Distributed Representations of Geographically Situated Language

David Bamman Chris Dyer Noah A. Smith
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dbamman,cdyer,nasmith}@cs.cmu.edu

Abstract

We introduce a model for incorporating contextual information (such as geography) in learning vector-space representations of *situated* language. In contrast to approaches to multimodal representation learning that have used properties of the *object* being described (such as its color), our model includes information about the *subject* (i.e., the speaker), allowing us to learn the contours of a word's meaning that are shaped by the context in which it is uttered. In a quantitative evaluation on the task of judging geographically informed semantic similarity between representations learned from 1.1 billion words of geo-located tweets, our joint model outperforms comparable independent models that learn meaning in isolation.

1 Introduction

The vast textual resources used in NLP – newswire, web text, parliamentary proceedings – can encourage a view of language as a disembodied phenomenon. The rise of social media, however, with its large volume of text paired with information about its author and social context, reminds us that each word is uttered by a particular person at a particular place and time. In short: language is *situated*.

The coupling of text with demographic information has enabled computational modeling of linguistic variation, including uncovering words and topics that are characteristic of geographical regions (Eisenstein et al., 2010; O'Connor et al., 2010; Hong et al., 2012; Doyle, 2014), learning correlations between words and socioeconomic variables (Rao et al., 2010; Eisenstein et al., 2011; Pennacchiotti and Popescu, 2011; Bamman et al., 2014); and charting how new terms spread geographically (Eisenstein et al., 2012). These models can tell us that *hella* was (at one time) used most often by a particular demographic group in northern California, echoing earlier linguistic studies (Bucholtz, 2006), and that *wicked* is used most often in New England (Ravindranath, 2011); and they have practical applications, facilitating tasks like text-based geolocation (Wing and Baldridge, 2011; Roller et al., 2012; Ikawa et al., 2012). One desideratum that remains, however, is how the *meaning* of these terms is shaped by geographical influences – while *wicked* is used throughout the United States to mean *bad* or *evil* ("he is a wicked man"), in New England it is used as an adverbial intensifier ("my boy's wicked smart"). In leveraging grounded social media to uncover linguistic variation, what we want to learn is how a word's meaning is shaped by its geography.

In this paper, we introduce a method that extends vector-space lexical semantic models to learn representations of geographically situated language. Vector-space models of lexical semantics have been a popular and effective approach to learning representations of word meaning (Lin, 1998; Turney and Pantel, 2010; Reisinger and Mooney, 2010; Socher et al., 2013; Mikolov et al., 2013, *inter alia*). In bringing in extra-linguistic information to learn word representations, our work falls into the general domain of multimodal learning; while other work has used visual information to improve distributed representations (Andrews et al., 2009; Feng and Lapata, 2010; Bruni et al., 2011; Bruni et al., 2012a; Bruni et al., 2012b; Roller and im Walde, 2013), this work generally exploits information about the object being described (e.g., *strawberry* and a picture of a strawberry); in contrast, we use information about the *speaker* to learn representations that vary according to contextual variables from the speaker's perspective. Unlike classic multimodal systems that incorporate multiple active modalities (such as gesture) from a user (Oviatt, 2003; Yu and

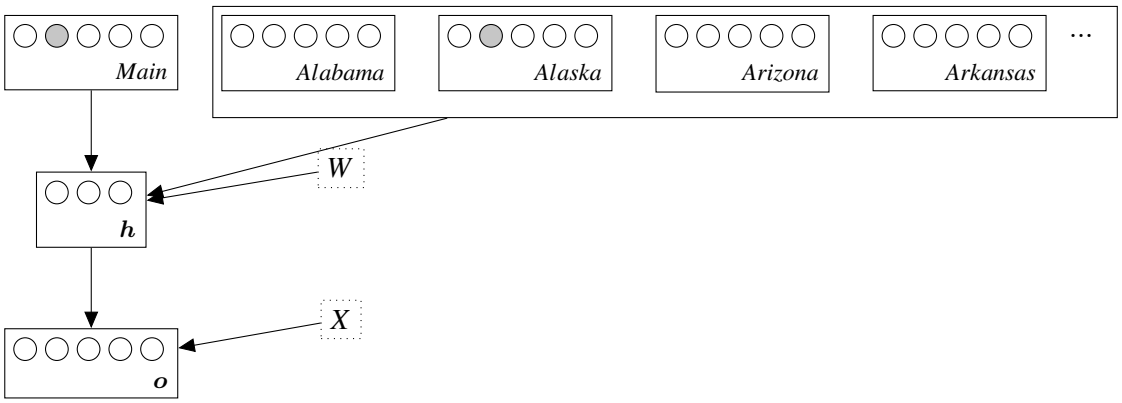

Figure 1: Model. Illustrated are the input dimensions that fire for a single sample, reflecting a particular word (vocabulary item #2) spoken in Alaska, along with a single output. Parameter matrix W consists of the learned low-dimensional embeddings.

