).

Using the structure of WordNet, we can now describe a generative process that produces a distribution over a multilingual vocabulary, which encourages correlations between words with similar mean-

ings regardless of what language each word is in. For each synset $h$, we create a multilingual word distribution for that synset as follows:

1. Draw transition probabilities $\beta_h \sim \text{Dir}\,(\tau_h)$
2. Draw stop probabilities $\omega_h \sim \text{Dir}\,(\kappa_h)$
3. For each language $l$, draw emission probabilities for that synset $\phi_{h,l} \sim \text{Dir}\,(\pi_{h,l})$.

For conciseness in the rest of the paper, we will refer to this generative process as *multilingual Dirichlet hierarchy*, or $\text{MULTDIRHIER}(\tau, \kappa, \pi)$.[2] Each observed token can be viewed as the end result of a sequence of visited synsets $\lambda$. At each node in the tree, the path can end at node $i$ with probability $\omega_{i,1}$, or it can continue to a child synset with probability $\omega_{i,0}$. If the path continues to another child synset, it visits child $j$ with probability $\beta_{i,j}$. If the path ends at a synset, it generates word $k$ with probability $\phi_{i,l,k}$.[3] The probability of a word being emitted from a path with visited synsets $r$ and final synset $h$ in language $l$ is therefore

$$p(w, \lambda = r, h|l, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\phi}) =$$
$$\left( \prod_{(i,j) \in r} \beta_{i,j} \omega_{i,0} \right) (1 - \omega_{h,1}) \phi_{h,l,w}. \quad (1)$$

Note that the stop probability $\omega_h$ is independent of language, but the emission $\phi_{h,l}$ is dependent on the language. This is done to prevent the following scenario: while synset $A$ is highly probable in a topic and words in language 1 attached to that synset have high probability, words in language 2 have low probability. If this could happen for many synsets in a topic, an entire language would be effectively silenced, which would lead to inconsistent topics (e.g.

---

Topic 1 is about baseball in English and about travel in German). Separating path from emission helps ensure that topics are consistent across languages.

Having defined topic distributions in a way that can preserve cross-language correspondences, we now use this distribution within a larger model that can discover cross-language patterns of use that predict sentiment.

## 1.2 The MLSLDA Model

We will view sentiment analysis as a regression problem: given an input document, we want to predict a real-valued observation $y$ that represents the sentiment of a document. Specifically, we build on supervised latent Dirichlet allocation (SLDA, (Blei and McAuliffe, 2007)), which makes predictions based on the topics expressed in a document; this can be thought of projecting the words in a document to low dimensional space of dimension equal to the number of topics. Blei et al. showed that using this latent topic structure can offer improved predictions over regressions based on words alone, and the approach fits well with our current goals, since word-level cues are unlikely to be identical across languages. In addition to text, SLDA has been successfully applied to other domains such as social networks (Chang and Blei, 2009) and image classification (Wang et al., 2009). The key innovation in this paper is to extend SLDA by creating topics that are globally consistent across languages, using the bridging approach above.

We express our model in the form of a probabilistic generative latent-variable model that generates documents in multiple languages *and* assigns a real-valued score to each document. The score comes from a normal distribution whose sum is the dot product between a regression parameter $\eta$ that encodes the influence of each topic on the observation and a variance $\sigma^2$. With this model in hand, we use statistical inference to determine the distribution over latent variables that, given the model, best explains observed data.

The generative model is as follows:

1. For each topic $i = 1 \ldots K$, draw a topic distribution $\{\boldsymbol{\beta}_i, \boldsymbol{\omega}_i, \boldsymbol{\phi}_i\}$ from $\text{MULTDIRHIER}(\tau, \kappa, \pi)$.
2. For each document $d = 1 \ldots M$ with language $l_d$:
   (a) Choose a distribution over topics $\theta_d \sim \text{Dir}\,(\alpha)$.

(b) For each word in the document $n = 1 \ldots N_d$, choose a topic assignment $z_{d,n} \sim \text{Mult}\left(\theta_d\right)$ and a path $\lambda_{d,n}$ ending at word $w_{d,n}$ according to Equation 1 using $\{\boldsymbol{\beta}_{z_{d,n}}, \boldsymbol{\omega}_{z_{d,n}}, \boldsymbol{\phi}_{z_{d,n}}\}$.

