

Topic Spotting in Dialogues using Knowledge Transfer

Rakesh Gupta
Honda Research Institute
Mountain View, CA, USA
rgupta@hra.com

Lev Ratinov
Department of Computer Science, University of
Illinois
Urbana, IL, USA
ratinov2@uiuc.edu

ABSTRACT

Multi-category classification of short dialogues is a common task performed by humans. For example when assigning a question to an expert, a customer service operator tries to classify the customer query into one of N different classes for which experts are available. Similarly, questions on the web (for example Yahoo Answers¹) can be automatically forwarded to a restricted group of people with a specific expertise. Typical questions are short and assume background world knowledge for correct classification. For example, the query *My focus doesn't start* assumes the world knowledge that *focus* is a car model.

We refer to the problem of applying external knowledge to a particular classification task as a *knowledge transfer* problem. We propose two novel knowledge-transfer algorithms with feature generation. We show that by using a wide array of existing and novel knowledge-transfer algorithms, with extremely large amounts of external data, we can significantly improve the classification accuracy. Using over 50GB of external data including Wikipedia, Google 5-grams collection, and Yahoo Answers, we show 17% error reduction in written dialogue classification in the Switchboard dataset. We implemented our approach in speech recognition system, improving on the previous results in spoken text categorization.

Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning; I.2.7 [Artificial Intelligence]: Natural Language Processing - Text analysis

General Terms

Algorithms, Experimentation

Keywords

Knowledge Transfer, Classification/Clustering methods, Machine Learning, Topic Detection

1. INTRODUCTION

Multi category classification is a common task performed by humans. For example when assigning a question to an expert, a customer service operator tries to classify the customer query into one of N different classes for which experts

¹<http://answers.yahoo.com/>

are available. Similarly questions on the web (for example Yahoo Answers) can be automatically forwarded to a restricted group of people with a specific expertise. These problems can be formulated as a multi-class classification with one caveat. The information in the customer query or the question is not sufficient and assumes background world knowledge for correct classification. For example, the query *My focus doesn't start* assumes the world knowledge that *focus* is a car model.

For written and transcribed text, recent works in topic detection and tracking [17, 8] solve the simultaneous problem of segmenting text and classifying each segment into a topic. Methods for finding topic boundaries include text tiling [11] and lexical cohesion. These systems have two problems for spoken dialogues. First, they need a lot of words - typically one or two order of magnitudes more than a typical dialogue (approximately 2-20 words). Secondly, these methods process the data in batch and are not suitable for real time topic detection.

For spoken text, Myers et al. [15] performed classification over conversations and dialogues spanning from several seconds to six minutes. The spoken dialogues were first converted to written texts using off-the-shelf speech recognition software. The resulting noisy texts were then classified to topics. Their work used the switchboard corpus and achieved recognition accuracy of 15.4% for audio dialogues over several seconds, a result which is only slightly better than the majority class baseline (13%) using audio data. The low performance can be explained by the insufficient information available in the short dialogues as well as by the errors introduced by speech recognition system. They report 50.9% classification accuracy using transcripts of the same audio data². We revisit this work using external data sources, better classification techniques and obtain significantly better baseline recognition. We further improve classification accuracy over this improved baseline using external knowledge.

External knowledge can be obtained from the web in many flavors: structured and unstructured, labeled and unlabeled. As an example of unstructured unlabeled external data set, we use the Google collection of 5-grams and their appearance count on the Web [20]. As an example of an external labeled data, we use the Yahoo Answers online data repository, which contains question/answer pairs cataloged by categories and sub-categories. As an example of unlabeled, but structured dataset, we use Wikipedia, which contains arti-

²Using transcripts of the audio data is equivalent to assuming a faultless speech recognition system.

cles organized by titles, and perhaps, loosely organized by categories. We note that unlike most similar work that utilizes external data sources, we do not assume that the external data and the data for the primary classification task come from the same distribution. Neither do we assume that the labeled external resources share a common labels with the primary dataset. Furthermore, we do not make the common assumption that the training set for the new task is limited, instead we show that even when thousands of labeled samples are available, external data resources can significantly improve the performance.

Humans tend to seamlessly utilize background knowledge coming from different and diverse sources in their classification. We are interested in duplicating this human capability in our work. A bag of words approach does not incorporate knowledge beyond words in the customer query or question and hence does not incorporate any background world knowledge. Another alternative is semantic understanding of natural language which is beyond the current state of art. Gabrilovich [5, 6, 7] proposed using feature generation using a single external source of knowledge like Open Directory Project (ODP) or Wikipedia. In contrast, our work leverages multiple labeled and unlabeled sources of knowledge as well as different data transfer in a single classifier in a real time system.

