An Information Theoretic Approach for Using Word Cluster Information in Natural Language Call Routing

Li Li, Feng Liu, Wu Chou

Avaya Labs Research, 233 Mt. Airy Rd., Basking Ridge, NJ 07920
{lli5, fengliu, wuchou}@research.avayalabs.com

10.21437/Eurospeech.2003-753

Abstract

In this paper, an information theoretic approach for using word clusters in natural language call routing (NLCR) is proposed. This approach utilizes an automatic word class clustering algorithm to generate word classes from the word based training corpus. In our approach, the information gain (IG) based term selection is used to combine both word term and word class information in NLCR. A joint latent semantic indexing natural language understanding algorithm is derived and studied in NLCR tasks. Comparing with word term based approach, an average performance gain of 10.7% to 14.5% is observed averaged over various training and testing conditions.

1. Introduction

Natural language call routing (NLCR) is to determine the caller's intension and transfer the caller to the desired destination through natural language based dialogue interaction. It is an application to use natural language to improve the service quality for interactive voice response (IVR), which is traditionally implemented by using highly constrained finite-state grammars derived from the service manual. Natural language call routing is related to natural language understanding and information retrieval, and it is well known that literally matching word terms in a user's query to a destination description can be a problem. This is because there are many ways to express a given concept (synonym), and the literal terms in a query may not match those of a relevant destination (document). This leads to the study and application of various natural language understanding and information retrieval techniques in NLCR, such as latent semantic indexing [2,3,4].

In natural language processing, word term classes (or clusters) are formed by clustering word terms that have some common properties or similar semantic meanings. They are regarded as more robust than word terms, because word class clustering process can be viewed as a mapping of the surface form representation in word terms to some generic concepts that should be more stable. One problem associated with word classes is that they may not be detailed enough to differentiate confusion cases in the NLP task. To effectively apply word classes in NLP is a challenging problem. This is because not all word classes are robust, especially when speech recognition is involved. Moreover, using word classes alone is not very good either, because it may lack some critical details in word terms that are needed to separate fine differences. In addition, word class clustering in NLP is an active research area of its own right. Most word class clustering is based on linguistic information or task dependent semantic analysis, which involves manual intervention, a costly, error prone and labor-intensive process.

In this paper, we study and provide solutions to the following two critical issues in applying word class in natural language call routing. The first one is how to apply a data driven automatic word class clustering method in NLCR where word classes are automatically derived from the word term statistics in the targeted NLP task without resolving to semantic knowledge based manual supervision. The second issue is how to combine both word classes and word terms information to enhance the robustness and performance in a natural language understanding task, such as NLCR.

The organization of this paper is as follows. We describe in Section 2 a data driven automatic word term clustering algorithm in our study that is applied to NLCR. In Section 3, we present an information theoretic approach to combine both word class and word term information for robust natural language understanding. The proposed approach is applied to call routing based on latent semantic indexing (LSI). Section 4 is devoted to experimental and comparative studies, where the proposed approach is studied in NLCR tasks. Findings and performance advantage of the proposed approach is summarized in Section 5.

2. Automatic Word Class Clustering

Data driven automatic word class clustering has the advantage of not requiring human intervention and the semantic knowledge of the task. But the original motivation of using automatic word class clustering is mainly from the need of building language model for speech recognition. In the language model for speech recognition, the number of possible n-grams is usually enormous, and even with a large training corpus, a significant number of events, i.e., word pairs and word triples, are rarely or never seen in the training data. Clustering words into equivalent classes, or word classes, can increase the number of observations and easy to generalize to unseen events. This leads to the word class based language model in which the probability of a given word depends on its class and on the classes of the preceding words [1,5].

In our study, we adopt the automatic word clustering algorithm used in [1] as a mean to automatically generate word classes for NLCR. Given the word term size W, the algorithm partitions word terms into a fixed number of word classes. The partition and word grouping is to find a class mapping function $G: w \rightarrow g_w$, which maps each word term w to its word class g_w such that the perplexity of the associated class based language model is minimized on the training corpus. The algorithm employs a technique of local

optimization by looping through each word in the vocabulary, moving it tentatively to each of the G word classes, searching for the best class membership assignment that gives the lower class based language model perplexity. The whole procedure is repeated until a stopping criterion is met. This so-called exchange algorithm works as follows [1]:

```
Set up initial word class mapping
    Compute the mapping perplexity on the training
    corpus
     Do until some stopping criterion is met
       Do for each word w in vocabulary W
         Remove w from class g_w
           Do for all existing classes g:
             Compute perplexity as if w were moved to g
           end-do-loop
         Assign w to the class with the lowest perplexity
       end-do-loop
    end-do-loop
Exit
```