3. Choose a response variable from $y \sim \text{Norm}\left(\eta^\top \bar{z}, \sigma^2\right)$, where $\bar{z}_d \equiv \frac{1}{N}\sum_{n=1}^{N} z_{d,n}$.

Crucially, note that the topics are not independent of the sentiment task; the regression encourages terms with similar effects on the observation $y$ to be in the same topic. The consistency of topics described above allows the same regression to be done for the entire corpus regardless of the language of the underlying document.

## 2 Inference

Finding the model parameters most likely to explain the data is a problem of statistical inference. We employ stochastic EM (Diebolt and Ip, 1996), using a Gibbs sampler for the E-step to assign words to paths and topics. After randomly initializing the topics, we alternate between sampling the topic and path of a word $(z_{d,n}, \lambda_{d,n})$ and finding the regression parameters $\eta$ that maximize the likelihood. We jointly sample the topic and path conditioning on all of the other path and document assignments in the corpus, selecting a path and topic with probability

$$p(z_n = k, \lambda_n = r | \boldsymbol{z}_{-n}, \boldsymbol{\lambda}_{-n}, w_n, \eta, \sigma, \Theta) =$$
$$p(y_d|\boldsymbol{z}, \eta, \sigma)p(\lambda_n = r|z_n = k, \boldsymbol{\lambda}_{-n}, w_n, \boldsymbol{\tau}, \boldsymbol{\kappa}, \boldsymbol{\pi})$$
$$p(z_n = k|\boldsymbol{z}_{-n}, \alpha). \tag{2}$$

Each of these three terms reflects a different influence on the topics from the vocabulary structure, the document's topics, and the response variable. In the next paragraphs, we will expand each of them to derive the full conditional topic distribution.

As discussed in Section 1.1, the structure of the topic distribution encourages terms with the same meaning to be in the same topic, even across languages. During inference, we marginalize over possible multinomial distributions $\beta$, $\omega$, and $\phi$, using the observed transitions from $i$ to $j$ in topic $k$; $T_{k,i,j}$, stop counts in synset $i$ in topic $k$, $O_{k,i,0}$; continue counts in synsets $i$ in topic $k$, $O_{k,i,1}$; and emission counts in synset $i$ in language $l$ in topic $k$, $F_{k,i,l}$. The
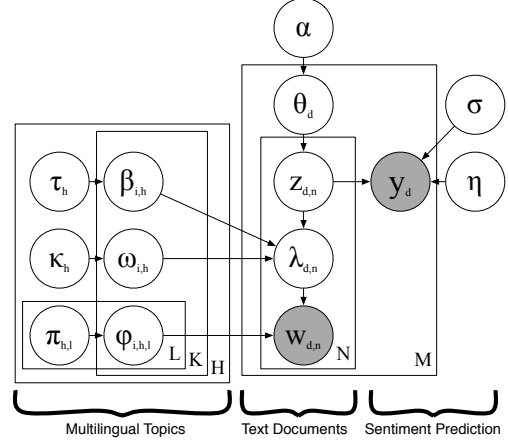


Figure 1: Graphical model representing MLSLDA. Shaded nodes represent observations, plates denote replication, and lines show probabilistic dependencies.

probability of taking a path $r$ is then

$$p(\lambda_n = r|z_n = k, \boldsymbol{\lambda}_{-n}) =$$
$$\underbrace{\prod_{(i,j)\in r} \left( \frac{B_{k,i,j} + \tau_{i,j}}{\sum_{j'} B_{k,i,j'} + \tau_{i,j}} \frac{O_{k,i,1} + \omega_i}{\sum_{s\in 0,1} O_{k,i,s} + \omega_{i,s}} \right)}_{\text{Transition}}$$
$$\underbrace{\frac{O_{k,r_{end},0} + \omega_{r_{end}}}{\sum_{s\in 0,1} O_{k,r_{end},s} + \omega_{r_{end},s}} \frac{F_{k,r_{end},w_n} + \pi_{r_{end},l}}{\sum_{w'} F_{r_{end},w'} + \pi_{r_{end},w'}}}_{\text{Emission}}.$$
$$\tag{3}$$

Equation 3 reflects the *multilingual* aspect of this model. The conditional topic distribution for SLDA (Blei and McAuliffe, 2007) replaces this term with the standard Multinomial-Dirichlet. However, we believe this is the first published SLDA-style model using MCMC inference, as prior work has used variational inference (Blei and McAuliffe, 2007; Chang and Blei, 2009; Wang et al., 2009).

