Importance of Semantic Representation: Dataless Classification

Ming-Wei Chang, Lev Ratinov, Dan Roth and **Vivek Srikumar**
Department of Computer Science
University of Illinois at Urbana-Champaign
{mchang21, ratinov2, danr, vsrikum2}@uiuc.edu

Abstract

Traditionally, text categorization has been studied as the problem of training of a classifier using labeled data. However, people can categorize documents into named categories without any explicit training because we know the meaning of category names. In this paper, we introduce *Dataless Classification*, a learning protocol that uses world knowledge to induce classifiers without the need for any labeled data. Like humans, a dataless classifier interprets a string of words as a set of semantic concepts. We propose a model for dataless classification and show that the label name alone is often sufficient to induce classifiers. Using Wikipedia as our source of world knowledge, we get 85.29% accuracy on tasks from the 20 Newsgroup dataset and 88.62% accuracy on tasks from a Yahoo! Answers dataset without *any labeled or unlabeled* data from the datasets. With unlabeled data, we can further improve the results and show quite competitive performance to a supervised learning algorithm that uses 100 labeled examples.

Introduction

Text categorization is traditionally seen as the problem of assigning a label to data. Often, semantic information available in the label is ignored while making this decision and the labels are treated as atomic identifiers. This necessitates the use of annotated data to train classifiers that map documents to these identifiers. On the other hand, humans can perform text categorization without seeing even one training example. For example, consider the task of deciding whether a Usenet post must be posted to *comp.sys.mac.os* or *sci.electronics*. We can do this without explicit training because we use the meaning of the labels to categorize documents.

In this paper, we introduce the notion of *Dataless Classification* – a model of classification that does not need annotated training data. Our approach is based on the use of a source of world knowledge to analyze both labels and documents from a semantic point of view, allowing us to learn classifiers. Such analysis by a semantic interpreter allows us to compare the *concepts* discussed by the document to perform classification.

Using a semantic knowledge source that is based on Wikipedia, we experimentally demonstrate that dataless classification can categorize text without any annotated data. We show the results of classification on the standard 20 Newsgroups dataset and a new Yahoo! Answers dataset. Without even looking at unlabeled instances, dataless classification outperforms supervised methods. Moreover, when *unlabeled* instances are available during training, our methods are comparable to the supervised methods that need 100 training examples.

We can perform *on the fly* text categorization for previously unseen labels, since our classifier was not trained on any particular labels. Furthermore, since we do not need previously labeled data, the model is not committed into any particular domain. Therefore, we can use dataless classification across different data sets. We experimentally show that our method works equally well across domains.

Semantic Representation

Traditionally, in most text categorization tasks, labels are treated as atomic identifiers, thereby losing their meaning. The task of text categorization is regarded as the task of learning a classifier that can distinguish between two abstract categories represented by these identifiers. However, labels often contain valuable semantic information that could be utilized for learning the classifiers.

The simplest approach to represent the semantics of a text fragment is to treat it as a vector in the space of words. We refer to this as the *bag of words (BOW)* representation. While the bag of words representation is one of the most commonly used representation for documents, documents *and* labels have semantic content which often goes beyond the words they contain. For example, though the phrase *'American politics'* can be treated as just the two words, it could connote discussion about a wide range of topics – Democrats, Republicans, abortion, taxes, homosexuality, guns, etc. This indicates that a text fragment can be treated as a vector in the space of concepts, if such a representation is available. The vector of concepts could be thought of as the semantic interpretation of the text fragment.

In order to obtain a more meaningful semantic interpretation of a text fragment, we use *Explicit Semantic Analysis (ESA)*, that was introduced in (Gabrilovich & Markovitch 2007) and uses Wikipedia as its source of world knowledge.

ESA was originally introduced to measure semantic relatedness between text fragments. Given a text fragment, the ESA algorithm generates a set of concepts that are weighted and ordered by their relevance to the input. Here, we provide a brief summary of this approach and refer the reader to (Gabrilovich & Markovitch 2007) and (Gabrilovich & Markovitch 2006) for more details.

The main assumption is that each article in Wikipedia corresponds to a concept. To get the ESA representation of a text fragment, for each word in the text, the interpreter identifies the concepts that contain it. These concepts are combined to form a weighted vector, where the weights are obtained by using the TFIDF representation of the original text. The list of concepts are ordered by the weight to get the final ESA representation.