Ballard, 2004), our primary input is textual data, supplemented with metadata about the author and the moment of authorship. This information enables learning models of word meaning that are sensitive to such factors, allowing us to distinguish, for example, between the usage of *wicked* in Massachusetts from the usage of that word elsewhere, and letting us better associate geographically grounded named entities (e.g, *Boston*) with their hypernyms (*city*) in their respective regions.

2 Model

The model we introduce is grounded in the distributional hypothesis (Harris, 1954), that two words are similar by appearing in the same kinds of contexts (where "context" itself can be variously defined as the bag or sequence of tokens around a target word, either by linear distance or dependency path). We can invoke the distributional hypothesis for many instances of regional variation by observing that such variants often appear in similar contexts. For example:

- my boy's *wicked* smart
- my boy's *hella* smart
- my boy's *very* smart

Here, all three variants can often be seen in an immediately pre-adjectival position (as is common with intensifying adverbs).

Given the empirical success of vector-space representations in capturing semantic properties and their success at a variety of NLP tasks (Turian et al., 2010; Socher et al., 2011; Collobert et al., 2011; Socher et al., 2013), we use a simple, but state-of-the-art neural architecture (Mikolov et al., 2013) to learn low-dimensional real-valued representations of words. The graphical form of this model is illustrated in figure 1.

This model corresponds to an extension of the "skip-gram" language model (Mikolov et al., 2013) (hereafter SGLM). Given an input sentence $\boldsymbol{s}$ and a context window of size t, each word s_i is conditioned on in turn to predict the identities of all of the tokens within t words around it. For a vocabulary V, each input word s_i is represented as a one-hot vector $\boldsymbol{w}_i$ of length $|V|$. The SGLM has two sets of parameters. The first is the representation matrix $W \in \mathbb{R}^{|V| \times k}$, which encodes the real-valued embeddings for each word in the vocabulary. A matrix multiply $\boldsymbol{h} = \boldsymbol{w}^\top W, \in \mathbb{R}^k$ serves to index the particular embedding for word $\boldsymbol{w}$, which constitutes the model's hidden layer. To predict the value of the context word y (again, a one-hot vector of dimensionality $|V|$), this hidden representation $\boldsymbol{h}$ is then multiplied by a second parameter matrix $X \in \mathbb{R}^{|V| \times k}$. The final prediction over the output vocabulary is then found by passing this resulting vector through the softmax function $\boldsymbol{o} = \text{softmax}(X\boldsymbol{h})$, giving a vector in the $|V|$-dimensional unit simplex. Backpropagation using (input x, output y) word tuples learns the values of W (the embeddings) and X (the output parameter matrix) that maximize the likelihood of y (i.e., the context words) conditioned on x (i.e., the s_i's). During backpropagation, the errors propagated are the difference between $\boldsymbol{o}$ (a probability distribution with k outcomes) and the true (one-hot) output y.

Let us define a set of contextual variables $\mathcal{C}$; in the experiments that follow, $\mathcal{C}$ is comprised solely of geographical state $\mathcal{C}_{state} = \{\text{AK}, \text{AL}, \ldots, \text{WY}\}$) but could in principle include any number of features, such as calendar

month, day of week, or other demographic variables of the speaker. Let $|\mathcal{C}|$ denote the sum of the cardinalities of all variables in $\mathcal{C}$ (i.e., 51 states, including the District of Columbia). Rather than using a single embedding matrix W that contains low-dimensional representations for every word in the vocabulary, we define a global embedding matrix $W_{main} \in \mathbb{R}^{|V| \times k}$ and an additional $|\mathcal{C}|$ such matrices (each again of size $|V| \times k$, which capture the effect that each variable value has on each word in the vocabulary. Given an input word $\boldsymbol{w}$ and set of active variable values $\mathcal{A}$ (e.g., $\mathcal{A} = \{state = \text{MA}\}$), we calculate the hidden layer $\boldsymbol{h}$ as the sum of these independent embeddings: $\boldsymbol{h} = \boldsymbol{w}^\top W_{main} + \sum_{a \in \mathcal{A}} \boldsymbol{w}^\top W_a$. While the word *wicked* has a common low-dimensional representation in $W_{main,wicked}$ that is invoked for every instance of its use (regardless of the place), the corresponding vector $W_{\text{MA},wicked}$ indicates how that common representation should shift in k-dimensional space when used in Massachusetts. Backpropagation functions as in standard SGLM, with gradient updates for each training example $\{x, y\}$ touching not only W_{main} (as in SGLM), but all active $W_{\mathcal{A}}$ as well.