As the amount of publicly available labeled and unlabeled datasets³ increases, using and reusing these valuable resources for new tasks is the central challenge in machine learning. In this paper, we denote this challenge as *knowledge transfer* task. Several branches of research, such as: domain adaptation [12, 13], transfer learning [19, 2], self-taught learning [18], semi-supervised learning [16, 3], have addressed this problem from different angles. In this paper we develop novel knowledge transfer techniques and leverage work from multiple existing branches of research for different labeled and unlabeled data categories, combining them to improve classification of short dialogues.

The heart of our framework is using *diverse* auxiliary datasets (labeled and unlabeled), *several orders of magnitude larger* than the primary classification task, as well as selected ideas from previous research. We focus on the task of classifying the dataset X_{test} given the training set X_{train} and auxiliary datasets $\{D_1, D_2, \dots, D_N\}$ as well as knowledge-transfer algorithms $\{A_1, A_2, \dots, A_M\}$ ⁴. We proceed as follows: for each auxiliary dataset D_i and for each knowledge-transfer algorithm A_j , we use X_{train} to learn a classifier $C_{i,j}$ that gives a "recommendation" regarding any $x \in \{X_{test} \cup X_{train}\}$. We then train a meta-classifier C that learns to combine the recommendations of the classifiers $\{C_{i,j}\}_{i,j}$. The approach is illustrated in figure 1.

We show that intelligently combining predictions of *diverse* knowledge-transfer classifiers trained on *diverse* auxiliary dataset significantly improves the performance, even when individual knowledge-transfer classifiers perform *close to the supervised baseline accuracy*. The reason for that may be that while we keep the hypothesis search space tractable, the predictions of the auxiliary classifiers can be viewed as

³For example: OpenMind, Wikipedia, Yahoo Answers, Google 5-grams Collection, CYC, WordNet, Part-of-speech and Question Answering datasets, etc.

⁴These can be any semi-supervised, transfer learning or self-taught learning algorithms that were developed in the research community

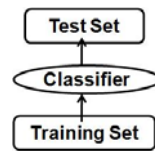


Figure 2: Traditional Classifier

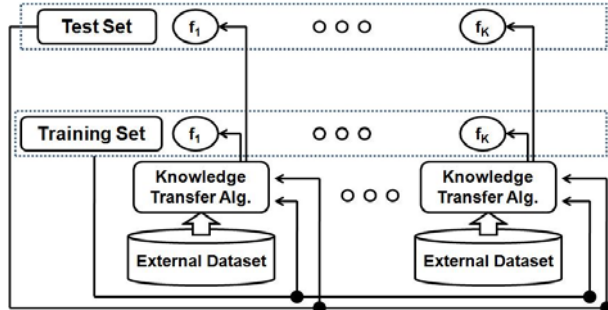


Figure 3: Feature generation Step

more expressive features. This allows us to learn more expressive functions that nevertheless do not overfit. For example, in our experiments, all the classifiers have linear decision boundary. However learning a global classifier on top of the local classifiers thresholds each linear decision boundary essentially learning a non-linear function.

Our problem is formulated as follows: given labeled data such as from Switchboard corpus and Google Question/Answers, use unlabeled and differently labeled data such as from Yahoo Question/Answers, Wikipedia, Google 5-grams, OpenMind to improve classification accuracy. We use two approaches. The first approach is category and cluster based feature generation, where we add features to the training set, injecting knowledge from external sources. Category based features are described in section 2.1 on Predictions as Features (PAF), and cluster based features are described in section 2.3 Structure Based Feature generation (SBFG). The second approach attempts to improve probability estimates obtained from the primary-task training set using external data (Yahoo Answers, Wikipedia and Google 5-gram datasets), and is similar in spirit to the traditional semi-supervised learning. The key difference is that we relax the assumption that the labeled and the unlabeled data come from the same distribution, modifying the algorithm accordingly. The resulting algorithm, Agnostic Semi Supervised Learning (ASSL), is described in section 2.2. Subsequent sections describe the datasets, experiments setting, processing on different corpuses, followed by sections on inter annotator agreement, results and conclusions. Our main results appear in section 5.