Since Wikipedia is the largest encyclopedic source of knowledge on the web, ESA representation is sufficient for many categorization tasks. Additionally, since Wikipedia was generated by humans, it provides a natural measure of relatedness between text fragments. Previous research has shown that semantic interpretation based on Wikipedia is a more reliable measure of distance between documents than the traditional bag-of-words approach. In the Discussion section, we present brief thoughts on why dataless classification works and what makes a resource of knowledge sufficient for this purpose.

In our experiments, we use the BOW and ESA representation schemes for both label names and the documents.

Data and Experimental Methodology

To evaluate our ideas, we need datasets where the labels contain rich semantic information. Here, we describe the two datasets that we used.

20 Newsgroups Dataset The 20 Newsgroups Dataset is a common benchmark used for testing classification algorithms. The dataset, introduced in (Lang 1995), contains approximately 20,000 newsgroup posts, partitioned (nearly) evenly across 20 different newsgroups. Some of the newsgroups are very closely related to each other (e.g. *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware*), while others are unrelated (for example, *misc.forsale* and *soc.religion.christian*). The 20 topics are organized into broader categories: computers, recreation, religion, science, forsale and politics. This dataset has been used for many learning tasks (Raina, Ng, & Koller 2006; McCallum *et al.* 1998).

Since our approach uses the content of label names, good label names are essential for the performance. We cleaned the existing labels by expanding the newsgroup names that were used in the original data. For example, we expanded *os* into *operating system* and *mac* to *macintosh apple*. We also removed irrelevant words – *misc, alt, rec* and *talk*. The amended label names given for each class are summarized in Table 1. This amendment is necessary because the existing labels are sometimes not real words or, as in the case of *for sale*, the expansion contains a stop word, which necessitates further expansion.

Newsgroup Name	Expanded Label
talk.politics.guns	politics guns
talk.politics.mideast	politics mideast
talk.politics.misc	politics
alt.atheism	atheism
soc.religion.christian	society religion christianity christian
talk.religion.misc	religion
comp.sys.ibm.pc.hardware	computer systems ibm pc hardware
comp.sys.mac.hardware	computer systems mac macintosh apple hardware
sci.electronics	science electronics
comp.graphics	computer graphics
comp.windows.x	computer windows x windowsx
comp.os.ms-windows.misc	computer os operating system microsoft windows
misc.forsale	for sale discount
rec.autos	cars
rec.motorcycles	motorcycles
rec.sport.baseball	baseball
rec.sport.hockey	hockey
sci.crypt	science cryptography
sci.med	science medicine
sci.space	science space

Table 1: Newsgroup label names. We expanded the names of the newsgroups to full words and removed some words like misc. This table lists the expanded newsgroup names.

Yahoo! Answers The second dataset that we used for our experiments is based on Yahoo! Answers and was collected for these tasks.[1] We extracted 189,467 question and answer pairs from 20 top-level categories from the Yahoo! Answers website (that is, about 10,000 question/answer pairs per category). These top-level categories have a total of 280 subcategories which refine the labels, though the distribution of question/answer pairs at the subcategory is not uniform. Table 2 shows a sample of category and subcategory names. For our experiments, we used the original category and subcategory names as label names.

Top-level Category	Subcategory
Arts And Humanities	Theater Acting
Business And Finance	Advertising Marketing
Business And Finance	Taxes
Computers And Internet	Security
Consumer Electronics	Play Station
Entertainment And Music	Jokes Riddles
Games And Recreation	Video Online Games
Sports	German Football Soccer
Sports	Rugby League

Table 2: Sample Yahoo! Answers categories and subcategories. In all, we collected 20 top level categories and 280 subcategories from Yahoo! Answers.

[1] Please contact the authors for the Yahoo! Answers data.

Id	Problem
1	Motorcycles Vs Ms-Windows
2	Baseball Vs Politics.misc
3	Religion Vs Politics.guns
4	Atheism Vs Autos
5	IBM hardware Vs Forsale
6	Politics.mideast Vs Sci.med
7	Christianity Vs Hockey
8	Space Vs Mac.Hardware
9	Windows.X Vs Electronics
10	Sci.Cryptography Vs Comp.graphics

Table 3: The set of 10 binary classification problems used in (Raina, Ng, & Koller 2006) for the 20 newsgroups data.