The additional W embeddings we add lead to an increase in the number of total parameters by a factor of $|\mathcal{C}|$. To control for the extra degrees of freedom this entails, we add squared ℓ_2 regularization to all parameters, using stochastic gradient descent for backpropagation with minibatch updates for the regularization term. As in Mikolov et al. (2013), we speed up computation using the hierarchical softmax (Morin and Bengio, 2005) on the output matrix X.

This model defines a joint parameterization over all variable values in the data, where information from data originating in California, for instance, can influence the representations learned for Wisconsin; a naive alternative would be to simply train individual models on each variable value (a "California" model using data only from California, etc.). A joint model has three *a priori* advantages over independent models: (i) sharing data across variable values encourages representations across those values to be similar; e.g., while *city* may be closer to *Boston* in Massachusetts and *Chicago* in Illinois, in both places it still generally connotes a *municipality*; (ii) such sharing can mitigate data sparseness for less-witnessed areas; and (iii) with a joint model, all representations are guaranteed to be in the same vector space and can therefore be compared to each other; with individual models (each with different initializations), word vectors across different states may not be directly compared.

3 Evaluation

We evaluate our model by confirming its face validity in a qualitative analysis and estimating its accuracy at the quantitative task of judging geographically-informed semantic similarity. We use 1.1 billion tokens from 93 million geolocated tweets gathered between September 1, 2011 and August 30, 2013 (approximately 127,000 tweets per day evenly sampled over those two years). This data only includes tweets that have been geolocated to state-level granularity in the United States using high-precision pattern matching on the user-specified location field (e.g., "new york ny" $\rightarrow$ NY, "chicago" $\rightarrow$ IL, etc.). As a preprocessing step, we identify a set of target multiword expressions in this corpus as the maximal sequence of adjectives + nouns with the highest pointwise mutual information; in all experiments described below, we define the vocabulary V as the most frequent 100,000 terms (either unigrams or multiword expressions) in the total data, and set the dimensionality of the embedding $k = 100$. In all experiments, the contextual variable is the observed US state (including DC), so that $|\mathcal{C}| = 51$; the vector space representation of word $\boldsymbol{w}$ in state s is $\boldsymbol{w}^\top W_{main} + \boldsymbol{w}^\top W_s$.

3.1 Qualitative Evaluation

To illustrate how the model described above can learn geographically-informed semantic representations of words, table 1 displays the terms with the highest cosine similarity to *wicked* in Kansas and Massachusetts after running our joint model on the full 1.1 billion words of Twitter data; while *wicked* in Kansas is close to other evaluative terms like *evil* and *pure* and religious terms like *gods* and *spirit*, in Massachusetts it is most similar to other intensifiers like *super*, *ridiculously* and *insanely*.

Table 2 likewise presents the terms with the highest cosine similarity to *city* in both California and New York; while the terms most evoked by *city* in California include regional locations like Chinatown, Los Angeles' South Bay and San Francisco's East Bay, in New York the most similar terms include *hamptons*, *upstate* and *borough*

Kansas		Massachusetts	
term	cosine	term	cosine
wicked	1.000	wicked	1.000
evil	0.884	super	0.855
pure	0.841	ridiculously	0.851
gods	0.841	insanely	0.820
mystery	0.830	extremely	0.793
spirit	0.830	goddamn	0.781
king	0.828	surprisingly	0.774
above	0.825	kinda	0.772
righteous	0.823	#sarcasm	0.772
magic	0.822	soooooooo	0.770

Table 1: Terms with the highest cosine similarity to *wicked* in Kansas and Massachusetts.

California		New York	
term	cosine	term	cosine
city	1.000	city	1.000
valley	0.880	suburbs	0.866
bay	0.874	town	0.855
downtown	0.873	hamptons	0.852
chinatown	0.854	big city	0.842
south bay	0.854	borough	0.837
area	0.851	neighborhood	0.835
east bay	0.845	downtown	0.827
neighborhood	0.843	upstate	0.826
peninsula	0.840	big apple	0.825

Table 2: Terms with the highest cosine similarity to *city* in California and New York.

