Detecting Cross-Cultural Differences Using a Multilingual Topic Model

E.D. Gutiérrez[1] **Ekaterina Shutova**[2] **Patricia Lichtenstein**[3]
Gerard de Melo[4] **Luca Gilardi**[5]
[1] University of California, San Diego
[2] Computer Laboratory, University of Cambridge
[3] University of California, Merced
[4] IIIS, Tsinghua University, [5] ICSI, Berkeley
edg@icsi.berkeley.edu es407@cam.ac.uk tricia1@uchicago.edu
gdm@demelo.org lucag@icsi.berkeley.edu

Abstract

Understanding cross-cultural differences has important implications for world affairs and many aspects of the life of society. Yet, the majority of text-mining methods to date focus on the analysis of monolingual texts. In contrast, we present a statistical model that simultaneously learns a set of common topics from multilingual, non-parallel data and automatically discovers the differences in perspectives on these topics across linguistic communities. We perform a behavioural evaluation of a subset of the differences identified by our model in English and Spanish to investigate their psychological validity.

1 Introduction

Recent years have seen a growing interest in text-mining applications aimed at uncovering public opinions and social trends (Fader et al., 2007; Monroe et al., 2008; Gerrish and Blei, 2011; Pennacchiotti and Popescu, 2011). They rest on the assumption that the language we use is indicative of our underlying worldviews. Research in cognitive and sociolinguistics suggests that linguistic variation across communities systematically reflects differences in their cultural and moral models and goes beyond lexicon and grammar (Kövecses, 2004; Lakoff and Wehling, 2012). Cross-cultural differences manifest themselves in text in a multitude of ways, most prominently through the use of explicit opinion vocabulary with respect to a certain topic (e.g. "policies that *benefit* the poor"), idiomatic and metaphorical language (e.g. "the company is *spinning its wheels*") and other types of figurative language, such as irony or sarcasm.

The connection between language, culture and reasoning remains one of the central research questions in psychology. Thibodeau and Boroditsky (2011) investigated how metaphors affect our decision-making. They presented two groups of human subjects with two different texts about *crime*. In the first text, crime was metaphorically portrayed as a *virus* and in the second as a *beast*. The two groups were then asked a set of questions on how to tackle crime in the city. As a result, while the first group tended to opt for preventive measures (e.g. stronger social policies), the second group converged on punishment- or restraint-oriented measures. According to Thibodeau and Boroditsky, their results demonstrate that metaphors have profound influence on how we conceptualize and act with respect to societal issues. This suggests that in order to gain a full understanding of social trends across populations, one needs to identify subtle but systematic linguistic differences that stem from the groups' cultural backgrounds, expressed both literally and figuratively. Performing such an analysis by hand is labor-intensive and often impractical, particularly in a multilingual setting where expertise in all of the languages of interest may be rare.

With the rise of blogging and social media, NLP techniques have been successfully used for a number of tasks in political science, including automatically estimating the influence of particular politicians in the US senate (Fader et al., 2007), identifying lexical features that differentiate political rhetoric of opposing parties (Monroe et al., 2008), predicting voting patterns of politicians based on their use of language (Gerrish and Blei, 2011), and predicting political affiliation of Twitter users (Pennacchiotti and Popescu, 2011). Fang et al. (2012) addressed

Transactions of the Association for Computational Linguistics, vol. 4, pp. 47–60, 2016. Action Editor: David Chiang.
Submission batch: 11/2015; Published 2/2016.
©2016 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

the problem of automatically detecting and visualising the contrasting perspectives on a set of topics attested in multiple distinct corpora. While successful in their tasks, all of these approaches focused on monolingual data and did not reach beyond literal language. In contrast, we present a method that detects fine-grained cross-cultural differences from multilingual data, where such differences abound, expressed both literally and figuratively. Our method brings together opinion mining and cross-lingual topic modelling techniques for this purpose. Previous approaches to cross-lingual topic modelling (Boyd-Graber and Blei, 2009; Jagarlamudi and Daumé III, 2010) addressed the problem of mining common topics from multilingual corpora. We present a model that learns such common topics, while simultaneously identifying lexical features that are indicative of the underlying differences in perspectives on these topics by speakers of English, Spanish and Russian. These differences are mined from multilingual, non-parallel datasets of Twitter and news data. In contrast to previous work, our model does not merely output a list of monolingual lexical features for manual comparison, but also automatically infers multilingual contrasts.

Our system (1) uses word-document co-occurrence data as input, where the words are labeled as *topic* words or *perspective* words; (2) finds the highest-likelihood dictionary between topic words in the two languages given the co-occurrence data; (3) finds cross-lingual topics specified by distributions over topic-words and perspective-words; and (4) automatically detects differences in perspective-word distributions in the two languages. We perform a behavioural evaluation of a subset of the differences identified by the model and demonstrate their psychological validity. Our data and dictionaries are available from the first author upon request.

2 Related work

View detection. Identifying different viewpoints is related to the well-studied area of subjectivity detection, which aims at exposing opinion, evaluation, and speculation in text (Wiebe et al., 2004) and attributing it to specific people (Awadallah et al., 2011; Abu-Jbara et al., 2012). In our work, we are less interested in explicit local forms of subjectivity, instead aiming at detecting more general contrasts across linguistic communities.

Another line of research has focused on inferring author attributes such as gender, age (Garera and Yarowsky, 2009), location (Jones et al., 2007), or political affiliation (Pennacchiotti and Popescu, 2011). Such studies make use of syntactic style, discourse characteristics, as well as lexical choice. The models used for this are typically binary classifiers trained in a fully supervised fashion. In contrast, in our task, we automatically infer the topic distributions and find topic-specific contrasts.

Probabilistic topic models. Probabilistic topic models have proven useful for a variety of semantic tasks, such as selectional-preference induction (Ó Séaghdha, 2010; Ritter et al., 2010), sentiment analysis (Boyd-Graber and Resnik, 2010) and studying the evolution of concepts and ideas (Hall et al., 2008). The goal of a topic model is to characterize observed data in terms of a much smaller set of unobserved, semantically coherent topics. A particularly popular probabilistic topic model is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Under its assumptions, each document has a unique mix of topics, and each topic is a distribution over terms in the vocabulary. A topic is chosen for every word token according to the topic mix of the document to which it belongs, and then the word's identity is drawn from the corresponding topic's distribution.