Experimental Methodology

(Raina, Ng, & Koller 2006) used the 20 Newsgroup dataset to construct ten binary classification problems, shown in Table 3. We used these problems for our experiments. In (Raina, Ng, & Koller 2006), they use a logistic regression classifier. As our supervised baseline, we trained a naive Bayes classifier using 10 labeled examples and observed similar accuracies.

For the Yahoo! Answers dataset, we generated 20 random binary classification problems at the subcategory level. Some of these problems are shown in 4. We report the average performance over all the problems and then focus on specific cases which highlight the strengths and weaknesses of our approach.

Intuitively, these are 'easy' classification problems for humans. However, with 10 training samples, (Raina, Ng, & Koller 2006) reported the error rates as high as 20% in 8 out of 10 problems, and even with 100 labeled samples, the error rate on the *religion vs. politics.guns* problem was above 20%. Using the on-the-fly classification technique described in the next section, we achieve error rates below 11.5% on 9 out of 10 problems with *no* labeled data at all.

Id	Description
⋮	⋮
14	Health Diet Fitness Health Allergies
15	Business And Finance Small Business Consumer Electronics Other Electronics
16	Consumer Electronics DVRs Pets Rodents
17	Business And Finance India Business And Finance Financial Services
18	Sports Mexican Football Soccer Social Science Dream Interpretation
19	Sports Scottish Football Soccer Pets Horses
20	Health Injuries Sports Brazilian Football Soccer

Table 4: Sample binary categorization problems for the Yahoo! Answers dataset – subcategory level

Dataset	Supervised Baseline	NN-BOW	NN-ESA
Newsgroups	71.71	65.73	85.29
Yahoo	84.34	66.79	88.62

Table 5: Average accuracy of dataless classification with the bag of words and ESA representations on the Newsgroups and Yahoo! Answers datasets. Note that in both domains, the use of ESA representation is better than using just the words of the labels and no other training data. In fact, the ESA representation, without any training data outperforms supervised naive Bayes classifiers which use ten labeled examples.

On-the-fly Classification

Classification with *no* data

We use the term "on-the-fly classification" to refer to the situation where category names are not known in advance and hence, no training can be done. Essentially, on-the-fly classification is a technique for classifier induction that does not use *any* data at all and uses no knowledge of the datasets either. We show that with no data other than the label names, we can get very good performance for classifying documents. The efficacy of this approach stems from the rich semantic representation discussed earlier. In order to demonstrate the effectiveness of the semantic representation, we use a very simple classification technique – we perform Nearest Neighbors classification on an appropriately selected feature space. Let $\varphi(t)$ denotes some vector representation of text t. For a document d and labels l_i, we pick the category i if $arg\min_i ||\varphi(l_i) - \varphi(d)||$.

Using the Bag of Words (BOW) representation, which represents text fragments as vectors in the space of words, we get the **NN-BOW** classifier. The ESA representation represents text as vectors in the space of concepts. Using the ESA representation gives us the **NN-ESA** classifier.

Results

The results of dataless classification on the binary categorization tasks for the Newsgroups and the Yahoo! Answers datasets are summarized in Table 5. It is clear that the **NN-BOW** algorithm can classify documents correctly only if the document contains words that appeared in the label names. Therefore, the recall of this method is quite limited. We can see **NN-ESA** performs much better than bags of words approach. This implies **NN-ESA** can categorize of documents even if they share no common words with the label name. Since the ESA representation maps a text snippet into high-dimensional semantic space, two documents may be very close in the ESA space even if they share no common terms. This results in an on-the-fly classifier that can use dynamic, ad-hoc labels.

In Table 5, we also show as baseline, the performance of a supervised naive Bayes classifier that was learned with ten examples. We can see that with both the Newsgroups and Yahoo! Answers datasets, the **NN-ESA** classifier outperforms the supervised classifier.

Semantics of Unlabeled Data

Modeling Unlabeled Data

In the previous section, we show that even without knowing anything about the data, a rich semantic representation can perform on par with supervised methods. However, often we have access to unlabeled documents belonging to the problem of interest. This allows us to take advantage of previous work in semi-supervised learning (Refer, for example, (Nigam *et al.* 2000)).

Bootstrapping A straightforward extension of the on-the-fly scheme is to bootstrap the learning process with only the label names as examples. Algorithm 1 presents the pseudo-code for learning a bootstrapped semi-supervised classifier using a feature representation φ. Note that though we do perform training, we still do not use any labeled data and use only the label names as the starting point in the training. (Steps 1 to 4 indicate this.) Using the BOW and ESA representations with this algorithm gives us the **BOOT-BOW** and **BOOT-ESA** classifiers respectively.