2. KNOWLEDGE-TRANSFER ALGORITHMS

The key to our approach is combining predictions of diverse knowledge-transfer classifiers trained on diverse auxiliary datasets. In this section we describe the different feature generation techniques depending on the type of external knowledge we are using. We start by considering a traditional classification approach, described in Figure 2. A training set (with some feature representation) is used to

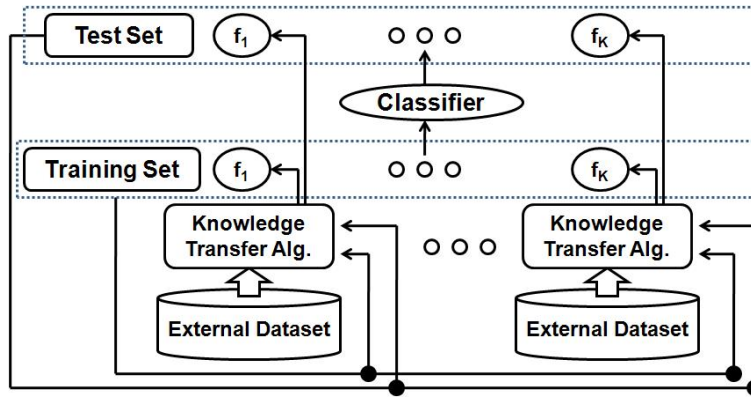


Figure 1: Overview of the proposed system.

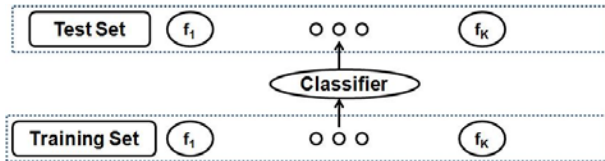


Figure 4: Classification with Expanded features

train a classifier, which is then used to classify the test set, represented in the same feature space. Stepping ahead, the final stage of our approach will be using this traditional supervised training paradigm, but with a richer feature space. The final stage of our approach is shown in Figure 4, where the classifier is trained and tested on the primary dataset, which has been augmented with expanded (expressive) features.

The key step in our approach is the feature generation step, shown in Figure 3. Multiple knowledge-transfer algorithms, using external knowledge sources, are applied to augment the training and the testing samples of the primary task with expanded expressive features. After the additional features are generated, we combine prediction from multiple knowledge sources by throwing all the features computed by different techniques along with the original features into one top-level classifier. The full picture is summarized in Figure 1. We also state at this stage that all the classifiers we used (both for top-level decision making and for feature generation) are *Naive Bayes* classifiers, unless explicitly mentioned otherwise. This completes the top-level overview of the proposed approach. The rest of the section describes the existing and the two novel individual components that are used in feature generation respectively.

2.1 Predictions as Features (PAF)

Given a labeled auxiliary dataset D_{aux} , the simplest way to transfer knowledge to the new task is to learn a classifier C_{aux} on D_{aux} and use the predictions of C_{aux} as additional features for data samples of the new task. The labels between the auxiliary task and the primary task need not be identical, as long as there is any correlation between the labels, the knowledge from the auxiliary task can be injected through this correlation. For example, consider a primary dataset D_{prim} , with a 3-way classification problem between *Travel*, *DiningOut*, *BeautyAndStyle*. Assume an auxiliary

dataset D_{aux} is available, where the documents in D_{aux} are organized in the following categories: *Restaurants*, *Bars*, *CarRentals*.

If we train a classifier C_{aux} on the auxiliary dataset, we expect that, when encountered with samples from D_{prim} , C_{aux} will tend to predict the label *CarRental* for samples from the *Travel* category and the labels *Restaurants* or *Bars* to the samples from the *DiningOut* category. We add the predictions of C_{aux} as features in D_{prim} , and train a classifier C_{prim} on the modified dataset. We note that at the feature generation stage, the additional features need not be the predicted labels, they can be similarity (or confidence) scores associated with each label from the auxiliary domain. In our experiments, we used a set of boolean features, indicating a set of labels from the auxiliary domain that can be assigned to the sample from the primary domain with confidence exceeding a preset threshold.

For example, consider the auxiliary Yahoo Answers dataset, organized into 822 subcategories (sample subcategories are shown in Table 1). For each data sample x_i in the primary dataset, we add 822 binary features f_{ij} if x_i is similar to subcategory j in Yahoo Answers above some threshold. The similarity is defined by cosine similarity of TF-IDF vectors. The subcategories in Yahoo Answers are represented by 20 top TF-IDF words appearing in the documents belonging to the subcategory. This stage is very similar to what was done in [5].