Handling multilingual corpora. LDA is designed for monolingual text and thus it lacks the structure necessary to model cross-lingually valid topics. While topic models can be trained individually on two languages and then the acquired topics can be matched, the correspondences between the topics for the two terms will be highly unstable. To address this, Boyd-Graber and Blei (2009) (MUTO) and Jagarlamudi and Daumé III (2010) (JOINTLDA) introduced the notion of cross-lingually valid concepts associated with different terms in different languages, using bilingual dictionaries to model topics across languages. Based on a model by Haghighi et al. (2008), MUTO is capable of learning translations–i.e., matching between terms in the different languages being compared. The Polylingual Topic Model of Mimno et al. (2009) is another approach to finding topics in multilingual corpora, but it requires tuples composed of compa-

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

rable documents in each language of the corpus.

Topic models for view detection. LDA also assumes that the distribution of each topic is fixed across all documents in a corpus. Therefore, a topic associated with, e.g., *war* will have the same distribution over the lexicon regardless of whether the document was taken from a pro-war editorial or an anti-war speech. However, in reality we may expect a single topic to exhibit systematic and predictable variations in its distribution based on authorship.

The cross-collection LDA model by Paul and Girju (2009) addresses this by specifically aiming to expose viewpoint differences across different document collections. Ahmed and Xing (2010) proposed a similar model for detecting ideological differences. Fang et al. (2012)'s Cross-Perspective Topic (CPT) model breaks up the terms in the vocabulary into topic terms and perspective terms with different generative processes, and differentiates between different collections of documents within the corpus. The topic terms are assumed to be generated as in LDA. However, the distribution of perspective terms in a document is taken to be dependent on both the topic mixture of the document as well as the collection from which the document is drawn.

Recent works proposed models for specific types of data. Qiu and Jiang (2013) use user identities and interactions in threaded discussions, while Gottipati et al. (2013) developed a topic model for Debatepedia, a semi-structured resource in which arguments are explicitly enumerated. However, all of these models perform their analyses on monolingual datasets. Thus, they are useful for comparing different ideologies expressed in the same language, but not for cross-linguistic comparisons.

3 Method

The goal of our model is to analyse large, non-parallel, multilingual corpora and present cross-lingually valid topics and the associated perspectives, automatically inferring the differences in conceptualization of these topics across cultures. Following Boyd-Graber and Blei (2009) and Jagarlamudi and Daumé III (2010), our distributions of latent topics range over latent, cross-lingual *topic concepts* that manifest themselves as language-specific *topic words*. We use bilingual dictionaries, containing words in one language and their translations in another language, to represent the topic concepts. These are represented as a bipartite graph, with each translation entry being an edge and each topic word in the two languages being a vertex. While the topic words are tied together by the translation dictionary, the perspective words can vary freely across languages. Following Fang et al. (2012), we treat nouns as topic words and verbs and adjectives as perspective words[1]. The model assumes that adjective and verb tokens in each document are assigned to topics in proportion to the topic assignments of the topic word tokens. Then, the perspective term for this topic is drawn depending on the topic assignment and the language of the speaker.

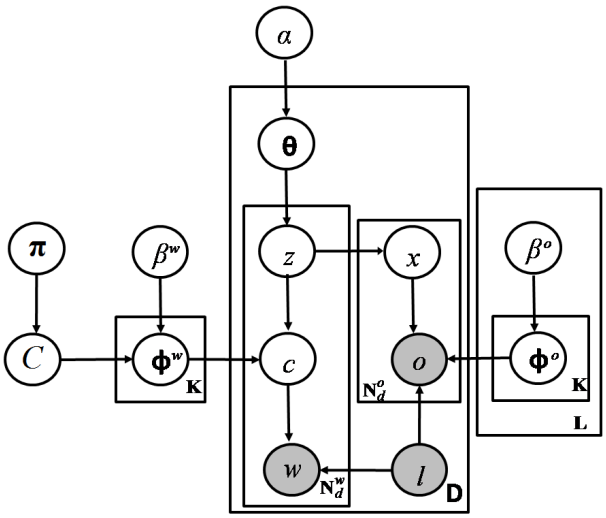

Figure 1: Basic generative model.

3.1 Basic Generative Model

Given the languages $\ell \in \{a, b\}$, our model infers the distributions of multi-lingual topics and language-specific perspective-words (Fig. 2), as follows:

1. Draw a set C of concepts (u, v) matching topic word u from language a to topic word v from language b, where the probability of concept (u, v) is proportional to a prior $\pi_{u,v}$ (e.g. based on information from a translation dictionary).

2. Draw multinomial distributions:

[1]This approximation was adopted for convenience, computational efficiency and ease of interpretation. However, in principle our method does not depend on it, since it can be applied with all content words as topic or perspective words.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

- For topic indices $k \in \{1, ..., K\}$, draw language-independent topic-concept distributions $\phi_k^w \sim \text{Dir}(\beta^w)$ over pairs $(w_a, w_b) \in C$.
- For topic indices $k \in \{1, ..., K\}$ and languages $\ell \in \{a, b\}$, draw language-specific perspective-term distributions $\phi_k^{\ell,o} \sim \text{Dir}(\beta^o)$ over perspective-terms in language ℓ.

3. For each document $d \in \{1, ..., D\}$ with lang. ℓ_d:
   - Draw topic weights $\theta_d \sim \text{Dir}(\alpha)$
   - For each topic-word index $i \in \{1, ..., N_d^w\}$ of document d:
     - Draw topic $z_i \sim \theta_d$
     - Draw topic concept $c_i = (w_a, w_b) \sim \phi_{z_i}^w$, and select w_{ℓ_d} as the member of that pair corresponding to language ℓ_d.
   - For each perspective-word index $j \in \{1, ..., N_d^o\}$ of document d:
     - Draw topic $x_j \sim \text{Uniform}(z_{w_1}, ..., z_{w_{N_d^o}})$
     - Draw perspective-word $o_j \sim \phi_{x_j}^{\ell,o}$

3.2 Model Variants

We have experimented with several variants of our model, in order to account for the translation of polysemous words, adapt the translation model to the corpus used, and to handle words for which no translation is found.

a) SINGLE variants of the model match each topic term in a language with at most one topic term in the other language.