(New York City's term of administrative division).

3.2 Quantitative Evaluation

As a quantitative measure of our model's performance, we consider the task of judging semantic similarity among words whose meanings are likely to evoke strong geographical correlations. In the absence of a sizable number of linguistically interesting terms (like *wicked*) that are known to be geographically variable, we consider the proxy of estimating the named entities evoked by specific terms in different geographical regions. As noted above, geographic terms like *city* provide one such example: in Massachusetts we expect the term *city* to be more strongly connected to grounded named entities like *Boston* than to other US cities. We consider seven categories for which we can reasonably expect the connotations of each term to vary by geography; in each case, we calculate the distance between two terms x and y using representations learned for a given state ($\delta_{state}(x, y)$).

1. *city*. For each state, we measure the distance between the word *city* and the state's most populous city; e.g., $\delta_{\text{AZ}}(city, phoenix)$.
2. *state*. For each state, the distance between the word *state* and the state's name; e.g., $\delta_{\text{WI}}(state, wisconsin)$.
3. *football*. For all NFL teams, the distance between the word *football* and the team name; e.g., $\delta_{\text{IL}}(football, bears)$.
4. *basketball*. For all NBA teams from a US state, the distance between the word *basketball* and the team name; e.g., $\delta_{\text{FL}}(basketball, heat)$.
5. *baseball*. For all MLB teams from a US state, the distance between the word *baseball* and the team name; e.g., $\delta_{\text{IL}}(baseball, cubs)$, $\delta_{\text{IL}}(baseball, white\ sox)$.
6. *hockey*. For all NHL teams from a US state, the distance between the word *hockey* and the team name; e.g., $\delta_{\text{PA}}(hockey, penguins)$.
7. *park*. For all US national parks, the distance between the word *park* and the park name; e.g., $\delta_{\text{AK}}(park, denali)$.

Each of these questions asks the following: what words are evoked for a given target word (like *football*)? While *football* may everywhere evoke similar sports like *baseball* or *soccer* or more specific football-related terms like *touchdown* or *field goal*, we expect that particular sports teams will be evoked more strongly by the word *football* in their particular geographical region: in Wisconsin, *football* should evoke *packers*, while in Pennsylvania, *football* evokes *steelers*. Note that this is not the same as simply asking which sports team is most frequently (or most characteristically) mentioned in a given area; by measuring the distance to a target word (*football*), we are attempting to estimate the varying strengths of association between concepts in different regions.

For each category, we measure similarity as the average cosine similarity between the vector for the target word for that category (e.g., *city*) and the corresponding vector for each state-specific answer (e.g., *chicago* for IL; *boston* for MA). We compare three different models:

1. JOINT. The full model described in section 2, in which we learn a global representation for each word along with deviations from that common representation for each state.
2. INDIVIDUAL. For comparison, we also partition the data among all 51 states, and train a single model for each state using only data from that state. In this model, there is no sharing among states; California has the most

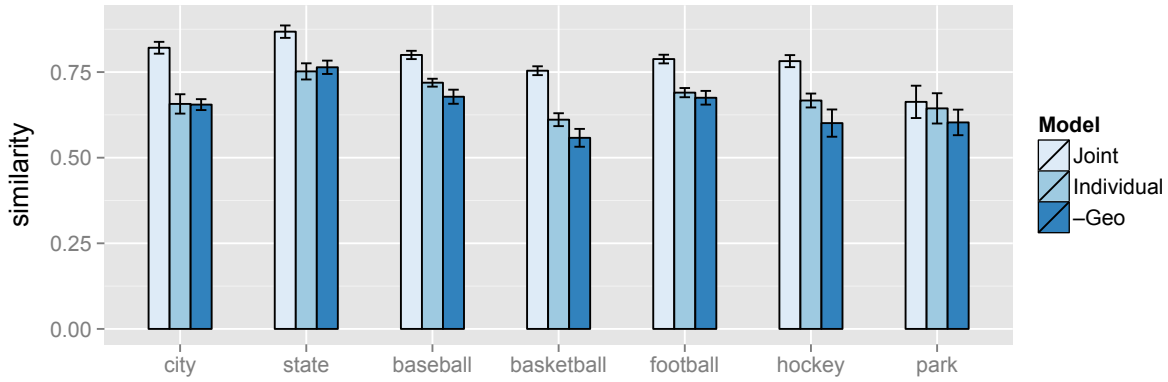

Figure 2: Average cosine similarity for all models across all categories, with 95% confidence intervals on the mean.

data with 11,604,637 tweets; Wyoming has the least with 47,503 tweets.