We note that, while the predictions of the classifier C_{aux} can be used directly as the extended expressive features for the global meta-classifier in Figure 4, we use the prediction of C_{prim} instead. The reason is that when using multiple auxiliary classifiers, each classifier C_{aux} will generate varying number of features (822 for the Yahoo Answers subcategory example). Thus, combining outputs of multiple auxiliary classifiers C_{aux} directly into the meta-classifier can lead to few auxiliary classifiers C_{aux} dominating and/or contaminating the generated features. By training the classifier C_{prim} , we make sure that each knowledge-transfer algorithm injects the same type of features to the meta-classifier.

2.2 Agnostic semi-supervised learning (ASSL)

When the auxiliary dataset D_{aux} is not labeled, another, rather simple approach to transfer knowledge from D_{aux} to the new task of classifying documents in the primary dataset D_{prim} is to treat the data in D_{aux} as unlabeled data

Categories (20)	Sub categories (822)
Society & Culture	Halloween Hanukkah Languages Mythology & Folklore Other Cultures & Groups Seniors
Arts & Humanities	...
Beauty & Style	...
Business & Finance	
Cars & Transportation	
Computers & Internet	
Consumer Electronics	
Dining Out	
Education & Reference	
Entertainment & Music	
Food & Drink	

Table 1: Examples of Yahoo Answers Categories and Sub-categories

from D_{prim} and to run a traditional semi-supervised learning algorithm. Unfortunately, a key assumption in semi-supervised learning is that the unlabeled data is drawn from the same distribution as the labeled data. In our case, this assumption is equivalent to assuming that D_{aux} and D_{prim} are drawn from the same distribution. Even when this assumption holds, semi-supervised learning has often been reported to *decrease* the performance compared to traditional supervised approach [3]. We therefore developed a novel *agnostic semi-supervised learning* protocol, which by being more conservative than traditional semi-supervised learning, can improve the performance even when D_{prim} and D_{aux} are drawn from different distributions, and which is extremely unlikely to decrease the performance. The protocol proceeds in four steps:

1. A Naive Bayes classifier C_{prim} is trained on D_{prim} .
2. C_{prim} is used to assign labels to D_{aux} .
3. C_{aux} is learned from the newly-labeled D_{aux} .⁵
4. The predictions of C_{aux} are used as additional features in D_{prim} as described in 2.1.

We note that adding the predictions of C_{aux} as features in D_{prim} allows us to directly learn to what extent D_{aux} can improve (or hurt) parameter estimation. In our work, we used three different auxiliary datasets for ASSL feature generation: Google 5-grams collection, Wikipedia articles, and Yahoo Answers.

2.3 Structure-based Feature generation(SBFG)

This approach can be seen as an instance of the approach described in section 2.1. The difference is that each document in the auxiliary dataset D_{aux} is treated as belonging to a distinct label. Thus, this approach operates on thousands (or even millions) of labels, with a single training instance for each label. The advantage of this approach is that the auxiliary dataset can be unlabeled.

This approach was originally proposed in [5]. Given an unlabeled auxiliary dataset D_{aux} , for each data sample $x \in$

⁵When D_{aux} is orders of magnitude larger than X_{train} , this may allow more reliable estimation of the parameters.

D_{aux} , a set of characteristic features (called 'attributes' in [5]) is selected. The selection in [5] is done with TF-IDF weighting. When learning a classifier on the primary task D_{prim} , each sample $x \in D_{prim}$ is extended with the identifiers of the samples in D_{aux} which attributes are similar to the features in x based on cosine similarity score.

This approach generates a very large number of features, for example Wikipedia has over 1 million concepts, each of which can be a potential generated feature [6]. While [6] reports improved results using this approach, we believe that a technique to capture this information with significantly fewer features is necessary. We propose to use clustering of unlabeled data to reduce the number of generated features. Another advantage of our approach is that it enables real time performance, since the clustering is done off-line just once, and at feature-generation stage only a small number of clusters (summarized with top-20 TF-IDF words) need to be considered.

To reduce the number of features, the documents in the unlabeled auxiliary dataset D_{aux} are clustered without supervision with agglomerative clustering to a prescribed number of clusters⁶. Each cluster can be considered as high-level concept/feature F , and is added as an additional feature to the samples in the primary dataset which have considerable feature overlap with F .

We use the CLUTO software for unsupervised clustering of both Yahoo Answers and Sikipedia datasets into 300-500 clusters⁷. We then summarize the resulting clusters with top-20 TF-IDF words and add the Yahoo-Cluster or Wikipedia-Cluster IDs as new features during feature generation as described in section 2.1.

