Semantic class induction and its application for a Chinese voice search system

Yali Li
ThinkIT laboratory, Institute of Acoustics, Chinese Academy of Sciences
liyali@hccl.ioa.ac.cn

Weiqun Xu
ThinkIT laboratory, Institute of Acoustics, Chinese Academy of Sciences
xuweiqun@hccl.ioa.ac.cn

Yonghong Yan
ThinkIT laboratory, Institute of Acoustics, Chinese Academy of Sciences
yyan@hccl.ioa.ac.cn

Abstract

In this paper, we propose a novel similarity measure based on co-occurrence probabilities for inducing semantic classes. Clustering with the new similarity measure outperformed that with the widely used distance measure based on Kullback-Leibler divergence in precision, recall and F1 evaluation. We then use the induced semantic classes and structures by the new similarity measure to generate in-domain data. At last, we use the generated data to do language model adaptation and improve the result of character recognition from 85.2% to 91%.

1 Introduction

Voice search (e.g. Wang et al., 2008) has recently become one of the major foci in spoken dialogue system research and development. In main stream large vocabulary ASR engines, statistical language models (n-grams in particular), usually trained with plenty of data, are widely used and proved very effective. But for a voice search system, we have to deal with the case where there is no or very little relevant data for language modeling. One of the conventional solutions to this problem is to collect and use some human-human or Wizard-of-Oz (WOZ) dialogue data. Once the initial system is up running, the performance can be further improved with human-computer data in a system-in-the-loop style. Another practical approach is to handcraft some grammar rules and generate some artificial data. But writing grammars manually is tedious and time-consuming and requires some linguistic expertise.

In this paper, we introduced a new similarity measure to induce semantic classes and structures. We then generated a large number of data using the induced semantic classes and structures to make language model adaptation. At the end, we give the conclusion and implied the future work.

2 Semantic Class Induction

The studies on semantic class induction in spoken language (or spoken language acquisition in general) have received some attention since the middle 90's. One of the earlier works is carried out by Gorin (1995), who employed an information -theoretic connectionist network embedded in a feedback control system to acquire spoken language. Later on Arai et al. (1999) further studied how to acquire grammar fragments in fluent speech through clustering similar phrases using Kullback-Leibler distance. Meng and Siu (2002) proposed to semi-automatically induce language structures from unannotated corpora for spoken language understanding, mainly using Kullback-Liebler divergence and mutual information. Pargellis et al. (2004) used similar measures (plus three others) to induce semantic classes for comparing domain concept independence and porting concepts across domains. Potamianos (2005, 2006, 2007) and colleagues conducted a series of studies to further improve semantic class induction, including combining wide and narrow context similarity measures, and adopting a soft-clustering algorithm (via a probabilistic class-membership function).

2.1 Clustering

In general, words and phrases which appear in similar context usually share similar semantics. E.g., 清华大学(Tsinghua University) and 北京大学(Peking University) in the following two utterances (literal translations are given in brackets) are both names of place or organisation.

请 找 **清华大学** 附近 的 银行。
Please/look for/Tsinghua University/near//bank
(Please look for banks near Tsinghua University.)

请 找 **北京大学** 附近 的 体育馆。
Please/look for/Peking University/nearby//gym
(Please look for gyms near Peking University.)

To automatically discover that the above two words have similar semantics from unannotated corpus, we try unsupervised clustering based on some similarity measures to induce semantic classes. Further details about similarity measures are given in section 2.2.

Before clustering, the utterances are segmented into phrases using a simple maximum matching against a lexicon. Clustering are conducted on phrases, which may be of a single word.

2.2 Similarity Measures

For lexical distributional similarity, several measures have been proposed and adopted, e.g., Meng and Siu (2002), Lin(1998), Dagan et al. (1999), Weeds et al. (2004).

We use two kinds of similarity measures in the experiments. One is similarity measure based on distance, and the other is a new similarity measure directly using the co-occurrence probabilities.