Because the observed response variable depends on the topic assignments of a document, the conditional topic distribution is shifted toward topics that explain the observed response. Topics that move the predicted response $\hat{y}_d$ toward the true $y_d$ will be favored. We drop terms that are constant across all

topics for the effect of the response variable,

$$p(y_d|\boldsymbol{z}, \eta, \sigma) \propto$$

$$\underbrace{\exp\left[\frac{1}{\sigma^2}\left(y_d - \frac{\sum_{k'} N_{d,k'}\eta_{k'}}{\sum_{k'} N_{d,k'}}\right)\frac{\eta_{z_k}}{\sum_{k'} N_{d,k'}}\right]}_{\text{Other words' influence}}$$

$$\underbrace{\exp\left[\frac{-\eta_{z_k}^2}{2\sigma^2 \sum_{k'} N_{d,k'}^2}\right]}_{\text{This word's influence}}. \qquad (4)$$

The above equation represents the *supervised* aspect of the model, which is inherited from SLDA.

Finally, there is the effect of the topics already assigned to a document; the conditional distribution favors topics already assigned in a document,

$$p(z_n = k|\boldsymbol{z}_{-n}, \alpha) = \frac{T_{d,k} + \alpha_k}{\sum_{k'} T_{d,k'} + \alpha_{k'}}. \qquad (5)$$

This term represents the *document* focus of this model; it is present in all Gibbs sampling inference schemes for LDA (Griffiths and Steyvers, 2004).

Multiplying together Equations 3, 4, and 5 allows us to sample a topic using the conditional distribution from Equation 2, based on the topic and path of the other words in all languages. After sampling the path and topic for each word in a document, we then find new regression parameters $\eta$ that maximize the likelihood conditioned on the current state of the sampler. This is simply a least squares regression using the topic assignments $\bar{z}_d$ to predict $y_d$.

Prediction on documents for which we don't have an observed $y_d$ is equivalent to marginalizing over $y_d$ and sampling topics for the document from Equations 3 and 5. The prediction for $y_d$ is then the dot product of $\eta$ and the empirical topic distribution $\bar{z}_d$.

We initially optimized all hyperparameters using slice sampling. However, we found that the regression variance $\sigma^2$ was not stable. Optimizing $\sigma^2$ seems to balance between modeling the language in the documents and the prediction, and thus is sensitive to documents' length. Given this sensitivity, we did not optimize $\sigma^2$ for our prediction experiments in Section 4, but instead kept it fixed at 0.25. We leave optimizing this variable, either through cross validation or adapting the model, to future work.

## 3 Bridges Across Languages

In Section 1.1, we described connections across languages as offered by semantic networks in a general way, using WordNet as an example. In this section, we provide more specifics, as well as alternative ways of building semantic connections across languages.

**Flat** First, we can consider a degenerate mapping that is nearly equivalent to running SLDA independently across multiple languages, relating topics only based on the impact on the response variable. Consider a degenerate tree with only one node, with all words in all languages associated with that node. This is consistent with our model, but there is really no shared semantic space, as all emitted words must come from this degenerate "synset" and the model only represents the output distribution for this single node.

**WordNet** We took the alignment of GermaNet to WordNet 1.6 (Kunze and Lemnitzer, 2002) and removed all synsets that were had no mapped German words. Any German synsets that did not have English translations had their words mapped to the lowest extant English hypernym (e.g. "beinbruch," a broken leg, was mapped to "fracture"). We stemmed all words to account for inflected forms not being present (Porter and Boulton, 1970). An example of the paths for the German word "wunsch" (wish, request) is shown in Figure 2(a).