Algorithm 1 Bootstrap-φ. *Training a bootstrapped classifier for a feature representation φ, where φ could be Bag of Words or ESA.*

1: Let training set $T = \emptyset$
2: **for all** labels l_i **do**
3: Add l_i to T with label i
4: **end for**
5: **repeat**
6: Train a naive Bayes classifier NB on T
7: **for all** d_i, a document in the document collection **do**
8: If $y = NB.classify(\varphi(d_i))$ with high confidence
9: Add d_i to T with label y
10: **end for**
11: **until** No new training documents are added.

Co-training The classifiers **Boot-BOW** and **Boot-ESA** are learned by bootstrapping on the bag of words and the ESA representations of the data respectively. The fact that BOW and ESA are parallel representation of the same data is ignored. Prior work ((Blum & Mitchell 1998)) has studied the scenario when two independent feature representations (or views, as they are called in (Blum & Mitchell 1998)) $\varphi_1(d)$ and $\varphi_2(d)$ are available for the same data and that if each feature representation is sufficient for correct classification of the data. In such a case, we can train two classifiers $c_{\{\varphi_1\}}$ and $c_{\{\varphi_2\}}$ that classify data better than chance. These two classifiers can train one another in a procedure called Co-training. We can apply a variant of this idea for our task. The algorithm for co-training is summarized in Algorithm 2. While the ESA representation is a function of the BOW representation, violating the 'view independence' assumption, we show that in practice, this procedure leads to a satisfactory performance. [2]

[2]We also experimented with concatenating the BOW and the ESA representations into a single BOW+ESA view as was done in (Gabrilovich & Markovitch 2005) for supervised classification. Then we bootstrapped the naive Bayes classifier on the BOW+ESA view, but it consistently performed worse than co-training.

Algorithm 2 Co-training *We use the fact that BOW and ESA can independently classify the data quite well to induce a new classifier.*

1: Let training set $T^{BOW} = \emptyset$,$T^{ESA} = \emptyset$.
2: **for all** labels l_i **do**
3: Add l_i to both T^{ESA} and T^{BOW} with label i
4: **end for**
5: **repeat**
6: Train a naive Bayes classifier NB^{BOW} on T^{BOW}.
7: Train a naive Bayes classifier NB^{ESA} on T^{ESA}.
8: **for all** d_i, a document in the document collection **do**
9: **if** Both NB^{BOW} and NB^{ESA} classify d_i with high confidence **then**
10: Add d_i to T^{BOW} with label from NB^{BOW}
11: Add d_i to T^{ESA} with label from NB^{ESA}
12: **end if**
13: **end for**
14: **until** No new training documents are added

Results

The performance of the different algorithms with the Newsgroup and Yahoo! Answers datasets is summarized in Table . Additionally, we also show the performance of a supervised baseline system that used a naive Bayes algorithm with 100 training examples. We can see from the table that both the **BOOT-ESA** and **Co-train** classifiers are competitive with the supervised classifier. Even the **BOOT-BOW** classifier is comparable to the supervised baseline. This indicates that using unlabeled data, we can train very strong classifiers if we use the semantics of the labels.

Discussion

Our results have showed that using a good representation can have a dramatic impact on classification. In this section, we discuss when and why simply using the label name is sufficient for classification and provide a justification for our method. Our goal is to develop an understanding of what makes a dataset a good semantic resource, so that it will be possible to apply our framework using resources other than Wikipedia.

Let W be the set of all words. Consider a classification task in which the categories are $c_1, c_2, \ldots, c_m$. We assume that there exist collections $G_1, G_2, \ldots, G_m$ of words that are 'good' for these categories. Informally, we interpret 'good' to have a discriminative meaning here. That is, a word is 'good' for c_i if it is more likely to appear in c_i-documents than in documents of other categories.

The notion of 'good' directly relates to a key assumption about the relation between category labels and the semantic knowledge resource that we use (in our experiments, Wikipedia). Specifically, if a user chooses l to be a label of a category c, and thus thinks that l is a 'good description' of the category, it means that Wikipedia articles that contain l should contain other words that are 'good' for c. Consequently, our approach for generating a semantic representation using l and the Wikipedia articles will bring a

Dataset	Supervised 10 Samples	Supervised 100 Samples	BOOT-BOW	BOOT-ESA	Co-train
Newsgroup	71.71	92.41	88.84	90.92	92.70
Yahoo! Answers	84.34	94.37	90.70	92.73	95.30

Table 6: Performance of classifiers when unlabeled data is available. The supervised baseline, in this case, is a naive Bayes classifier that is trained on 100 labeled examples. The last three classifiers do not use any labeled examples at all.

document that belongs to category c closer to the representation of l, than it would a document that does not belong to c. Even if the approximation by the label name is not of a high quality, given a good representation, classifiers can be discriminating.