The perplexity (PP) of the class based language model can be calculated as follows:

$PP = 2^{LP}$, where LP can be estimated as

$$LP = \frac{-1}{T}[\sum_{w} N(w)\log N(w) + \sum_{g_w, g_v} N(g_w, g_v)\log \frac{N(g_w, g_v)}{N(g_w)N(g_v)}]$$

where T is the length of the training text, and $N(\cdot)$ is the number of occurrences of event given in the parentheses in the training data.

This is a data driven statistical clustering method. The number of word classes is a design parameter to control the word class clustering. However, this method is used in building class based language model, and it is not related to NLCR. Many of the clusters generated in this automatic process can be poorly formed. In order to make a meaningful use of these word classes, a well-founded statistical framework is needed to robustly integrate word class information in NLCR.

3. An Information Theoretic Framework for Using Word Clusters in NLCR

As pointed out in the introduction, in natural language processing, not every word term has detailed information that is salient for natural language understanding, and not every word class is robust and useful either. This situation is more acute with word classes obtained from a data driven automatic clustering process. In this section, we first present an information theoretic framework to select salient word terms and word classes for natural language understanding. Then we show how this approach can be used in latent semantic indexing (LSI) based natural language understanding to improve the robustness and the performance of NLCR.

3.1. An information theoretic term selection framework for NLCR

NLCR is to classify the user query, which is a sequence of word terms, into one of the N categories (or destinations). Among various term selection criteria, information gain (IG) based term selection is very unique. It provides an information theoretic framework to term selection. In IG based term selection approach, the IG score of a term is the degree of certainty gained about which category is "transmitted" when the term is "received" or not "received." The significance of the term is determined by the average entropy variations on the categories, which relates to the perplexity of the classification task.

The IG score of a term t_i , $IG(t_i)$, is calculated according to the following formulas:

$$IG(t_i) = H(C) - H(C \mid t_i) - H(C \mid \bar{t_i}) \quad (1)$$

$$H(C) = -\sum_{j=1}^{n} p(c_j)\log(p(c_j)) \quad (2)$$

$$H(C \mid t_i) = -p(t_i)\sum_{j=1}^{n} p(c_j \mid t_i)\log(p(c_j \mid t_i)) \quad (3)$$

$$H(C \mid \bar{t_i}) = -p(\bar{t_i})\sum_{j=1}^{n} p(c_j \mid \bar{t_i})\log(p(c_j \mid \bar{t_i})) \quad (4)$$

where n is the number of categories, and

- $H(C)$: the entropy of the categories
- $H(C \mid t_i)$: the conditional category entropy when t_i is present
- $H(C \mid \bar{t_i})$: the conditional entropy when t_i is absent
- $p(c_j)$: the probability of category c_j
- $p(c_j \mid t_i)$: the probability of category c_j given t_i
- $p(c_j \mid \bar{t_i})$: the probability of c_j without t_i

The right side of Formula (1) can be transformed to the following formula [4]:

$$\sum_{j=1}^{n}\left[\begin{array}{l} p(t_i c_j)\log\left(\frac{p(t_i c_j)}{p(c_j)p(t_i)}\right) + \\ (p(c_j) - p(t_i c_j))\log\left(\frac{p(c_j) - p(t_i c_j)}{p(c_j)(1-p(t_i))}\right)\end{array}\right]$$

where we have:

- $p(t_i)$: the probability of term t_i
- $p(t_i c_j)$: the joint probability of t_i and c_j

The IG based term selection provides a unified approach to select salient features from multiple information sources. It is applied in [4] for word term selection in NLCR application. We show in the rest of this section that a new joint classifier based on both word term and word class information.

3.2. Joint word term and word class based LSI algorithm

In this subsection, we describe the approach and implementation of an IG enhanced joint word term and word class based LSI classifier. The focus is on the joint word terms and word class IG extension part in the proposed approach and we refer to [2,3,4] for other details of LSI based classifier.