**Dictionaries** A dictionary can be viewed as a many to many mapping, where each entry $e_i$ maps one or more words in one language $s_i$ to one or more words $t_i$ in another language. Entries were taken from an English-German dictionary (Richter, 2008) a Chinese-English dictionary (Denisowski, 1997), and a Chinese-German dictionary (Hefti, 2005). As with WordNet, the words in entries for English and German were stemmed to improve coverage. An example for German is shown in Figure 2(b).

**Algorithmic Connections** In addition to hand-curated connections across languages, one could also consider automatic means of mapping across languages, such as using edit distance or local context (Haghighi et al., 2008; Rapp, 1995) or using a lexical translation table obtained from parallel text (Melamed, 1998). While we experimented
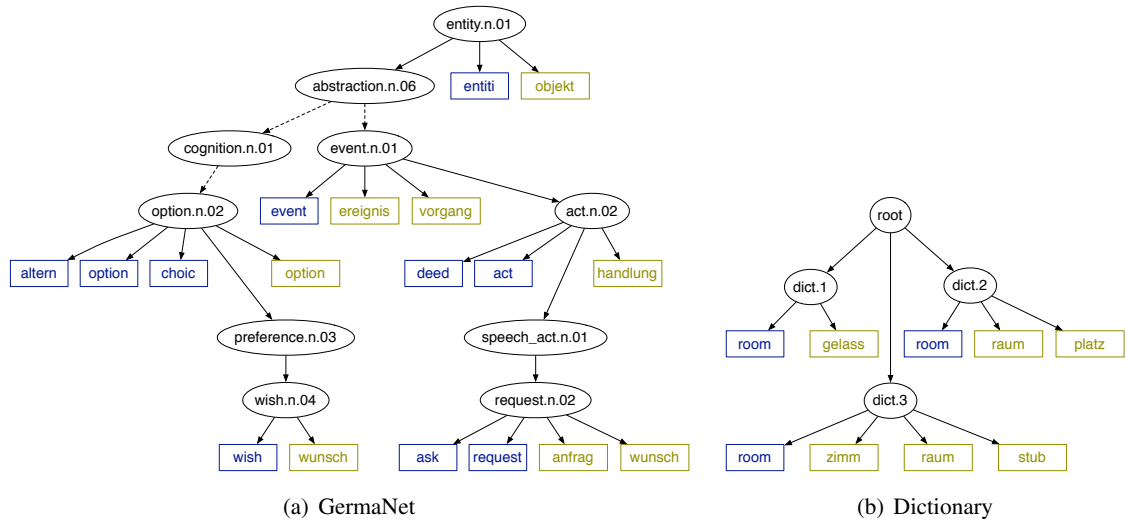
(a) GermaNet        (b) Dictionary

Figure 2: Two methods for constructing multilingual distributions over words. On the left, paths to the German word "wunsch" in GermaNet are shown. On the right, paths to the English word "room" are shown. Both English and German words are shown; some internal nodes in GermaNet have been omitted for space (represented by dashed lines). Note that different senses are denoted by different internal paths, and that internal paths are distinct from the per-language expression.

with these techniques, constructing appropriate hierarchies from these resources required many arbitrary decisions about cutoffs and which words to include. Thus, we do not consider them in this paper.

## 4 Experiments

We evaluate MLSLDA on three criteria: how well it can discover consistent topics across languages for matching parallel documents, how well it can discover sentiment-correlated word lists from non-aligned text, and how well it can predict sentiment.

### 4.1 Matching on Multilingual Topics

We took the 1996 documents from the Europarl corpus (Koehn, 2005) using three bridges: GermaNet, dictionary, and the uninformative flat matching.[4] The model is unaware that the translations of documents in one language are present in the other language. Note that this does not use the supervised framework

(as there is no associated response variable for Europarl documents); this experiment is to demonstrate the effectiveness of the multilingual aspect of the model. To test whether the topics learned by the model are consistent across languages, we represent each document using the probability distribution $\theta_d$ over topic assignments. Each $\theta_d$ is a vector of length $K$ and is a language-independent representation of the document.

For each document in one language, we computed the Hellinger distance between it and all of the documents in the other language and sorted the documents by decreasing distance. The translation of the document is somewhere in that set; the higher the normalized rank (the percentage of documents with a rank lower than the translation of the document), the better the underlying topic model connects languages.