Let $\varphi(.)$ be the semantic representation which receives a document and generates the semantic representation. First, we assume that the representation is good. This means that there exits a an oracle document w^i for category i such that all documents that belong to c_i are close to w^i. More formally,

$$\|w^i - \varphi(d)\| \leq \|w^j - \varphi(d)\| - \gamma, \forall j \neq i \qquad (1)$$

where γ can be viewed as margin.

Note that our approximation for w^i is by $\varphi(\{l_i\})$ which is the representation for the label name of category c_i. We say that the approximation is reasonable, which means the distance between $\varphi(\{l_i\})$ and w^i is not too far. Formally,

$$\|w^i - \varphi(\{l_i\})\| < \eta.$$

By triangle inequality,

$$\begin{aligned} \|\varphi(d) - \varphi(\{l_i\})\| &\leq \|w^i - \varphi(d)\| + \|w^i - \varphi(\{l_i\})\| \\ &= \|\varphi(d) - w^i\| + \eta. \end{aligned} \qquad (2)$$

Combining (1) and (2), it follows that

$$\begin{aligned} \|\varphi(d) - \varphi(\{l_i\})\| &\leq \|w^i - \varphi(d)\| + \eta \\ &\leq \|w^j - \varphi(d)\| - \gamma + \eta, \forall j \neq i, \\ &\leq \|\varphi(d) - \varphi(\{l_j\})\| - \gamma + 2\eta, \forall j \neq i \end{aligned}$$

The final step of the above inequality implies that if the representation is good (that is, if γ is large enough), then η can be as large as $\gamma/2$ without changing the classification result. This justifies the intuition that if the distance between the categories is large enough in some space, then the approximation that the labels provide for the categories need not be perfect for good classification performance.

The previous discussion suggests that often, with the appropriate representation, dataless classification is *easy* since the categories are far apart. We believe that, often, when the problem seems to be hard for dataless classification, it could be due to a very specific definition of a category, that may be different than the one expressed in our resources (Wikipedia).

Consider, for example, the difficult pair of *Christianity vs. Religion* in the 20 Newsgroup data. We have listed the top five discriminative words (that is, highest weighted words for both classes according to a discriminative classifier) for this task in Table 7. It is clear that (i) it is difficult to distinguish the document based on these words even for human, and (ii) there are some unrepresentative words which happened to be 'good' in this task. Since it is unlikely we can find a document in Wikipedia mentions *Christianity* and *nntp* at the same time, it is hard for dataless classification to perform well. Therefore, it is crucial to the success of the task is that the 'good' words for the specific classification task should be representative to their label names.

Newsgroup name	Discriminative words
Christianity	morality, host, nntp, writes, koresh
Religon	thanks, quite, assume, clh, rutgers

Table 7: The 'discriminative words' found by labeled data of the task *Christianity vs. Religion* in 20 newsgroup.

Dataless Classification for Domain Adaptation

The problem of discriminating two concepts across several domains is an important open problem in machine learning called *domain adaptation*, which recently has received a lot of attention (Daumé & Marcu 2006; Jiang & Zhai 2007). In this section, we claim that the semantic representation of documents are 'universal' and works across domains. Therefore, in document classification, universal representation diminishes the necessity of domain adaptation.

In traditional approaches to text categorization, when a classifier is trained on a given domain, it may not categorize documents well in a new domain. Informally, the classifier has learned the vocabulary used to express the category's documents in a specific domain because different domains might use different words to express the same ideas.

On the other hand, using the dataless approach simplifies the problem of domain adaptation. First of all, the label names are the same from different domains. Furthermore, documents of the same category from different domains should project to similar concepts because of the wide coverage of Wikipedia. Therefore, projecting the documents onto the space of concepts works well across domains. Our interpretation is that semantic representation categorizes documents as belonging to a category based on their *meaning* rather than the surface representation.