Table 2 shows examples of clusters found in Wikipedia and Yahoo Question Answers datasets. These clusters reduce noise in the external knowledge and make the feature-generation tractable for a real-time system. Other techniques, such as principal component analysis [1] can be used to extract high-level features from the auxiliary unlabeled dataset. This is the approach essentially proposed in [18], where a more sophisticated high-level concept extraction mechanism was used. For more details, the reader is referred to [18].

2.4 Preprocessing Auxiliary Datasets

Google 5-grams is a 33GB collection, containing all the 5-grams which appear on the Web. Many of the 5-grams do not contain legal English words, but rather contain only HTML tags. Some contain segments taken across sentence boundaries (e.g.: 'fish . He was angry'). To make feature generation using the auxiliary dataset data tractable, and in order to remove noise, we pruned the Wikipedia and the Google 5-grams datasets. We reduce Wikipedia data from 2million concepts (12 GB) to 200k concepts (1 GB) using OpenMind Indoor Common Sense project [10]. We look at all the titles and keep all concepts that have any word from OpenMind Indoor Common Sense database. The idea is to get rid of titles such as *Dirichlet distribution* which are too scientific and unlikely to occur in everyday conversations. We also prune concepts related to people names,

⁶Experimentally, good results are achieved when the number of clusters was KN , where N is the number of labels in the primary dataset and K was set to 50.

⁷CLUTO is available online at <http://glaros.dtc.umn.edu/gkhome/views/cluto>

Wikipedia clusters	Yahoo Clusters
churches neighbor wedding window cleaning dining feeding customer neighbors nursing restaurant lawn invitations seek cafe onion activities patients charge diner	recipe festival root flavor calories draft dark spices styles ale wine mild beer brand drink owned company cream steam taste
car insurance cars highway grand muscle driver road roads wheel sport streets recall mile traffic speed drive miles gas driving	bike downhill division france local wine city county district saint pop government village rural spa town map ski arms population
department drive texas california driving license florida virginia georgia district insurance arizona required state laws requirements driver education training minor	earth series rose doctor jane ninth adventure lord stories tells finish novels master fiction story television seventh
students student junior science classes activities grade college writing schools semester score act studying university colleges study senior scores sat	bowie albums tonight apple album beatles love song songs solo guitar pepper lyrics imagine tour bass strawberry track yesterday piano

Table 2: Examples of Wikipedia and Yahoo clusters

place names, concepts with very long titles (more than 40 words in title), images and redirect pages from Wikipedia. Similarly Google 5gram data is reduced from 33GB to 2.4GB by removing uncommon words that do not appear in OpenMind Indoor Common Sense database, and 5-grams that contain two or more punctuation symbols (hence 3 or fewer words). Yahoo Question Answer data is originally 120 MB and is not reduced.

3. PRIMARY DATASETS

We evaluated our approach on two datasets: the Switchboard corpus [9], a benchmark dataset for topic spotting and speech recognition, and the Google Answers dataset, a collection of short questions cataloged by topics, which we extracted from the Web⁸.

The Google Answers dataset is a collection of 8,850 questions pertaining to 9 top-level categories extracted from the Web (around 1000 questions per category). The list of the categories is given in Table 3. Before Google Answers was discontinued, Google required a payment to be made for a question which was answered by an expert in Google Answers. Therefore, typically, the questions in this dataset are directed to experts, and require a lot of prior knowledge to be correctly categorized. Below are some examples of the questions in the category Arts And Entertainment:

1. *In 1998, Henry Rollins did a spoken word engagement gig in/near Venice beach... i'd like to know the date.*
2. *Please provide general information including best photos of beautiful "antigua town" of Guatemala country.*
3. *Looking for Boy Goergoe manager's phone number.*

The Switchboard corpus is a multi-speaker corpus of conversational speech with about 2500 conversations by 500 speakers from around the US. These conversations are transcribed by speaker terms and span 70 topics (like camping,

⁸Google Answers are available online at <http://answers.google.com/answers/>. The extracted dataset will be available on the Web soon

Category name
Arts And Entertainment
Family And Home
Relationships And Society
Business And Money
Health
Science
Computers
Reference And Education And News
Sports And Recreation

Table 4: Google Answers categories

taxes and recycling). A sample of the transcribed data is given below:

- A.1: *[Laughter].*
 B.2: *Uh-huh.*
 A.3: *Um, I guess were supposed to talk about music.*
 B.4: *Okay.*
 A.5: *And, uh, let me go ahead and push one here. [tone]*
Uh, do you ha-, are you a musician yourself?
 B.6: *Uh, well, I sing.*
 A.7: *Uh-huh.*
 B.8: *I dont play an instrument.*
 A.9: *Uh-huh.*
Where, do you sing in, in a choir or a choral group?
 B.10: *Oh, not right now.*

The Switchboard dataset was previously used for classifying spoken text in [15]. Following [15], we manually map selected topics to ten categories as shown in Figure 3, and consider the task of classifying each individual speaker turn to one out of 10 categories (see [15] for details). Our 10 category corpus has 46,000 utterances, giving an average of 71.7 utterances per conversation. Since many of speaker turns do not carry meaningful information, We filter out stop words and sentences that contain less than 10 words to get approximately 24,000 sentences. To test whether the resulting dataset can be successfully classified, we asked two human annotators to annotate the sentences. We describe

Category	Switchboard categories	Number of conversations
books	books and literature; magazines	48
fitness	exercise and fitness	63
movies	movies	56
pets	pets	63
sports	football; baseball; basketball	69
family	family life; family reunions; care of the elderly	153
food	recipes food and cooking	51
music	music	58
restaurants	restaurants	41
weather	weather climate	39

Table 3: Number of conversations in different categories in our corpus

Category	Switchboard Label	Annotator I	Annotator II
books	47	34	47
fitness	41	40	42
movies	39	44	39
pets	38	35	38
sports	72	71	72
family	99	117	98
food	45	58	46
music	43	40	43
restaurants	46	31	45
weather	30	29	30

Table 5: Labels given by two annotators on 500 randomly selected sentences from transcripts. Same numbers in a category do not imply same labeling

the result in section 4.

4. INTER ANNOTATOR AGREEMENT

Agreement among annotators gives us an estimate of consistency among people on this classification task. We select a random set of 500 sentences from the 24,000 transcript sentences and ask two annotators to categorize them into the ten categories.

Table 5 shows the number of sentences in each category labeled by both annotators. The strength of the agreement among the annotators is measured using the Fleiss’ kappa score [4]. The Fleiss’ kappa score of 79.2% reflects substantial agreement among the annotators [14]. Measuring human inter-annotator agreement for the switchboard data allowed us to determine a reasonable level of performance we can expect from our system for this dataset.

5. WRITTEN TOPIC SPOTTING RESULTS

As discussed in section 3 we apply our approach to two domains: the Switchboard corpus and the Google Answers. Our auxiliary datasets include:

1. A raw Yahoo Answers dataset, which is used to train an ASSL classifier, *ASSL(Yahoo)*.
2. Yahoo Answers, organized by subcategories, which are used to train a PAF classifier, *PAF(Yahoo)*.
3. Yahoo Answers, automatically clustered into 300 clusters, which are used to train an SBFGE classifier:

SBFG(Yahoo).

4. Pruned raw Wikipedia articles dataset, which is used to train an ASSL classifier, *ASSL(Wiki)*.
5. Pruned Wikipedia articles automatically clustered into 500 clusters, which are used to train an SBFGE classifier, *SBFG(Wiki)*.
6. Pruned Google 5-grams collection, which is used to train an ASSL classifier, *ASSL(Google)*.

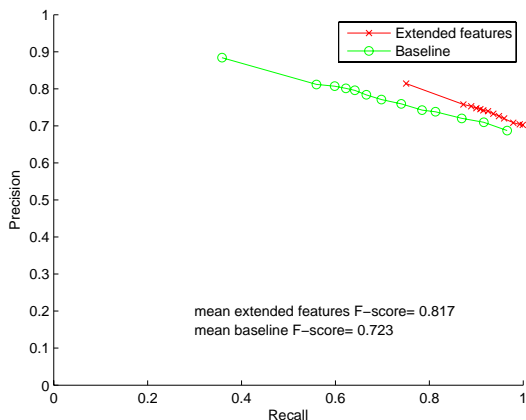
The results for using each of the six resulting classifiers and their combination on the two primary classification datasets are summarized in Table 6. Note that while each individual feature extension method only marginally improves the performance, the net effect is 17% error reduction on the Switchboard dataset. The results suggest that feature extension improves classification accuracy.