We compare three bridges against what is to our knowledge the only other topic model for unaligned text, Multilingual Topics for Unaligned Text (Boyd-Graber and Blei, 2009).[5]

---

[4]For English and German documents in all experiments, we removed stop words (Loper and Bird, 2002), stemmed words (Porter and Boulton, 1970), and created a vocabulary of the most frequent 5000 words per language (this vocabulary limit was mostly done to ensure that the dictionary-based bridge was of manageable size). Documents shorter than fifty content words were excluded.

[5]The bipartite matching was initialized with the dictionary weights as specified by the Multilingual Topics for Unaligned Text algorithm. The matching size was limited to 250 and the bipartite matching was only updated on the initial iteration then held fixed. This yielded results comparable to when the matching
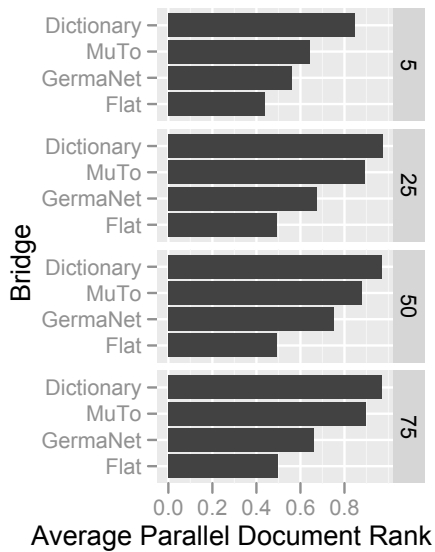
Figure 3: Average rank of paired translation document recovered from the multilingual topic model. Random guessing would yield 0.5; MLSLDA with a dictionary based matching performed best.

Figure 3 shows the results of this experiment. The dictionary-based bridge had the best performance on the task, ranking a large proportion of documents (0.95) below the translated document once enough topics were available. Although GermaNet is richer, its coverage is incomplete; the dictionary structure had a much larger vocabulary and could build a more complete multilingual topics. Using comparable input information, this more flexible model performed better on the matching task than the existing multilingual topic model available for unaligned text. The degenerate flat bridge did no better than the baseline of random guessing, as expected.

## 4.2 Qualitative Sentiment-Correlated Topics

One of the key tasks in sentiment analysis has been the collection of lists of words that convey sentiment (Wilson, 2008; Riloff et al., 2003). These resources are often created using or in reference to resources like WordNet (Whitelaw et al., 2005; Baccianella and Sebastiani, 2010). MLSLDA provides a method for extracting topical and sentiment-correlated word lists from multilingual corpora. If

---

was updated more frequently.

a WordNet-like resource is used as the bridge, the resulting topics are distributions over synsets, not just over words.

As our demonstration corpus, we used the Amherst Sentiment Corpus (Constant et al., 2009), as it has documents in multiple languages (English, Chinese, and German) with numerical assessments of sentiment (number of stars assigned to the review). We segmented the Chinese text (Tseng et al., 2005) and used a classifier trained on character n-grams to remove English-language documents that were mixed in among the Chinese and German language reviews.

Figure 4 shows extracted topics from German-English and German-Chinese corpora. MLSLDA is able to distinguish sentiment-bearing topics from content bearing topics. For example; in the German-English corpus, "food" and "children" topics are not associated with a consistent sentiment signal, while "religion" is associated with a more negative sentiment. In contrast, in the German-Chinese corpus, the "religion/society" topic is more neutral, and the gender-oriented topic is viewed more negatively. Negative sentiment-bearing topics have reasonable words such as "pages," "kǒng pà" (Chinese for "I'm afraid that . . . ") and "tuo" (Chienese for "discard"), and positive sentiment-bearing topics have reasonable words such as "great," "good," and "juwel" (German for "jewel").

The qualitative topics also betray some of the weaknesses of the model. For example, in one of the negative sentiment topics, the German word "gut" (good) is present. Because topics are distributions over words, they can encode the presence of negations like "kein" (no) and "nicht" (not), but not collocations like "nicht gut." More elaborate topic models that can model local syntax and collocations (Johnson, 2010) provide options for addressing such problems.