As our running example, we choose to focus on discriminating documents pertaining to *baseball* and *hockey*. The categorization accuracy with 5-fold cross-validation are 0.97 for the Newsgroup domain, and 0.93 for the Yahoo domain using BOW approach (training data and testing data are from the same domain).

The results for domain adaptation are shown in Table 8.

	Model	Features	Accuracy
20NG → 20 NG	Supervised	BOW	0.97
Yahoo → 20 NG	Supervised	BOW	0.60
20NG → 20 NG	Supervised	ESA	0.96
Yahoo → 20 NG	Supervised	ESA	0.90
$\emptyset$ → 20NG	Dataless	ESA	0.96
Yahoo → Yahoo	Supervised	BOW	0.93
20NG → Yahoo	Supervised	BOW	0.89
Yahoo → Yahoo	Supervised	ESA	0.97
20NG → Yahoo	Supervised	ESA	0.96
$\emptyset$ → Yahoo	Dataless	ESA	0.94

Table 8: Analysis for adaptation from $Source \rightarrow Target$ for different domains.

When we applied the NB classifier trained on the BOW representation of 20NG to Yahoo data , the accuracy dropped from 0.93 down to 0.89, and when we applied the classifier trained on Yahoo to the Newsgroup domain, the accuracy dropped down significantly from 0.97 to 0.60. This shows that the BOW classifiers are very sensitive to data distribution. In contrast, when ESA representation is used instead of BOW, 5-fold cross validation was 0.96 on 20NG and 0.97 for Yahoo. When the NB classifier trained on the ESA representation of Yahoo documents was applied to 20NG, the performance dropped only slightly to 0.90. When we applied the classifier trained on ESA representation of 20NG documents to Yahoo, the accuracy was 0.96.

However, the most significant result is that when we applied the dataless algorithm **NN-ESA**, presented in Section (where we used only the label name), the performance was 0.94 on the 20NG dataset and 0.96 on the Yahoo dataset, which are very competitive with the result by supervised learning algorithm with *in-domain* training data.

These results demonstrate the good adaptation properties of the ESA-representation-based approaches in general, and the dataless NN-ESA approach presented in this paper, which uses the universal cross-domain semantic information present in the label to classify data across domains.

Conclusions and Future Work

Quite often, classification tasks are defined by specifying labels that carry meaning. Text categorization is a prime example in that the labels identify a category with words that describe the category. In this paper, we develop the notion of *Dataless Classification* that exploits this situation to produce on-the-fly classifiers without the need for training. Furthermore, using the unlabeled data, we can improve our training to get highly accurate classifiers without looking at any labeled examples.

We note that, unlike traditional learning, where we need at least two classes to learn a classifier, the idea of using a semantic interpreter could be applied to create a one-class classifier too. For example, we could think of a 'baseball-classifier', which identifies documents that discuss baseball. This bridges the notions of classification and information retrieval.

The semantic interpreter plays an important role in the performance of our system. An appealing aspect of ESA is that it covers a wide-ranging set of topics. However, we could replace Wikipedia with a different source of world knowledge and define our semantic interpreter using this source. For example, in (Gabrilovich & Markovitch 2005), the authors use the Open Directory Project to construct a semantic interpreter. Additionally, we could create semantic interpreters with specialized data sources if we are interested in categorizing documents related to a specific area.

Acknowledgments

We wish to thank Evgeniy Gabrilovich for providing his implementation of ESA for our experiments. This work was supported by NSF grant NSF SoD-HCER-0613885 and by MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC.

References

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *COLT*, 92–100.

Daumé, H., and Marcu, D. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence* 26:101–126.

Gabrilovich, E., and Markovitch, S. 2005. Feature generation for text categorization using world knowledge. In *IJCAI*, 1048–1053.

Gabrilovich, E., and Markovitch, S. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI*, 1301–1306.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJCAI*, 1606–1611.

Jiang, J., and Zhai, C. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the annual meeting of ACL*, 264–271.

Lang, K. 1995. NewsWeeder: learning to filter netnews. In *ICML*, 331–339.

McCallum, A. K.; Rosenfeld, R.; Mitchell, T. M.; and Ng, A. Y. 1998. Improving text classification by shrinkage in a hierarchy of classes. In *ICML*, 359–367.

Nigam, K.; McCallum, A. K.; Thrun, S.; and Mitchell, T. M. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39(2/3):103–134.

Raina, R.; Ng, A. Y.; and Koller, D. 2006. Constructing informative priors using transfer learning. In *ICML*, 713–720.