   MULTIPLE variants allow each term to match to multiple other words in the other language.

b) INFER variants allow higher-likelihood matchings to be inferred from the data.

   STATIC variants treat the matchings as fixed, which is equivalent to assigning a probability of 0 or 1 to every edge in our bipartite graph C.

c) RELEGATE variants relegate all unmatched words in each language to a single separate background topic distinct from the topics that are learned for the matched topic words. This is akin to forcing the probability for currently unmatched words to 0 in all topics except for one, and forcing the probability of all currently matched words to 0 in this topic.

   INCLUDE variants do not restrict the assignment unmatched words; they are assigned to the same set of topics as the matched words.

We test the following six variants: SINGLESTATICRELEGATE, SINGLESTATICINCLUDE, SINGLEINFERRELEGATE, SINGLEINFERINCLUDE, MULTIPLESTATICRELEGATE, and MULTIPLESTATICINCLUDE. We do not test MULTIPLEINFER variants because of the complexity of inferring a multiple matching in a bipartite graph.

3.3 Learning & Inference

For all variants, a collapsed Gibbs sampler can be used to infer topics $\phi^{\ell,o}$ and ϕ^w, per-document topic distributions θ, as well as topic assignments $\mathbf{z}$ and $\mathbf{x}$. This corresponds to the S-step below. For INFER variants, we follow Boyd-Graber and Blei in using an M-step involving a bipartite graph matching algorithm to infer the matching $\mathbf{m}$ that maximizes the posterior likelihood of the matching.

S-Step: Sample topics for words in the corpus using a collapsed Gibbs sampler. For topic-word $w_i = u$ belonging to document d, if the word occurs in concept $c_i = (u, v)$, then sample the topic and entry according to:

$$p(z_i = k, c_i = (u, v) \mid w_i = u, \mathbf{z}_{-i}, C) \propto \frac{N_{dk} + \alpha_k}{\sum_j (N_{dj} + \alpha_j)} \times \frac{N_{k(u,v)} + \beta_k^w}{\sum_{v'} \left(N_{k(u,v')} + \beta_k^w\right)}$$

where the sum in the denominator of the first term is over all topics, and in the second term is over all words matched to u. N_{dk} is the count of topic-words of topic k in document d, $N_{k(u,v)}$ is the count of topic-words either of type u or of type v assigned to topic k in all the corpora.[2] For perspective-word $o_i = n$, sample the topic according to:

$$p(z_i = k | o_i = n, \mathbf{z}_{-i}, C) \propto \frac{N_{dk}}{\sum_j N_{dj}} \times \frac{N_{kv}^{\ell_d} + \beta_k^o}{\sum_m \left(N_{km}^{\ell_d} + \beta_k^o\right)}$$

[2] In RELEGATE variants, for u unmatched z_i is sampled as: $p(z_i = k | w_i = u, \mathbf{z}_{-i}, C) \propto \frac{N_{dk} + \alpha_k}{\sum_k (N_{dk} + \alpha_k)}$, which can be seen as $\beta_{u\cdot}^w \to \infty$ for unmatched terms.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

where the sum in the second term of the denominator is over the perspective-word vocabulary of language ℓ_d; N_{dk} is the count of *topic* words in document d with topic k; and $N_{km}^{\ell_d}$ is the count of perspective-word m being assigned topic k in language ℓ_d. Note that in all the counts above, the current word token i is omitted from the count.

Given our sampling assignments, we can then estimate θ^d, $\phi^{\ell,o}$, and ϕ^w as follows:

$$\hat{\theta}_{kd} = \frac{N_{dk} + \alpha_k}{\sum_k (N_{dk} + \alpha_k)},$$

$$\hat{\phi}^w_{k(u,v)} = \frac{N_{k(u,v)} + \beta^w_{(u,v)}}{\sum_{v'} \left(N_{k(u,v')} + \beta^w_{(u,v')}\right)},$$

$$\hat{\phi}^{\ell,o}_{nk} = \frac{N_{kn} + \beta^o_n}{\sum_m \left(N^\ell_{km} + \beta^o_n\right)}.$$

M-Step: *(for* INFER *variants only)*: Run the Jonker-Volgenant (Jonker and Volgenant, 1987) bipartite matching algorithm to find the optimal matching C given some weights. For topic-term u from language a and topic-term v from language b, our weights correspond to the log of the posterior odds that the occurrences of u and v come from a matched topic distribution, as opposed to coming from unmatched distributions:

$$\mu_{u,v} = \sum_{k \setminus \{a*, b*\}} \left(N_{k(u,v)} \log \hat{\phi}^w_{k(u,v)}\right) - N_u \log \hat{\phi}^w_{k(u,\cdot)} - N_v \log \hat{\phi}^w_{k(\cdot,v)} + \pi_{u,v},$$

where N_u is the count of topic-term u in the corpus. This expression can also be interpreted as a kind of pointwise mutual information (Haghighi et al., 2008). The Jonker-Volgenant algorithm has time complexity of at most $O(V^3)$, where V is the size of the lexicon (Jonker and Volgenant, 1987).

3.4 Inference of Perspective-Word Contrasts

Having learned our model and inferred how likely perspective-terms are for a topic in a given language, we seek to know whether these perspectives differ significantly in the two languages. More precisely, can we infer whether word m in language a and the equivalent word n in language b have significantly different distributions under a topic k? To do this, we make the assumption that the perspective-words in languages a and b are in one-to-one correspondence to each other. Recall that, for a given topic k and language ℓ, N^ℓ_{km} is the count for term m and $\phi^{\ell,o}_{k,m}$ is the probability for word m in language ℓ. Just as we collect the probabilities into word-topic distribution vectors $\phi^{\ell,o}_k$, we collect the counts into word-topic count vectors $[N^\ell_{k1}, N^\ell_{k2}, ..]$. Then, since our model assumes a prior over the parameter vectors $\phi^{\ell,o}_k$, we can infer the likelihood for that observed word-topic counts N^a_{km} and N^b_{kn} were drawn from a single word-topic-distribution prior denoted by $\breve{\phi} := \phi^{a,o}_{km} = \phi^{b,o}_{kn}$. Below all our probabilities are conditioned implicitly on this event as well as on N^a_k and N^b_k being fixed.