We do not report the results for sentiment prediction for this corpus because the baseline of predicting a positive review is so strong; most algorithms do extremely well by always predicting a positive review, ours included.

## 4.3 Sentiment Prediction

We gathered 330 film reviews from a German film review site (Vetter et al., 2000) and combined them with a much larger English film review corpus of over

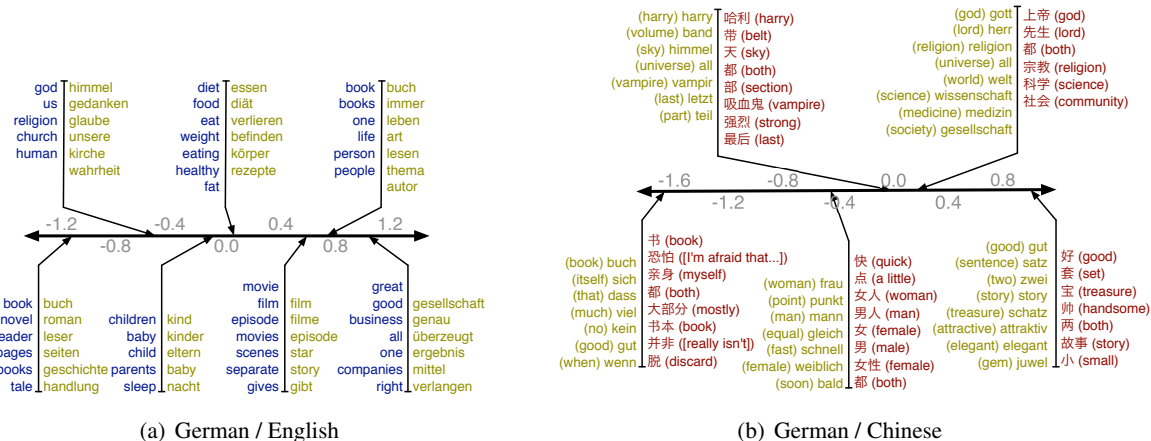(a) German / English          (b) German / Chinese

Figure 4: Topics, along with associated regression coefficient $\eta$ from a learned 25-topic model on German-English (left) and German-Chinese (right) documents. Notice that theme-related topics have regression parameter near zero, topics discussing the number of pages have negative regression parameters, topics with "good," "great," "hǎo" (good) and "überzeugt" (convinced) have positive regression parameters. For the German-Chinese corpus, note the presence of "gut" (good) in one of the negative sentiment topics, showing the difficulty of learning collocations.

| Train | Test | GermaNet | Dictionary | Flat |
|-------|------|----------|------------|------|
| DE | DE | 73.8 | 24.8 | 92.2 |
| EN | DE | 7.44 | 2.68 | 18.3 |
| EN + DE | DE | **1.17** | **1.46** | **1.39** |

Table 1: Mean squared error on a film review corpus. All results are on the same German test data, varying the training data. Over-fitting prevents the model learning on the German data alone; adding English data to the mix allows the model to make better predictions.

5000 film reviews (Pang and Lee, 2005) to create a multilingual film review corpus.[6]

The results for predicting sentiment in German documents with 25 topics are presented in Table 1. On a small monolingual corpus, prediction is very poor. The model over-fits, especially when it has the entire vocabulary to select from. The slightly better performance using GermaNet and a dictionary as topic priors can be viewed as basic feature selection, removing proper names from the vocabulary to

---

[6]We followed Pang and Lee's method for creating a numerical score between 0 and 1 from a star rating. We then converted that to an integer by multiplying by 100; this was done because initial data preprocessing assumed integer values (although downstream processing did not assume integer values). The German movie review corpus is available at http://www.umiacs.umd.edu/~jbg/static/downloads_and_media.html

prevent over-fitting.

One would expect that prediction improves with a larger training set. For this model, such an improvement is seen even when the training set includes *no* documents in the target language. Note that even the degenerate flat bridge across languages provides useful information. After introducing English data, the model learns to prefer smaller regression parameters (this can be seen as a form of regularization).