The training corpus for LSI based classifier is a collection of documents with corresponding categories labeled. It is usually first processed by a linguistic analysis module to convert words in the document into a sequence of raw terms. This module is often based on morphological rules, such as Porter stemming, and linguistic resources such as root dictionary, ignore word list and stop word list, etc. The terms used in LSI analysis can be based on term unigram, bigram and trigram that correspond to raw terms, raw term pairs and raw term triplets.

The joint word term and word class LSI classifier in our approach is based on the union of terms obtained from word terms as in standard LSI, and the terms obtained from the word classes. The terms from the word classes are obtained from the training corpus that maps each word w into its corresponding word class. The word class mapping can be constructed by hand, by automatic clustering, or by the mixture of both. The automatic clustering algorithm described in Section 2, makes it possible that such a joint LSI classifier is always achievable even without any side information on the task. The inclusion of terms from multiple resources leads to a huge increase in the dimension of term-category matrix in the LSI classifier, and many of them can be very noisy and of poor quality. The IG based term selection is critical here to cut down the term space dimension and only the top percentile terms are selected based on the joint word term and word class information. In our approach, this process follows the information gain based information theoretic framework that is well suited for NLCR. It eliminates heuristic term selection procedures, and the whole process can be made automatic or semi-automatic.

The term-category matrix M in our approach is formed by terms from IG based joint term selection. It can be word terms or word classes, depending on the IG score which describes the discriminative information of the term in NLCR task. The $M[i,j]$ cell of the term-category matrix is the sample count that the i-th selected term occurs in j-th category. An $m \times k$ term matrix T and a $n \times k$ category matrix C are derived by decomposing M through the SVD process, such that row $T[i]$ is the term vector for the i-th term, and row $C[i]$ is the category vector for the i-th category as typical in LSI based approach [2-4].

In our proposed approach, terms are selected and used in the term-category matrix based on their discriminative power according to IG criterion given the joint information of both word terms and word classes. It consists of the following steps: (1) sort all (word, word class) terms by their IG values in a descending order; (2) select top p percentile of terms according to the IG score distribution; (3) construct the term-category matrix and perform LSI analysis based on terms selected from (2). To categorize an unknown document, the user input is processed into a sequence of words. It is mapped to a query vector X according to the order and mapping from word sequence to each selected terms in the joint word term and word class LSI classifier. If both word w and its word class g_w are selected by the joint IG based term selection process, both entries in the query vector will have non-zero term counts.

Before leaving this section, we would like to point out that the proposed IG enhanced joint LSI classifier approach apply to the case of having more than one word class mappings, and to the case of using multiple raw term resources beyond word classes. The joint word term and word class LSI classifier is established for all applications, because the word class generation process can be made automatic without any linguistic or task dependent knowledge.

4. Experimental and Comparative Studies

Experimental studies were performed on a natural language call routing task based on the transcriptions of spoken dialog data [4]. The training session of the database consists of 3510 training documents in 23 categories. The independent test session consists of 307 test documents in 21 categories (2 categories are not observed). The average document length, counted by words, is 8.1 in the training set and 14.5 words in the testing set. The number of documents in each category is highly unbalanced. The standard deviation of number of documents in each category is 33.09 for the test set and it is 393.76 for the training set.

4.1. Experimental study setup

In order to study the effectiveness and robustness of natural language understanding based on cluster-boosted LSI classifier, experiments were conducted on two focus group studies: 1) reduced training data size; and 2) reduced linguistic processing. In the focus group study of reducing training data size, two sub-corpora of 1755 (sub-corpus-1) and 1404 (sub-corpus-2) utterances each were constructed from the original training corpus of 3510 transcribed spoken dialog utterances. The original disjoint test corpus of 307 utterances was used throughout all experiments.

In all experiments, three type of LSI classifiers were constructed:

- baseline: LSI classifier trained and tested on the word corpus;
- cluster: LSI classifier trained and tested based on word class corpus;
- joined: LSI classifier trained and tested on the joined corpus.

The word class corpus was obtained by mapping words in the corresponding word corpus into their word classes. 10 different IG term selection thresholds varying from 1% to 40% were used to study the behavior of the classifier. The classifier performance comparison is based on recall scores [4].

