Science	health	Sports & Recreation	Business & Money
explosive material	Top Games Card-Games	Entertainment & Music-Horoscopes	Business & Finance-Australia & New Zealand
spacecraft propulsion	Dining Out-Halifax	Top Shopping Tools	Business & Finance-Marketing & Sales
gravity	medicine	Top Shopping Tobacco	Top Business Healthcare
hydrogen	Top Games Dice	Top Reference Maps	Top Business Mining-and-Drilling
water (molecule)	Dining Out-Gualeguaychu	Education&Reference-Primary&Secondary Education	bank
black hole	prostate cancer	Top Shopping Health	Citigroup
sewage treatment	hormone replacement therapy trans	sports timeline	Top Arts Video
temperature	cancer	DiningOut-Zurich	investment bank
sw	non-hodgkin's lymphoma	Entertainment & Music-Blues	mergers and acquisitions
titanium	Dining Out-Hamburg	Entertainment & Music-Jokes	mutual fund

Table 4: Top 10 features using Mutual Information for Google Answers dataset. Best viewed in color. ESA features are shown green, PAF(Yahoo) features are shown red. PAF(ODP) features are shown blue. No bag-of-words features were selected in this case

Results

As discussed previously, we apply our approach to two datasets: the Switchboard corpus and the Google Answers. Our features come from:

1. **BOW**: bag of words after stop-word removal.
2. **ESA**: Features extracted from Wikipedia using ESA.
3. **PAF(Yahoo)**: Yahoo Answers, organized into 822 sub-categories.
4. **PAF(ODP)**: Open Directory Project, organized into 494 sub-categories.
5. **SBFG(Wiki)**: Pruned Wikipedia articles automatically clustered into 500 clusters.
6. **SBFG(Yahoo)**: Yahoo Answers, automatically clustered into 300 clusters.
7. **SBFG(ODP)**: Open Directory Project, automatically clustered into 500 clusters.

The results from using different feature combinations on the two primary classification datasets are summarized in this section. Figure 1 shows variation of accuracy with number of features/class for the Google Answers dataset. Note that the x axis is on a logarithmic scale. We note that Google dataset has 112052 total features from all techniques with 11430 BOW features. The best results are obtained using top 200-500 features. Using features from all six auxiliary data technique pairs with bag of words (BOW) shows substantial improvements over the baseline technique using only BOW and BOW+ESA techniques. Switchboard dataset with a total of 85703 features from all techniques has similar results.

The performance of different feature generation techniques for top 200 features/class are shown in Table 5. BOW+ESA+PAF refers to all techniques using words and category data from all auxiliary datasets, and BOW+SBFG refers to all techniques using words and clusters based on all auxiliary datasets. Using all six auxiliary data technique pairs with BOW features gives a 64% error reduction in Switchboard dataset and 48% error reduction in the Google Answers dataset over the BOW+ESA(Wiki) baseline technique.

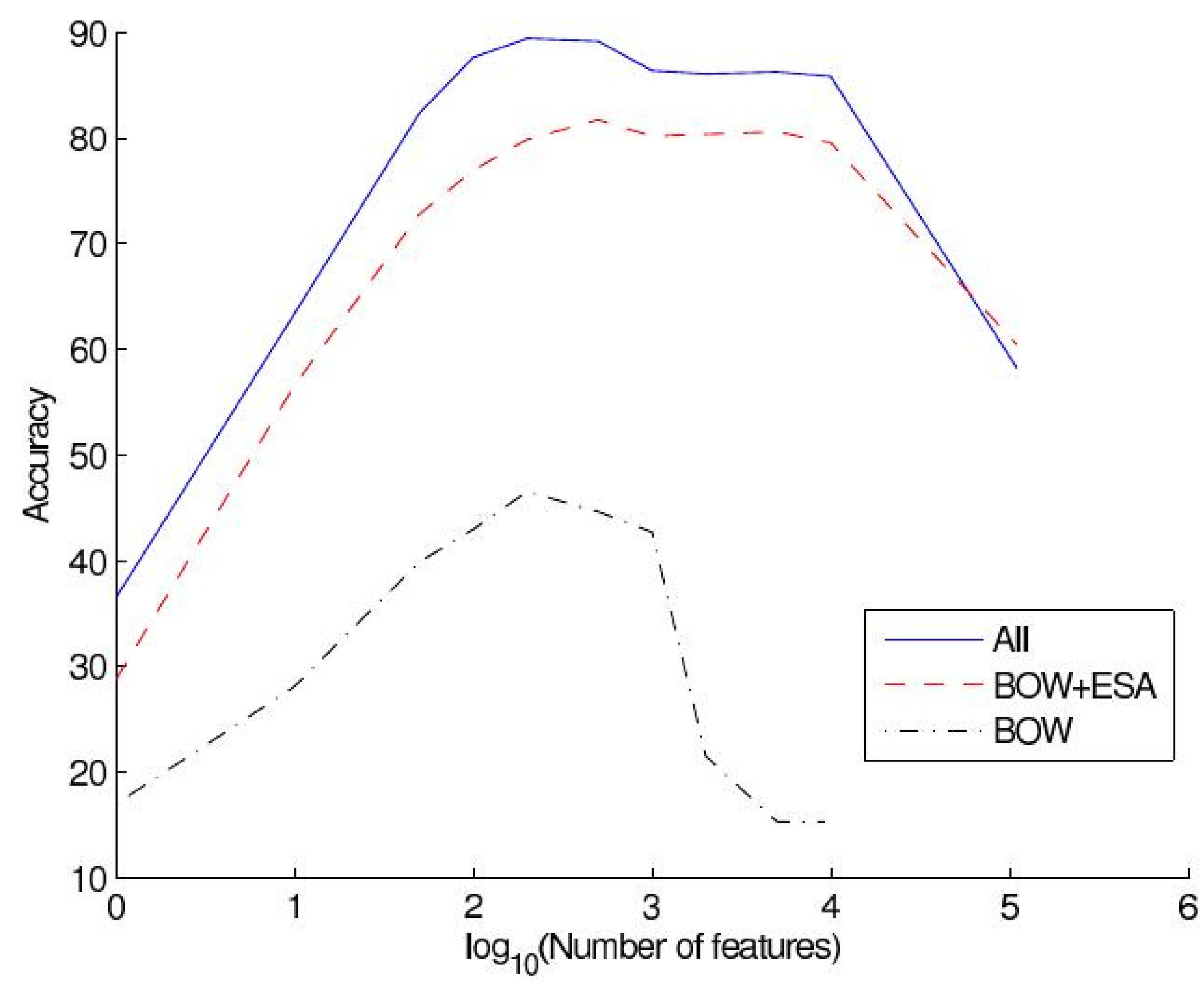

Figure 1: Variation of accuracy with number of features/class. Note that X axis is on log scale

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. It is important to increase the precision of the classifier by paying the price of reduced recall (refusal to make automated decision). Figure 2 compares the precision/recall curves for the baseline classifier and our approach for Google Answers dataset. 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 the baseline classifier.

Conclusions

In this paper we propose a feature-generation approach to knowledge transfer. We combine various sources of unlabeled and labeled data to make human-like decisions that