Performance is best when a reasonably large corpus is available including some data in the target language. For each bridge, performance improves dramatically, showing that MLSLDA is successfully able to incorporate information learned from both languages to build a single, coherent picture of how sentiment is expressed in both languages. With the GermaNet bridge, performance is better than both the degenerate and dictionary based bridges, showing that the model is sharing information both through the multilingual topics and the regression parameters. Performance on English prediction is comparable to previously published results on this dataset (Blei and McAuliffe, 2007); with enough data, a monolingual model is no longer helped by adding additional multilingual data.

## 5 Relationship to Previous Research

The advantages of MLSLDA reside largely in the assumptions that it makes and does not make: documents need not be parallel, sentiment is a normally distributed document-level property, words are exchangeable, and sentiment can be predicted as a regression on a K-dimensional vector.

By not assuming parallel text, this approach can be applied to a broad class of corpora. Other multilingual topic models require parallel text, either at the document (Ni et al., 2009; Mimno et al., 2009) or word-level (Kim and Khudanpur, 2004; Zhao and Xing, 2006). Similarly, other multilingual sentiment approaches also require parallel text, often supplied via automatic translation; after the translated text is available, either monolingual analysis (Denecke, 2008) or co-training is applied (Wan, 2009). In contrast, our approach requires fewer resources for a language: a dictionary (or similar knowledge structure relating words to nodes in a graph) and comparable text, instead of parallel text or a machine translation system.

Rather than viewing one language through the lens of another language, MLSLDA views all languages through the lens of the topics present in a document. This is a modeling decision with pros and cons. It allows a language agnostic decision about sentiment to be made, but it restricts the expressiveness of the model in terms of sentiment in two ways. First, it throws away information important to sentiment analysis like syntactic constructions (Greene and Resnik, 2009) and document structure (McDonald et al., 2007) that may impact the sentiment rating. Second, a single real number is not always sufficient to capture the nuances of sentiment. Less critically, assuming that sentiment is normally distributed is not true of all real-world corpora; review corpora often have a skew toward positive reviews. We standardize responses by the mean and variance of the training data to partially address this issue, but other response distributions are possible, such as generalized linear models (Blei and McAuliffe, 2007) and vector machines (Zhu et al., 2009), which would allow more traditional classification predictions.

Other probabilistic models for sentiment classification view sentiment as a word level feature. Some models use sentiment word lists, either given or learned from a corpus, as a prior to seed topics so that they attract other sentiment bearing words (Mei et al., 2007; Lin and He, 2009). Other approaches view sentiment or perspective as a perturbation of a log-linear topic model (Lin et al., 2008). Such techniques could be combined with the multilingual approach presented here by using distributions over words that not only bridge different languages but also encode additional information. For example, the vocabulary hierarchies could be structured to encourage topics that encourage correlation among similar sentiment-bearing words (e.g. clustering words associated with price, size, etc.). Future work could also more rigorously validate that the multilingual topics discovered by MLSLDA are sentiment-bearing via human judgments.

In contrast, MLSLDA draws on techniques that view sentiment as a regression problem based on the topics used in a document, as in supervised latent Dirichlet allocation (SLDA) (Blei and McAuliffe, 2007) or in finer-grained parts of a document (Titov and McDonald, 2008). Extending these models to multilingual data would be more straightforward.

## 6 Conclusions

MLSLDA is a "holistic" statistical model for multilingual corpora that does not require parallel text or expensive multilingual resources. It discovers connections across languages that can recover latent structure in parallel corpora, discover sentiment-correlated word lists in multiple languages, and make accurate predictions across languages that improve with more multilingual data, as demonstrated in the context of sentiment analysis.

More generally, MLSLDA provides a formalism that can be used to incorporate the many insights of topic modeling-driven sentiment analysis to multilingual corpora by tying together word distributions across languages. MLSLDA can also contribute to the development of word list-based sentiment systems: the topics discovered by MLSLDA can serve as a first-pass means of sentiment-based word lists for languages that might lack annotated resources.

MLSLDA also can be viewed as a sentiment-informed multilingual word sense disambiguation (WSD) algorithm. When the multilingual bridge is an explicit representation of sense such as WordNet, part

of the generative process is an explicit assignment of every word to sense (the path latent variable $\lambda$); this is discovered during inference. The dictionary-based technique may be viewed as a disambiguation via a transfer dictionary. How sentiment prediction impacts the implicit WSD is left to future work.