The word classes are generated automatically based on the algorithm described in Section 2. It was applied to each of the three training corpuses (sub-corpus-1, sub-corpus-2, original-corpus) without linguistic preprocessing. Different number of word classes, which was a parameter in the data driven

automatic word clustering algorithm, were chosen for each training corpus based on some initial studies. Number of word classes depends on the size of the training data, and the following number of word classes was used in the experiments:

Training/Test	# Cluster
3510/307	802
1755/307	532
1404/307	472

In the second focus group study of reducing linguistic pre-processing resources, three conditions were considered which corresponds to different levels of linguistic resource dependencies: (1) using all linguistic resources, including ignore list, stop list, roots, and Porter stemming; (2) using only Porter stemming; (3) not using any linguistic resource. This study would indicate the portability of the classifier for different NLCR tasks.

4.2. Experimental results and comparisons

One issue in classifier performance comparison is that for the same IG term selection threshold, the number of selected terms (words and clusters) in each of the three LSI classifiers (baseline, cluster and joined) was quite different. This is because they were based on different pools of terms. Therefore, it may not be fair to do the pair wise recall scores comparison between classifiers with the same IG threshold. We took the approach of classifier performance comparison based on the average performance of the classifier over its normal IG threshold operating range of 1% to 40%. However, a standard arithmetic mean is not always fair, especially for the baselines under low IG thresholds when too few terms are selected and some documents are not analyzed at all. Therefore, two alternative methods were used to measure the average error rate reduction on recall for each cluster-boosted classifier against the others:

- Truncated mean: the mean of the 10 recalls is calculated after the max and min are removed; by this method, the cluster-boosted LSI has a 14.8% average error reduction over the baselines.

- Quartile mean: the mean of the first, second and third quartiles of the 10 recalls is used; by this method, the cluster-boosted LSI has a 10.7% average error reduction over the baselines.

These experimental results are tabulated in Fig. 1 and Fig. 2 for the two focus group studies.

5. Summary

In this paper, an information theoretic approach for using word clusters in natural language call routing was opposed. This approach utilizes an automatic word class clustering algorithm to generate word classes from the word based training corpus. The information theoretic approach based on information gain (IG) was used to combine both word term and word class information in NLCR. A joint latent semantic indexing natural language understanding algorithm was derived. Comparing with word term based approach, an average performance gain of 10.7% to 14.5% were observed over various training and testing conditions.

training	linguistics	cluster	joined
3510	no ling	3.30%	6.18%
	stemmer only	5.26%	15.34%
	all ling	-7.80%	17.89%
1755	no ling	-26.21%	3.61%
	stemmer only	-24.70%	9.52%
	all ling	-34.29%	23.08%
1404	no ling	-14.93%	6.67%
	stemmer only	-16.17%	13.75%
	all ling	-3.11%	37.56%
average		-13.18%	14.84%

Figure 1: Error Rate Reductions over baselines by truncated mean

training	linguistics	cluster	joined
3510	no ling	5.38%	7.53%
	stemmer only	5.16%	16.20%
	all ling	-18.12%	9.73%
1755	no ling	-24.61%	5.71%
	stemmer only	-24.85%	9.86%
	all ling	-59.04%	9.55%
1404	no ling	-17.34%	3.83%
	stemmer only	-15.23%	14.87%
	all ling	-32.58%	19.17%
average		-20.14%	10.72%

Figure 2: Error Rate Reductions over baselines by quartile mean

6. References

[1] Martin, S., Liermann, J. and Ney, H., "Algorithms for bigram and trigram word clustering", *Speech Communication* 24(1998) 19-37.

[2] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer T.K. Harshman, R., "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, 41:391—407, 1990.

[3] Chu-Carrol, J. and Carpenter B., "Vector-Based Natural Language Call Routing", *Computational Linguistics*, 25(3):361—389, 1999.

[4] Li, L. and Chou, W., "Improving Latent Semantics Indexing Based Classifier with Information Gain", *Proc. of the 7th International Conference on Spoken Language Processing,* 2:1141-1144, Sept. 2002.

[5] P.F.Brown, V.J.Della Pietra, P.V. deSouza, J.C.Lai, and R.L.Mercer. "Class-based n-gram models of natural language", *Computational Linguistics*, 18(4), 1992.

[6] Gorin, A.L., "Processing of Semantic Information in Fluently Spoken Language**",** *Proc. of the 4th International Conference on Spoken Language Processing*, 2:1001—1004, 1996.