2.3 Distance based similarity measures

The relative entropy between two probability mass functions $p(x)$ and $q(x)$ is defined by (Cover and Thomas, 2006) as:

$$D(p \| q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(x)}{q(x)} \quad (1)$$

The relative entropy, as an asymmetric distance between two distributions, measures the inefficiency of assuming that the distribution is q when the true distribution is p.

It is commonly used as a statistical distance and can be symmetry as follows:

$$div(p,q) = D(p \| q) + D(q \| p) \quad (2)$$

For two words in a similar context, e.g., in the sequence $\{..., w_{-1}, w, w_1, ...\}$, where w can be word a or b, the right bigram $D_1(a^R \| b^R)$ and $D_1(b^R \| a^R)$ are defined as:

$$D_1(a^R \| b^R) = \sum_{w_1 \in W} p(w_1 | a) \log \frac{p(w_1 | a)}{p(w_1 | b)} \quad (3)$$

and

$$D_1(b^R \| a^R) = \sum_{w_1 \in W} p(w_1 | b) \log \frac{p(w_1 | b)}{p(w_1 | a)} \quad (4)$$

where W is the set of words or phrases.

And the symmetric divergence is

$$div_1(a^R, b^R) = D_1(a^R \| b^R) + D_1(b^R \| a^R) \quad (5)$$

The left bigram symmetric divergence can be similarly defined.

Using both left and right symmetric divergences, the distance between a and b is

$$d_1(a,b) = div_1(a^L, b^L) + div_1(a^R, b^R) \quad (6)$$

So the KL distance becomes:

$$\begin{aligned} KL(a,b) &= div(a^L, b^L) + div(a^R, b^R) \\ &= \sum_{w_{-1} \in W} p(w_{-1} | a) \log \frac{p(w_{-1} | a)}{p(w_{-1} | b)} \\ &+ \sum_{w_{-1} \in W} p(w_{-1} | b) \log \frac{p(w_{-1} | b)}{p(w_{-1} | a)} \quad (7) \\ &+ \sum_{w_1 \in W} p(w_1 | a) \log \frac{p(w_1 | a)}{p(w_1 | b)} \\ &+ \sum_{w_1 \in W} p(w_1 | b) \log \frac{p(w_1 | b)}{p(w_1 | a)} \end{aligned}$$

This is the widely used distance measure for lexical semantic similarity, e.g., Dagan et al. (1999); Meng and Siu (2002); Pargellis et al (2004). We can also see the IR distance and L1 distance below:

$$IR(a,b) = \sum_{w_{-1} \in W} p(w_{-1} \mid a) \log \frac{2p(w_{-1} \mid a)}{p(w_{-1} \mid a) + p(w_{-1} \mid b)} + \sum_{w_{-1} \in W} p(w_{-1} \mid b) \log \frac{2p(w_{-1} \mid b)}{p(w_{-1} \mid a) + p(w_{-1} \mid b)} + \sum_{w_1 \in W} p(w_1 \mid a) \log \frac{2p(w_1 \mid a)}{p(w_1 \mid a) + p(w_1 \mid b)} + \sum_{w_1 \in W} p(w_1 \mid b) \log \frac{2p(w_1 \mid b)}{p(w_1 \mid a) + p(w_1 \mid b)} \quad (8)$$

We can see from the IR metric that it is similar to the KL distance. Manhattan-norm (L1) distance :

$$L1(a,b) = \sum_{w_{-1} \in W} | p(w_{-1} \mid a) - p(w_{-1} \mid b) | + \sum_{w_1 \in W} | p(w_1 \mid a) - p(w_1 \mid b) | \quad (9)$$

In Pargellis et al. (2004), the lexical context is further extended from bigrams to trigrams as follows. For the sequence:

$..., w_{-2}, w_{-1}, w, w_1, w_2, ...$

where w can be word a or b, the trigram KL between a and b is:

$$KL_2(a,b) = \sum_{w_{-2}, w_{-2} \in W} p(w_{-2} w_{-1} \mid a) \log \frac{p(w_{-2} w_{-1} \mid a)}{p(w_{-2} w_{-1} \mid b)} + \sum_{w_{-2}, w_{-2} \in W} p(w_{-2} w_{-1} \mid b) \log \frac{p(w_{-2} w_{-1} \mid b)}{p w_{-2} w_{-1} \mid a)} + \sum_{w_1, w_2 \in W} p(w_1 w_2 \mid a) \log \frac{p(w_1 w_2 \mid a)}{p(w_1 w_2 \mid b)} + \sum_{w_1, w_2 \in W} p(w_1 w_2 \mid b) \log \frac{p(w_1 w_2 \mid b)}{p(w_1 w_2 \mid a)} \quad (10)$$

Since more information is taken into account in $KL_2(a,b)$, more constraints are imposed on the similarity measure. This is expected to improve the precision of clustering but may lead to a lower recall.

2.4 Co-occurrence Probability based similarity measures

After a close investigation of the corpus, we came up with an intuitive similarity measure directly based on the co-occurrence probability. The key idea is that the more common neighbouring words or phrases any two words or phrases in question share, the more similar they are to each other. Therefore, for each left or right neighboring word or phrase, we take the lower conditional probability into account.

Thus we have the following similarity measures:

Similarity using the bigram context

$$S_1(a,b) = \sum_{w_{-1} \in W} \min(p(w_{-1} \mid a), p(w_{-1} \mid b)) + \sum_{w_1 \in W} \min(p(w_1 \mid a), p(w_1 \mid b)) \quad (11)$$

Similarity using the trigram context

$$S_2(a,b) = \sum_{w_{-2}, w_{-1} \in W} \min(p(w_{-2} w_{-1} \mid a), p(w_{-2} w_{-1} \mid b)) + \sum_{w_1, w_2 \in W} \min(p(w_1 w_2 \mid a), p(w_1 w_2 \mid b)) \quad (12)$$

Similarity extending $S_1(a,b)$, taking both left and right contexts into account simultaneously

$$S_3(a,b) = S_1 + \sum_{w_{-1}, w_1 \in W} \min(p(w_{-1} w_1 \mid a), p(w_{-1} w_1 \mid b)) \quad (13)$$

After pairs of words or phrases are clustered above, those pairs with common members are further merged.

2.5 Comparison of measures

The KL distances emphasize on the difference of two probability but the new measure take the probability itself into account. Take the right bigram context the similarity measure for example:

$$KL_R(a,b) = \sum_{w_1 \in W} (p(w_1 \mid a) \log \frac{p(w_1 \mid a)}{p(w_1 \mid b)} + p(w_1 \mid b) \log \frac{p(w_1 \mid b)}{p(w_1 \mid a)}) \quad (14)$$

seeing $P(w_1 \mid a)$ as x and seeing $P(w_1 \mid b)$ as y, the equation changed to:

$$KL_R(x, y) = \sum (x \log \frac{x}{y} + y \log \frac{y}{x}) \quad (15)$$

and $S_R(x,y)$ becomes to:

$$S_R(x, y) = \sum \min(x, y) \quad (16)$$

We can also get the $IR_R(x,y)$ and $L1_R(x,y)$|

$$IR_R(x, y) = \sum (x \log \frac{2x}{x+y} + y \log \frac{2y}{x+y}) \quad (17)$$

and $L1_R(x,y) = |x - y|$ (18)

We can see the space distribution in Figure.1.