3. –GEO. We also train a single model on all of the training data, but ignore any state metadata. In this case the distance δ between two terms is their overall distance within the entire United States.

As one concrete example of these differences between individual data points, the cosine similarity between *city* and *seattle* in the –GEO model is 0.728 (*seattle* is ranked as the 188th most similar term to *city* overall); in the INDIVIDUAL model using only tweets from Washington state, $\delta_{WA}(city, seattle) = 0.780$ (rank #32); and in the JOINT model, using information from the entire United States with deviations for Washington, $\delta_{WA}(city, seattle) = 0.858$ (rank #6). The overall similarity for the city category of each model is the average of 51 such tests (one for each city).

Figure 2 present the results of the full evaluation, including 95% confidence intervals for each mean. While the two models that include geographical information naturally outperform the model that does not, the JOINT model generally far outperforms the INDIVIDUAL models trained on state-specific subsets of the data.[1] A model that can exploit all of the information in the data, learning core vector-space representations for all words along with deviations for each contextual variable, is able to learn more geographically-informed representations for this task than strict geographical models alone.

[1]This result is robust to the choice of distance metric; an evaluation measuring the Euclidean distance between vectors shows the JOINT model to outperform the INDIVIDUAL and –GEO models across all seven categories.

4 Conclusion

We introduced a model for leveraging situational information in learning vector-space representations of words that are sensitive to the speaker's social context. While our results use geographical information in learning low-dimensional representations, other contextual variables are straightforward to include as well; incorporating effects for time – such as time of day, month of year and absolute year – may be a powerful tool for revealing periodic and historical influences on lexical semantics.

Our approach explores the degree to which geography, and other contextual factors, influence word *meaning* in addition to frequency of usage. By allowing all words in different regions (or more generally, with different metadata factors) to exist in the same vector space, we are able compare different points in that space – for example, to ask what terms used in Chicago are most similar to *hot dog* in New York, or what word groups shift together in the same region in comparison to the background (indicating the shift of an entire semantic field). All datasets and software to support these geographically-informed representations can be found at: `http://www.ark.cs.cmu.edu/geoSGLM`.

5 Acknowledgments

The research reported in this article was supported by US NSF grants IIS-1251131 and CAREER IIS-1054319, and by an ARCS scholarship to D.B. This work was made possible through the use of computing resources made available by the Open Cloud Consortium, Yahoo and the Pittsburgh Supercomputing Center.

References

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2).

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. In *Proc. of the Workshop on Geometrical Models of Natural Language Semantics*.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012a. Distributional semantics in technicolor. In *Proc. of ACL*.

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. 2012b. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proc. of the ACM International Conference on Multimedia*.

Mary Bucholtz. 2006. Word up: Social meanings of slang in California youth culture. In Jane Goodman and Leila Monaghan, editors, *A Cultural Approach to Interpersonal Communication: Essential Readings*, Malden, MA. Blackwell.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *Proc. of EACL*.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proc. of EMNLP*.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proc. of ACL*.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2012. Mapping the geographical diffusion of new words. *arXiv*, abs/1210.5268.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Proc. of NAACL*.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the Twitter stream. In *Proc. of WWW*.

Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. 2012. Location inference using microblog messages. In *Proc. of WWW*.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of ICLR*.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proc. of AISTATS*.

Brendan O'Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. 2010. Discovering demographic language variation. In *NIPS Workshop on Machine Learning and Social Computing*.

Sharon Oviatt. 2003. Multimodal interfaces. In Julie A. Jacko and Andrew Sears, editors, *The Human-computer Interaction Handbook*, pages 286–304, Hillsdale, NJ, USA. L. Erlbaum Associates Inc.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, Republicans and Starbucks afficionados: User classification in Twitter. In *Proc. of KDD*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proc. of the Workshop on Search and Mining User-generated Contents*.

Maya Ravindranath. 2011. A wicked good reason to study intensifiers in New Hampshire. In *NWAV 40*.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proc. of NAACL*.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proc. of EMNLP*.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proc. of EMNLP-CoNLL*.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proc. of EMNLP*.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with compositional vector grammars. In *Proc. of ACL*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proc. of ACL*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, January.

Benjamin P. Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proc. of ACL*.

Chen Yu and Dana H. Ballard. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception*, 1(1):57–80.