Denote the total count of word tokens in topic k from language ℓ by $N^\ell_k = \sum_m N^\ell_{km}$. Now, we derive the probability that we observe a ratio greater than δ between the proportion of words in topic k that belong to word type m in language a and to corresponding word type n in language b:

$$p\left(\frac{N^a_{km}}{N^a_k}\frac{N^b_k}{N^b_{kn}} \geq \delta\right) + p\left(\frac{N^b_{kn}}{N^b_k}\frac{N^a_k}{N^a_{km}} \geq \delta\right) \quad (1)$$

By symmetry, it suffices to derive an expression for the first term. We note that the inequality in the probability is equivalent to a sum over a range of values of N^a_{km} and N^b_{kn}. By rearranging terms, applying the law of conditional probability to condition on the term $\breve{\phi}$, and exploiting the conditional independence of N^a_{km} and N^b_{km} given $\breve{\phi}$, N^a_k, and N^b_k, we can rewrite this first term as

$$\sum_{x=0}^{N^b_k} \sum_{y=x\delta N^{a/b}}^{N^a_k} \int p(N^b_{kn} = x|\breve{\phi}) p(N^a_{km} = y|\breve{\phi}) p(\breve{\phi}) d\breve{\phi},$$

where $N^{a/b} = \frac{N^a_k}{N^b_k}$. Recall that $\phi^{\ell,o}_k \sim \text{Dir}(\beta^o)$ under our model. Assume a symmetric Dirichlet distribution for simplicity. It can then be shown that the marginal distribution of $\breve{\phi}$ is $\breve{\phi} \sim \text{Beta}(\beta^o, (V-1)\beta^o)$, where V is the total size of the perspective-word vocabulary. Similarly, it can be shown that the marginal distribution of N^ℓ_{km} given $\phi^{\ell,o}_k$ is $N^\ell_{km} \sim \text{Binom}(N^\ell_k, \phi^{\ell,o}_i)$ for $\ell \in \{a, b\}$. Therefore, the integrand above is proportional to the beta-binomial distribution with number of trials $N^a_k + N^b_k$, successes $x + y$, and parameters β^o and $(V-1)\beta^o$, but with partition function $\binom{N^a_k}{y}\binom{N^b_k}{x}$. Denote the PMF of this

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

distribution by $f(N_k^a + N_k^b, x+y, \beta^o)$. Then expression (1) above becomes:

$$\sum_{x=0}^{N_k^b} \sum_{y=x\delta N^{a/b}}^{N_k^a} f(N_k^a + N_k^b, x+y, \beta^o) + \sum_{x=0}^{N_k^a} \sum_{y=x\delta N^{b/a}}^{N_k^b} f(N_k^a + N_k^b, x+y, \beta^o). \quad (2)$$

We cannot observe N_{kb}^a, N_{kn}^b, N_k^a and N_k^b explicitly, but we can estimate them by obtaining posterior samples from our Gibbs sampler. We substitute these estimates into expression (2).

4 Experiments

4.1 Data

Twitter Data. We gathered Twitter data in English, Spanish and Russian during the first two weeks of December 2013 using the Twitter API. Following previous work (Puniyani et al., 2010), we treated each Twitter user account as a document. We then tagged each document for part-of-speech, and divided the word tokens in it into topic-words and perspective-words. We constructed a lexicon of 2,000 topic terms and 1,500 perspective-terms for each language by filtering out any terms that occurred in more than 10% of the documents in that language, and then selecting the remaining terms with the highest frequency. Finally, we kept only documents that contained 4 or more topic words from our lexicon. This left us with 847,560 documents in English (4,742,868 topic-word and 1,907,685 perspective-word tokens); 756,036 documents in Spanish (4,409,888 topic-word and 1,668,803 perspective-word tokens); and 260,981 documents in Russian (1,621,571 topic-word and 981,561 perspective-word tokens).

News Data. We gathered all the articles published online during the year 2013 by the state-run media agencies of the United States (Voice of America or "VOA"–English), Russia (RIA Novosti or "RIA"–Russian), and Venezuela (Agencia Venezolana de Noticias or "AVN"–Spanish). These three news agencies were chosen because they not only provide media in three distinct languages, but they are guided by the political world-views of three distinct governments. We treated each news article as a document, and removed duplicates. Once again, we constructed a lexicon of 2,000 topic terms and 1,500 perspective-terms using the same criteria as for Twitter, and kept only documents that contained 4 or more topic words from our lexicon. This left us with 23,159 articles (10,410,949 tokens) from VOA, 41,116 articles (11,726,637 tokens) from RIA, and 8,541 articles (2,606,796 tokens) from AVN.

Dictionaries. To create the translation dictionaries, we extracted translations from the English, Spanish, and Russian editions of Wiktionary, both from the translation sections and the gloss sections if the latter contained single words as glosses. Multi-word expressions were universally removed. We added inverse translations for every original translation. From the resulting collection of translations, we then created separate translation dictionaries for each language and part-of-speech tag combination.

In order to give preference to more important translations, we assigned each translation an initial weight of $1 + \frac{1}{r}$, where r was the rank of the translation within the page. Since a translation (or its inverse) can occur on multiple pages, we aggregated these initial weights and then assigned final weights of $1 + \frac{1}{r'}$, where r' was the rank after aggregation and sorting in descending order of weights.

4.2 Experimental Conditions

To evaluate the different variants of our model, we held out 30,000 documents (test set) during training. We plugged in the estimates of ϕ^w and C acquired during training using the rest of the corpus to produce a likelihood estimate for these held-out documents. All models were initialized with the prior matching determined by the dictionary data. For each number of topics K, we set α to $50/K$and the β variables to 0.02, as in Fang et al. (2012). For the MULTIPLE variants, we set $\pi_{i,j} = 1$ if i and j share an entry and 0 otherwise. For INFER variants, only three M-steps were performed to avoid overfitting, at 250, 500, and 750 iterations of Gibbs sampling, following the procedure in Boyd-Graber and Blei (2009).

4.3 Comparison of model variants

In order to compare the variants of our model, we computed the perplexity and coherence for

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

each variant on TWITTER and NEWS, for English–Spanish and English–Russian language pairs.