Better capturing local syntax and meaningful collocations would also improve the model's ability to predict sentiment and model multilingual topics, as would providing a better mechanism for representing words not included in our bridges. We intend to develop such models as future work.

# 7 Acknowledgments

# References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *ICML*.

Andrea Esuli Stefano Baccianella and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *EMNLP*.

David M. Blei and John D. Lafferty. 2005. Correlated topic models. In *NIPS*.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *NIPS*. MIT Press.

David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3:993–1022.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *UAI*.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *EMNLP*.

Jonathan Chang and David M. Blei. 2009. Relational topic models for document networks. In *AISTATS*.

Shay B. Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *NAACL*.

Noah Constant, Christopher Davis, Christopher Potts, and Florian Schwarz. 2009. The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung*, 33(1–2).

Kerstin Denecke. 2008. Using SentiWordNet for multilingual sentiment analysis. In *ICDEW 2008*.

Paul Denisowski. 1997. CEDICT. http://www.mdbg.net/chindict/.

Jean Diebolt and Eddie H.S. Ip, 1996. *Markov Chain Monte Carlo in Practice*, chapter Stochastic EM: method and application. Chapman and Hall, London.

Alexander Geyken. 2007. The DWDS corpus: A reference corpus for the German language of the 20th century. In *Idioms and Collocations: Corpus-based Linguistic, Lexicographic Studies*. Continuum Press.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *NAACL*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *PNAS*, 101(Suppl 1):5228–5235.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, Columbus, Ohio.

Jan Hefti. 2005. HanDeDict. http://chdw.de.

Hitoshi Isahara, Fransis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *LREC*.

Mark Johnson. 2010. PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In *ACL*.

Woosung Kim and Sanjeev Khudanpur. 2004. Lexical triggers and latent semantic analysis for cross-lingual language model adaptation. *TALIP*, 3(2):94–112.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*. http://www.statmt.org/europarl/.

Claudia Kunze and Lothar Lemnitzer. 2002. Standardizing WordNets in a web-compliant format: The case of GermaNet. In *Workshop on Wordnets Structures and Standardisation*.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *CIKM*.

Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. 2008. A joint topic and perspective model for ideological discourse. In *ECML PKDD*.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. In *Tools and methodologies for teaching*. ACL.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *ACL*.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*.

Ilya Dan Melamed. 1998. *Empirical methods for exploiting parallel texts*. Ph.D. thesis, University of Pennsylvania.

George A. Miller. 1990. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *WWW*.

Noam Ordan and Shuly Wintner. 2007. Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*.

Bo Pang and Lillian Lee. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Martin Porter and Richard Boulton. 1970. Snowball stemmer. http://snowball.tartarus.org/credits.php.

Matthew Purver, Konrad Körding, Thomas L. Griffiths, and Joshua Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *ACL*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *ACL*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453.

Frank Richter. 2008. Dictionary nice grep. http://www-user.tu-chemnitz.de/ fri/ding/.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *NAACL*.

Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *UAI*.

Benoît Sagot and Darja Fišer. 2008. Building a Free French WordNet from Multilingual Resources. In *OntoLex*.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *ACL*.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *SIGHAN Workshop on Chinese Language Processing*.

University of Oxford. 2006. British National Corpus. http://www.natcorp.ox.ac.uk/. http://www.natcorp.ox.ac.uk/.

Tobias Vetter, Manfred Sauer, and Philipp Wallutat. 2000. Filmrezension.de: Online-magazin für filmkritik. http://www.filmrezension.de.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *ACL*.

Chong Wang, David Blei, and Li Fei-Fei. 2009. Simultaneous image classification and annotation. In *CVPR*.

Xing Wei and Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *SIGIR*.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *CIKM*.

Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. thesis, University of Pittsburgh.

Bing Zhao and Eric P. Xing. 2006. BiTAM: Bilingual topic admixture models for word alignment. In *ACL*.

Jun Zhu, Amr Ahmed, and Eric P. Xing. 2009. Medlda: maximum margin supervised topic models for regression and classification. In *ICML*.