In many applications, such as call placement, the cost of error is high (annoyed customers who reach the wrong extension), therefore, it is common to refuse making an automated decision, and redirect the call to human operator. Therefore, it is important to be able to increase the precision of the classifier by paying the price of reduced recall (refusal to make automated decision). Figure 5 compares the precision/recall curves for the baseline classifier and the proposed approach. The results indicate that the proposed approach allows a significantly better increase in precision while paying a smaller recall penalty relative to the baseline classifier. Note that the mean F1 score for the classifier trained with the extended features is higher than that of a baseline classifier.

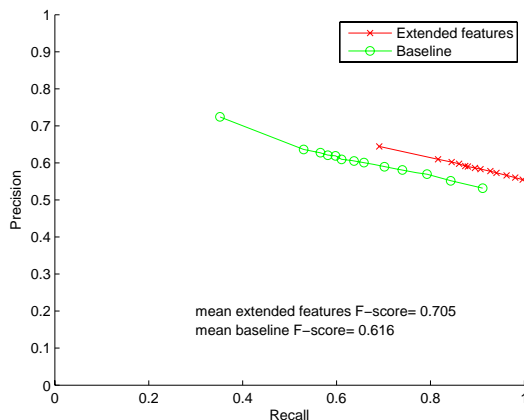
Finally, we have analyzed the concrete utterances which are miss-classified by the baseline classifier, but are correctly classified using the feature generation approach. Table 7 shows examples from Google Answers dataset that are misclassified in the baseline but corrected in the classifier using extended features. Table 8 shows examples from the Switchboard dataset. In both datasets common sense knowledge is needed to imply the correct label.

6. SPOKEN TEXT RECOGNITION

We implemented our approach in a speech recognition system. In our audio tests we used 641 conversations of about 6 minutes each from the Switchboard dataset. We split each conversation into 32 parts. We used these 20,512 audio clips in 5-fold topic classification experiments using audio data without the transcripts. We used the off-the-shelf



(a) Switchboard.



(b) Google Answers.

Figure 5: Comparison of the precision/recall tradeoff on the Switchboard and the Google datasets for the baseline classifier and the classifier that uses the extended features. The extended features allow to increase precision with less sacrifice in recall. The mean F1 score for the classifier trained with the extended features is higher than that of a baseline classifier.

Algorithm	Accuracy/ (Err. Reduction) on Switchboard	Accuracy/ (Err. Reduction) on Google Answers
<i>Baseline</i>	66.3/(0.0)	49.77/(0.0)
<i>ASSL(Wiki)</i>	66.71/(1.21)	50.72/(1.89)
<i>SBF(Wiki)</i>	68.43/(6.33)	50.81/(2.07)
<i>SBF(Yahoo)</i>	69.38/(9.15)	50.44/(1.32)
<i>PAF(Yahoo)</i>	69.09/(8.28)	50.93/(2.29)
<i>ASSL(Yahoo)</i>	68.96/(7.89)	50.98/(2.4)
<i>ASSL(Google)</i>	68.09/(5.33)	50.35/(1.14)
<i>Combined</i>	72.33/(17.91)	53.06/(6.54)

Table 6: Results of 5 fold cross validation. While each individual feature extension method only marginally improves the performance, the net effect is 17% error reduction on Switchboard data.

Nuance Speech Recognition system with a Statistical Language Model with large vocabulary grammar of 13917 words from transcriptions of the switchboard corpus. Both the switchboard and Yahoo Answers are used in computation of trigram probabilities. Nuance Speech Recognition is used with largely untuned parameters, hence baseline recognition could be further improved. Below we list a sample set of utterances from the category *books* as they are transcribed by the speech recognition system. It is hard even for the human to classify the noisy transcriptions.

- *does anyone else for what was the decision to listen to rap ...*
- *uh there is a magazine that like that its real @reject@ not ...*
- *and uh was called la want to just go through your exercise um ...*
- *return someone lives but women feel was one of the time and um uh ...*
- *@reject@ true well um our society yeah that but it does anyone ...*

- *yeah anytime i really enjoyed the magazine about i really dark ...*
- *and i @reject@ know i am having a party doll we had @reject@ ...*

[15] reports 15% accuracy on this dataset, which is only marginally better than majority classification, which leads to 13% accuracy. Our baseline system resulted is 28.47%, and with using external knowledge, the result improves to 29.97%. The low increase in accuracy using extended features can be explained by the fact that common-sense background knowledge is hard (if possible) to apply on such noisy transcripts.

7. CONCLUSIONS

In this paper we proposed a feature-generation approach to knowledge transfer. We combine various sources of unlabeled and labeled data to make human-like decisions that incorporate world knowledge. We proposed two novel feature generation algorithms and leverage selected existing techniques to make large scale use of multiple external knowledge sources in a single system. While each individual approach does not lead to a significant performance improvement, when a large array of feature-generation techniques are used with massive external data sources, the net effect of the feature-generators is significant.