Perplexity is a measure of how well a model trained on a training set predicts the co-occurrence of words on an unseen test set $\mathcal{H}$. Lower perplexity indicates better model fit. We evaluate the held-out perplexity for topic words w_i and perspective-words o_i separately. For topic words, the perplexity is defined as $exp(-\sum_{w_i \in \mathcal{H}} logp(w_i)/N^w)$. As for standard LDA, exact inference of $p(w_i)$ is intractable under this model. Therefore we adapted the estimator developed by Murray and Salakhutdinov (2009) to our models.

Coherence is a measure inspired by pointwise mutual information (Newman et al., 2010). Let $D(v)$ be the the number of documents with at least one token of type v and let $D(v, w)$ be the number of documents containing at least one token of type v and at least one token of type w. Then Mimno et al. (2011) define the coherence of topic k as

$$\frac{1}{\binom{M}{2}} \sum_{m=2}^{M} \sum_{\ell=1}^{m-1} \log \frac{D(v_m^{(k)}, v_\ell^{(k)}) + \epsilon}{D(v_\ell^{(k)})},$$

where $V^{(k)} = (v_1^{(k)}, ..., v_M^{(k)})$ is a list of the M most probable words in topic k and ϵ is a small smoothing constant used to avoid taking the logarithm of zero. Mimno et al. (2011) find that coherence correlates better with human judgments than do likelihood-based measures. Coherence is topic-specific measure, so for each model variant we trained, we computed the median topic coherence across all the topics learned by the model. We set $\epsilon = 0.1$.

Model performance and analysis. Fig. 2 shows perplexity for the variants as a function of the number of iterations of Gibbs sampling on the English-Spanish NEWS corpus. The figure confirms that 1000 iterations of Gibbs sampling on the NEWS corpus was sufficient for convergence across model variants. We omit figures for English-Russian and for the TWITTER corpus, since the patterns were nearly identical. Figure 3 shows how perplexity varies as a function of the number of topics. We used this information to choose optimal models for the different corpora. The optimal number of topics was $K = 175$ for the English-Spanish NEWS corpus, $K = 200$ for the English-Russian NEWS, $K = 325$ for the English-Spanish TWITTER, and $K = 300$ for the English-Russian TWITTER. Although the optimal number of topics varied across corpora, the relative performance of the different models was the same. In all of our corpora, the MULTIPLE variants provided better fits than their corresponding SINGLE variants. There are several explanations for this. For one, the MULTIPLE variants are able to exploit the information from multiple translations, unlike the SINGLE variants, which discarded all but one translation per word. For another, the matchings produced by the SINGLEINFER variants can be purely coincidental and the result of overfitting (see some examples below). INCLUDE variants performed markedly better than RELEGATE variants. INFER variants improved model fit compared to STATIC variants, but required more topics to produce optimal fit.

Recall that we performed an M-step in the INFER variants 3 times, at 250, 500, and 750 iterations. As noted in §3.3, the M-step in the INFER variants maximizes the posterior likelihood of the matching. However, Fig. 2 shows that this maximization causes held-out perplexity to increase substantially just after the first matching M-step, around 250 iterations, before decreasing again after about 50 more iterations of Gibbs sampling. We believe that this happens because the M-step is maximizing over expectations that are approximate, since they are estimated using Gibbs sampling. If the sampler has not yet converged, then the M-step's maximization will be unstable. We found support for this explanation when we re-ran the INFER variants using 1000 iterations between M-steps, giving the Markov chain enough time to converge. After this change, perplexity went down immediately after the M-step and kept decreasing monotonically, rather than increasing after the M-step before decreasing. However, this did not result in a significantly lower final perplexity or coherence and thus did not change the relative performance of the models. In addition, Fig. 2 suggests that the second and third M-steps (at 500 and 750 iterations, respectively) had little effect on perplexity. In light of the high computational expense of each inference step, this suggests in practice a single inference step may be sufficient.

Fig. 4 shows that the MULTIPLESTATICINCLUDE variant was also the superior model as measured by

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

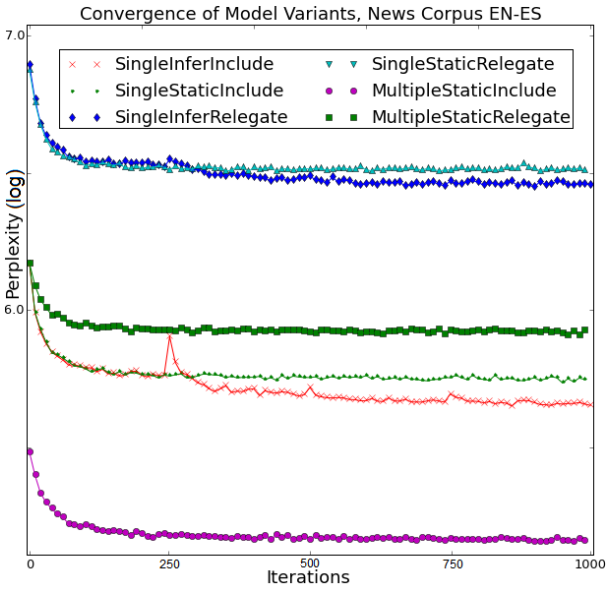

Figure 2: Perplexity of different model variants for different numbers of iterations at K=175.

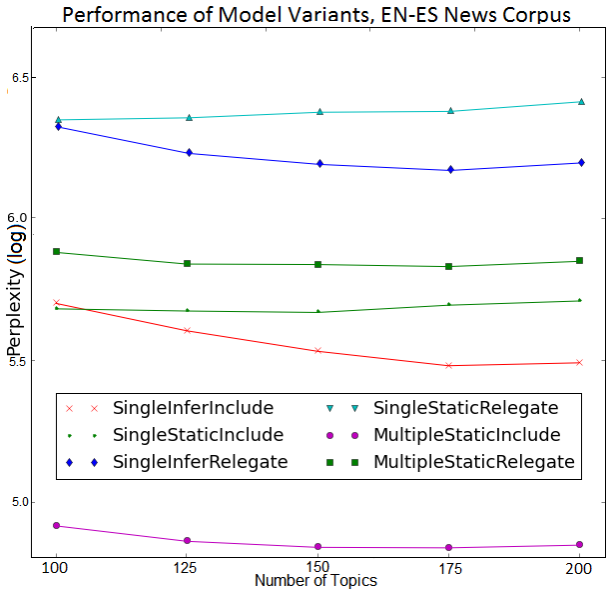

Figure 3: Perplexity of different model variants.

median topic coherence. Once again, this general pattern held true for the English-Russian pair and TWITTER corpora. Overall, the results show that MULTIPLESTATICINCLUDE provides superior performance across measures, corpora, topic numbers, and languages. We therefore used this variant in further data analysis and evaluation. Incidentally, the observed decrease in topic coherence as K increases is expected, because as K increases, lower-likelihood topics tend to be more incoherent (Mimno et al., 2011). Experiments by Stevens et al. (2012) show that this effect is observed for LDA-, NMF-, and SVD-based topic models.

Cross-linguistic matchings. The matchings inferred by the SINGLEINFERINCLUDE variant were of mixed quality. Some of the matchings corrected low-quality translations in the original dictionary. For instance, our prior dictionary matched *passage* in English to *pasaje* in Spanish. Though technically correct, the dominant meaning of *pasaje* is *[travel] ticket*. The TWITTER model correctly matched *passage* to *ruta* instead. Many of the matchings learned by the model did not provide technically correct translations, yet were still revelatory and interesting. For instance, the dictionary translated the Spanish word *pito* as *cigarette* in English. However, in informal usage this word refers specifically to cannabis cigarettes, not tobacco cigarettes. The TWITTER model matches *pito* to the English slang word *weed* instead. The Spanish word *Siria* (Syria) was unmatched in the prior dictionary; the NEWS model matched it to the word *chemical*, which makes sense in the context of extensive reporting of the usage of chemical weapons in the ongoing Syrian conflict.

4.4 Data analysis and discussion

We have conducted a qualitative analysis of the topics, perspectives and contrasts produced by our models for English–Spanish and English–Russian, TWITTER and NEWS datasets. While the topics were coherent and consistent across languages, sets of perspective words manifested systematic differences revealing interesting cross-cultural contrasts. Fig. 5 and 7 show the top perspective words discovered by the model for the topic of *finance* and *economy* in English and Spanish NEWS and TWITTER corpora, respectively. While some of the perspective words are neutral, mostly literal and occur in both English and Spanish (e.g. *balance* or *authorize*), many others represent metaphorical vocabulary (e.g. *saddle, gut, evaporate* in English, or *incendiar, sangrar, abatir* in Spanish) pointing at distinct models of conceptualization of the topic. When we applied the contrast detection method (described in §3.4) to these perspective words, it highlighted the differences in metaphorical perspectives, rather than the literal ones, as shown in Fig. 6 and 8. En-

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

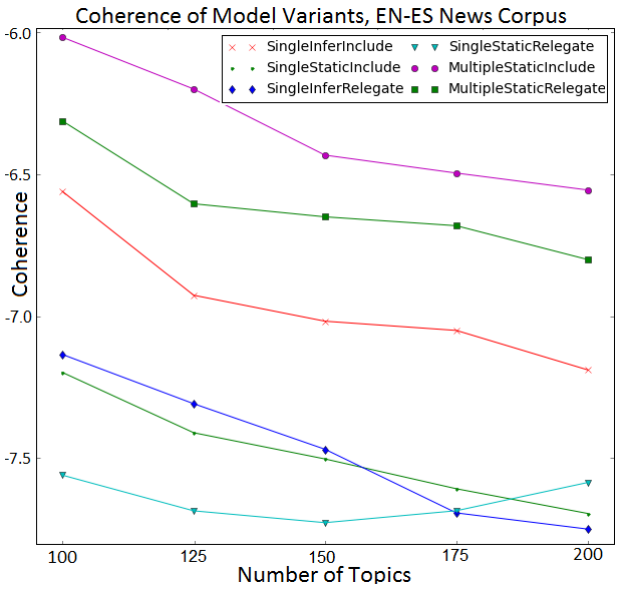

Figure 4: Coherence of different model variants.

glish speakers tend to discuss economic and financial processes using motion terms, such as "*slow, drive, boost* or *sluggish*", or a related metaphor of *horse-riding*, e.g. "*rein* in debt", "*saddle* with debt", or even "*breed* money". In contrast, Spanish speakers tend to talk about the economy in terms of size rather than motion, using verbs such as *ampliar* or *disminuir*, and other metaphors, such as *sangrar* (to bleed) and *incendiar* (to light up). These examples demonstrate coherent conceptualization patterns that differ in the two languages. Interestingly, this difference manifested itself in both NEWS and TWITTER corpora and echoes the findings of a previous corpus-linguistic study of Charteris-Black and Ennis (2001), who manually analysed metaphors used in English and Spanish financial discourse and reported that motion and navigation metaphors that abound in English were rarely observed in Spanish.

For the majority of the topics we analysed the model revealed interesting cross-cultural differences. For instance, the Spanish corpora exhibited metaphors of *battle* when talking about *poverty* (with *poverty* seen as an enemy), while in the English corpus *poverty* was discussed more neutrally as a social problem that needs a practical solution. English-Russian NEWS experiments revealed a surprising difference with respect to the topic of *protests*. They suggested that while US media tend to use stronger metaphorical vocabulary, such as

Topic_EN budget debt deficit reduction spend balance cut increase limit downtown tax stress addition planet

Topic_ES presupuesto deficit deuda reduccion equilibrio disminucion gasto aumentacion tasa sacerdote

Perspective_EN balance default triple *rein* accumulate accrue *trim* incur *saddle slash* prioritize avert *gut* burden *evaporate* borrow pile *cap cut tackle*

Perspective_ES renegociar mejora etiquetado *desplomar recortar* endeudar *incendiar* destinar asignar autorizar aprobado ascender *sangrar* augurar *abatir*

Figure 5: Top perspectives in system output for the topic of *finance* in the NEWS corpus (metaphors in red italics).

Contrasts_EN: rein [in debt], saddle [with debt], cap [debt], breed [money], gut [budget], [debt] hit, tackle [debt], boost, slow, drive, sluggish [economy], spur

Contrasts_ES: sangrar [dinero], ampliar, disminuir [la economía], superar [la tasa], emitir [deuda]

Figure 6: Contrasts identified by the model in NEWS.

clash, erupt or *fire*, in Russian protests are discussed more neutrally. Generally, the NEWS corpora contained more abstract topics and richer information about conceptual structure and sentiment in all languages. Many of the topics discovered in TWITTER related to everyday concepts, such as *pets* or *concerts*, with fewer topics covering societal issues. Yet, a few TWITTER-specific contrasts could be observed: e.g., the *sports* topic tends to be discussed using *war* and *battle* vocabulary in Russian to a greater extent than in English.