We tested the proposed approach for classification of short conversational snippets and text data in two domains (Switchboard transcripts and Google Answers). We built a real time system for dialogue classification that classifies short utterances with significantly higher accuracy than previously published work on the same data [15]. For text transcripts of the switchboard corpus we obtain up to 17% error reduction using external knowledge.

8. REFERENCES

- [1] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

Text snippet	Baseline category	Extended Feature category
is ornette coleman's "fifth of beethoven" based on beethoven's fifth symphony in some way? if so, how?	Business and Money	Arts and Entertainment
i'm looking for suppliers of educational software from year 4 to year 12, subjects;maths, science, physics for schools in melbourne, australia.	Relationships and Society	Reference and Education and News
where can i get information on used car sales on a state to state basis listing make model milage and selling price for the purpose of comparison with other states (as in buy low and sell high)	Sports and Recreation	Business and Money
how much of the war on terror is real, based on hard evidence, and how much is made up by politician's, to scare us into supporting them and reducing our civil liberties? also, how can we wage war on a noun?	Science	Relationships and Society
what is the diameter of a cadmium atom in angstroms? this may or may not be an odd unit of measurement for atomic sizes , but i 'm trying to get an idea of the relative size between the two. an angstrom equaling 1 x 10-8 centimeters , which would equal roughly .00000004 inches (if my math isn 't too shot). please correct any errors. thanks, patrice	Computers	Science
stop birds nesting on my roof	Science	Family and Home

Table 7: Examples of questions in Google Answers, for which the right category is determined using external data.

- [2] R. Caruana. Multitask learning. In *Machine Learning*, 28, 41-75, 1997.
- [3] F. G. Cozman, I. Cohen, and M. C. Cirelo. Semi-supervised learning of mixture models and bayesian networks. In *International conference of Machine Learning (ICML)*, 2003.
- [4] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.
- [5] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, August 2005.
- [6] E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with. In *Proceedings of the Association for Advancement of Artificial Intelligence (AAAI)*, July 2006.
- [7] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, January 2007.
- [8] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL)*, July 7-12 2003.
- [9] J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 517–520, March 23-26 1992.
- [10] R. Gupta and M. Kochenderfer. Common sense data acquisition for indoor mobile robots. In *Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, July 25-29 2004.
- [11] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23:33–64, March 1997.
- [12] H. D. III and D. Marcu. Domain adaptation for statistical classifiers. In *J. Artificial Intelligence*, 26:101-126, 2006.
- [13] J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of ACL*, 2007.
- [14] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.
- [15] K. Myers, M. Kearns, S. Singh, and M. A. Walker. A boosting approach to topic spotting on subdialogues. In P. Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 655–662, 2000.
- [16] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.
- [17] L. Pevzner and M. A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, March 2002.
- [18] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the twenty-fourth International Conference on Machine Learning (ICML)*, 2007.
- [19] R. Raina, A. Y. Ng, and D. Koller. Transfer learning by constructing informative priors. In *Proceedings of the Twenty-third International Conference on Machine Learning (ICML)*, 2006.

Transcript	Baseline category	Extended Feature category
yes since TI has uh uh instituted that um uh walkabout i have gone out and started walking even more bought me the you know the proper shoes and everything to get started and uh park out there you know way out there in the boondocks	sports	fitness
well i think we've we've covered the subject i've got some interesting information about crawfish i was in fact i was it's good cause i was curious about that today when i saw those for sale	family	food
and uh it kind of makes me listen a little closer to to the piano when i hear it at other places or when i you know see some see and hear someone playing i sort of watch their technique too and and see if it corresponds up to what she's learning	family	music
drop a ball that cost a couple runs ended up costing the game so he could you know he could be both brilliant and with a bat and just a disaster in the field	fitness	sports
okay okay well i'm not to say that all folks from Brooklyn are thugs but these two were definitely thugs and they were from Brooklyn so i'm kind of hoping i i i guess uh anymore i pretty much pull for the Rangers though they were uh they're they're they're they they have the i think they have the best facilities in the major leagues	family	sports
right right you i've seen hail you know but they're usually the size you know of a tiny tiny pebbles you know really small you know but this was uh really large	sports	weather

Table 8: Example transcripts in Switchboard, where the right category is determined using external data

[20] A. F. Thorsten Brants. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, 2006.