Our models tend to identify two general kinds of differences: (1) cross-corpus differences representing world views of particular populations whom the corpora characterize (such differences exist both across and within languages, e.g. the metaphors used in the progressive *New York Times* would be different from the ones in the more conservative *Wall Street Journal*); and (2) deeply entrenched cross-linguistic differences, such as the *motion* versus *expansion* metaphors for the economy in English and Spanish. Such systematic cross-linguistic contrasts can be associated with contrastive behavioural patterns across the different linguistic communities (Casasanto and Boroditsky, 2008; Fuhrman et al., 2011). In both NEWS and TWITTER data, our model effectively identifies and summarises such contrasts simplifying the manual analysis of the data

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

Topic_EN economy growth rate percent bank economist interest reserve market policy

Topic_ES economía crecimiento tasa banco poltica mercado interés inflacin empleo economista

Perspective_EN economic financial grow global expect remain *cut boost* low *slow drive*

Perspective_ES económico mundial agregar financiero informal *pequeño* significar interno *bajar*

Figure 7: Top perspectives in system output for the *economy* topic in TWITTER (metaphors in red).

Contrasts_EN: slow [the economy], push [the economy], strong [economy], weak [economy], stable [economy], boost [the economy]

Contrasts_ES: caer [la economía], disminuir, superar [la economía], ampliar [el crecimiento]

Figure 8: Contrasts identified by the model in TWITTER.

by highlighting linguistic trends that are indicative of the underlying conceptual differences. However, the conceptual differences are not straightforward to evaluate based on the surface vocabulary alone. In order to investigate this further, we conducted a behavioural experiment testing a subset of the contrasts discovered by our model.

5 Behavioural evaluation

We assessed the relevance of the contrasts through an experimental study with native English-speaking and native Spanish-speaking human subjects. We focused on a linguistic difference in the metaphors used by English speakers versus Spanish speakers when discussing changes in a nation's economy. While English speakers tend to use metaphors involving both locative motion verbs (e.g. *slow*) as well as expansive/contractive motion verbs (e.g. *shrink*), Spanish speakers preferentially employ expansive/contractive motion verbs (e.g. *disminuir*) to describe changes in the economy. These differences could reflect linguistic artefacts (such as collocation frequencies) or could reflect entrenched conceptual differences. Our experiment addresses the question of whether such patterns of behaviour arise cross-linguistically in response to non-linguistic stimuli. If the linguistic differences are indicative of entrenched conceptual differences, then we expect to see responses to the non-linguistic stimuli that correspond to the usage differences in the two languages.

5.1 Experimental setup

We recruited 60 participants from one English-speaking country (the US) and 60 participants from three Spanish-speaking countries (Chile, Mexico, and Spain) using the CrowdFlower crowdsourcing platform. Participants first read a brief description of the experimental task, which introduced them to a fictional country in which economists are devising a simple but effective graphic for "representing change in [the] economy". They then completed a demographic questionnaire including information about their native language. Results from 9 US and 3 non-US participants were discarded for failure to meet the language requirement.

Participants navigated to a new page to complete the experimental task. Stimuli were presented in a 1200×700-pixel frame. The center of the frame contained a sphere with a 64-pixel diameter. For each trial, participants clicked on a button to activate an animation of the sphere which involved (1) a positive displacement (in rightward pixels) of 10% or 20%, or a negative displacement (in leftward pixels) of 10% or 20%;[3] and, (2) an expansion (in increased pixel diameter) of 10% or 20%, or a contraction (in decreased pixel diameter) of 10% or 20%.[4]

Participants saw each of the resulting conditions 3 times. The displacement and size conditions were drawn from a random permutation of 16 conditions using a Fisher-Yates shuffle (Fisher and Yates, 1963). Crucially, half of the stimuli contained conflicts of information with respect to the size and displacement metaphors for economic change (e.g. the sphere could both grow and move to the left). Overall we expected the Spanish speakers' responses to be more closely associated with changes in diameter due to the presence and salience of the size metaphor, and the English speakers' responses to be influenced by both conditions. We expected these differences to be most prominent in the con-

[3]The use of leftward/rightward horizontal displacement to represent decreases/increases in magnitude is supported by research in numerical cognition showing that people associate smaller magnitudes with the left side of space and larger magnitudes with the right side (Dehaene, 1992; Fias et al., 1995).

[4]A demonstration of the English experimental interface can be accessed at `http://goo.gl/W3YVfC`. The Spanish interface is identical, but for a direct translation of the guidelines provided by a native Spanish/fluent English speaker.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

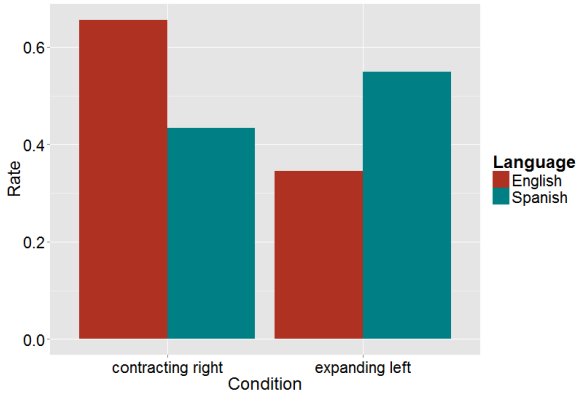

Figure 9: "Economy Improved" response rate in conflicting stimulus conditions.

flicting trials, which force English speakers (unlike Spanish speakers) to choose between two available metaphors. We focus on these conflicting trials in our analysis and discussion of the results.

5.2 Results

In trials in which stimuli moving rightward were simultaneously contracting, English speakers responded that the economy improved 66% of the time, whereas Spanish speakers judged the economy to have improved 43% of the time. In trials in which stimuli moving leftward were simultaneously expanding, English speakers judged the economy to have improved 34% of the time, and Spanish speakers responded that the economy improved 55% of the time. The results are illustrated in Figure 9.

These results indicate three effects: (1) English speakers exhibit a pronounced bias for using horizontal displacement rather than expansion/contraction during the decision-making process; (2) Spanish speakers are more biased toward expansion/contraction in formulating a decision; and, (3) across the two languages the responses show contrasting patterns. The results support our expectation on the relevance of different metaphors when reasoning about the economy by the English and Spanish speakers.

To examine the significance of these effects, we fit a binary logit mixed effects model[5] to the data. The full analysis modeled judgment with native language, displacement, and size as fully crossed fixed effects and participant as a random effect. This analysis confirmed that native language was associated with judgments about economic change. In particular, it indicated that changes in size affected English speakers' judgments and Spanish speakers' judgments differently ($p < 0.001$), with an increase in size increasing the odds ($e^{\beta} = 2.5$) of a judgment of IMPROVED by Spanish speakers and decreasing the odds ($e^{\beta} = 0.44$) of a judgment of IMPROVED by English speakers. A Type II Wald test revealed the interaction between language and size to be highly statistically significant ($\chi^2(1) < 0.001$).

In summary, the patterns we see in the behavioural data are consistent with the patterns uncovered in the output of our model. While much territory remains to be investigated to delimit the nature of this relationship, our results represent a first step toward establishing an association between information mined from large textual data collections and information observed through behavioural responses on a human scale.

6 Conclusion

We presented the first model that detects common topics from multilingual, non-parallel data and automatically uncovers differences in perspectives on these topics across linguistic communities. Our data analysis and behavioural evaluation offer evidence of a symbiotic relationship between ecologically sound corpus experiments and scientifically controlled human subject experiments, paving the way for the use of large-scale text mining to inform cognitive linguistics and psychology research.

We believe that our model represents a good foundation for future projects in this area. A promising area for further work is in developing better methods for identifying contrasts in perspective terms. This could perhaps involve modifying the generative process for perspective terms or incorporating syntactic dependency information. It would also be interesting to investigate the effect of dictionary quality and corpus size on the relative performance of STATIC and INFER variants. Finally, we note that the model can be applied to identify contrastive perspectives in monolingual as well as multilingual data, providing a general tool for the analysis of subtle, yet important, cross-population differences.

[5] See Fox and Weisberg (2011) for a discussion of such models including application of the Type II Wald test.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

Acknowledgments

We would like to thank the anonymous reviewers as well as the TACL editors, Sharon Goldwater and David Chiang, for helpful comments on an earlier draft of this paper. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. Ekaterina Shutova's research is supported by the Leverhulme Trust Early Career Fellowship. Gerard de Melo's research is supported by China 973 Program Grants 2011CBA00300, 2011CBA00301, and NSFC Grants 61033001, 61361136003, 61550110504.

References

Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 399–409, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amr Ahmed and Eric P. Xing. 2010. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1140–1150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2011. OpinioNetIt: Understanding the Opinions-People network for politically controversial topics. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2481–2484, New York, NY, USA. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09)*, pages 75–82. Arlington, VA, USA: AUAI Press.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: multilingual supervised latent Dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 45–55.

Daniel Casasanto and Lera Boroditsky. 2008. Time in the mind: Using space to think about time. *Cognition*, 106(2):579–593.

Jonathan Charteris-Black and Timothy Ennis. 2001. A comparative study of metaphor in Spanish and English financial reporting. *English for Specific Purposes*, 20:249–266.

Stanislas Dehaene. 1992. Varieties of numerical abilities. *Cognition*, 44:1–42.

Anthony Fader, Dragomir Radev, Burt L. Monroe, and Kevin M. Quinn. 2007. MavenRank: Identifying influential members of the US senate using lexical centrality. In *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 658–666.

Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. 2012. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*, pages 63–72, New York. New York: ACM.

Wim Fias, Marc Brysbaert, Frank Geypens, and Géry d'Ydewalle. 1995. The importance of magnitude information in numerical processing: evidence from the SNARC effect. *Mathematical Cognition*, 2(1):95–110.

Ronald A. Fisher and Frank Yates. 1963. *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver and Boyd, Edinburgh.

John Fox and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. SAGE Publications, CA: Los Angeles.

Orly Fuhrman, Kelly McCormick, Eva Chen, Heidi Jiang, Dingfang Shu, Shuaimei Mao, and Lera Boroditsky. 2011. How linguistic and cultural forces shape conceptions of time: English and Mandarin time in 3D. *Cognitive Science*, 35:1305–1328.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 710–718, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sean M. Gerrish and David M. Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of ICML*.

Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. 2013. Learning topics and positions from Debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1858–1868, Seattle,

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

Washington, USA, October. Association for Computational Linguistics.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, ACL-'08:HLT, pages 771–779, Columbus, Ohio, USA.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language processing*, pages 363–371. Association for Computational Linguistics.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, and Suzanne Little, editors, *Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR'2010)*, pages 444–456. Springer-Verlag, Berlin.

Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2007. "I know what you did last summer": Query logs and user privacy. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 909–914, New York, NY, USA. ACM.

Roy Jonker and Anton Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Zoltán Kövecses. 2004. Introduction: Cultural variation in metaphor. *European Journal of English Studies*, 8:263–274.

George Lakoff and Elisabeth Wehling. 2012. *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic*. Free Press, New York.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, pages 880–889. Association for Computational Linguistics.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Iain Murray and Ruslan R. Salakhutdinov. 2009. Evaluating probabilities under high-dimensional latent variable models. In *Advances in Neural Information Processing Systems*, pages 1137–1144.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala, Sweden. Association for Computational Linguistics.

Michael Paul and Roxana Girju. 2009. 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: Volume 3 - Volume 3*, EMNLP '09, pages 1408–1417, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, Republicans and Starbucks afficionados: user classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 430–438.

Kriti Puniyani, Jacob Eisenstein, Shay Cohen, and Eric P. Xing. 2010. Social links from latent topics in microblogs. In *Proceedings of the NAACL/HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 19–20. Association for Computational Linguistics.

Minghui Qiu and Jing Jiang. 2013. A latent variable model for viewpoint discovery from threaded forum posts. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040, Atlanta, Georgia, June. Association for Computational Linguistics.

Alan Ritter, Mausam Etzioni, and Oren Etzioni. 2010. A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434. Association for Computational Linguistics.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961, Jeju Island, Korea.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023

Paul H. Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE*, 6(2):e16782.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Comput. Linguist.*, 30(3):277–308, September.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00082/1567352/tacl_a_00082.pdf by guest on 04 October 2023