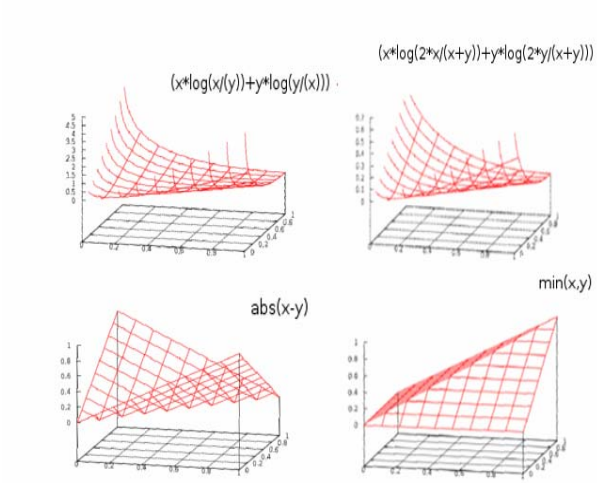

Figure 1. Space distribution of different metrics

$$x = y \quad (19)$$

$$z = 0$$

$$x = y = z \quad (20)$$

We can see from the four figures (the space distribution of four bigram metrics) that four curve surface are all symmetric. The curve surface of the three distance (KL,IR, L1) all contain the curve of (19), and curve surface of the minimum similarity contains the curve of (20). We say that the KL distances, IR distances and L1 distances all emphasize only on the distances and don't take the probability itself into account.

We take the right context of two pairs (a_1, b_1) and (a_2, b_2) for example. If

$$p(w|a_1) = 0.1, \quad p(w_1|a_1) = 0.9$$

$$p(w|b_1) = 0.1, \quad p(w_2|b_1) = 0.9$$

$$p(w'|a_2) = 0.9, \quad p(w_3|a_2) = 0.1$$

$$p(w'|b_2) = 0.9, \quad p(w_4|b_2) = 0.1$$

The calculation is shown as follows:

$$KL_R(a_1,b_1) = p(w|a_1)\log\frac{p(w|a_1)}{p(w|b_1)} + p(w|b_1)\log\frac{p(w|b_1)}{p(w|a_1)} = 0.1 * \log\frac{0.1}{0.1} + 0.1 * \log\frac{0.1}{0.1} = 0$$

$$KL_R(a_2,b_2) = p(w'|a_2)\log\frac{p(w'|a_2)}{p(w'|b_2)} + p(w'|b_2)\log\frac{p(w'|b_2)}{p(w'|a_2)} = 0.9 * \log\frac{0.9}{0.9} + 0.9 * \log\frac{0.9}{0.9} = 0$$

$$S_R(a_1,b_1) = \min(p(w|a_1), p(w|b_1)) = \min(0.1,0.1) = 0.1$$

$$S_R(a_2,b_2) = \min(p(w'|a_2), p(w'|b_2)) = \min(0.9,0.9) = 0.9$$

The KL calculation result of two pairs is the same but the new similarity calculated that (a_2, b_2) is more similar than (a_1, b_1) because they have more similar context probability 0.9.

3 Experiments and Results

3.1 Data

In our experiments, four types of corpora are exploited in different stages and different ways.

- T: A large collection of text corpus is used to train a general n-gram language model.
- H: Some WOZ dialogues were collected before the system is built, using a similar scenario where users talked in Chinese to a service provider (human) via telephone to search for local information, or information about some local points of interest (POI). These dialogues were manually transcribed and used for language model training. This is the best data we could get before the

system is built though it is not the real but near in-domain data.

- C: After the initial system was up running, some real human-computer dialogues were collected and transcribed. These dialogues were split into three sets. One (C1) is used for semantic class and structure induction. One (C2) is used as test data. The other (C3) is reserved.
- A: Domain information (domain entities) is used in conjunction with the induced semantic classes and structures from C1 to generate a large amount of in-domain corpus for language model adaptation. In Table 1, we give some statistics in terms of the number of utterances(no. u) and Chinese characters(no. c) for the above corpora.

corpus	no. u	no. c
T	38, 636	8, 706, 340
H	6, 652	151, 460
C1	658	15, 434
C2	1, 000	19, 284
C3	411	8, 014
A	14, 205	365, 576

Table 1. statistics of different corpus

3.2 Semantic Clustering

We conducted clustering with the above similarity measures on the data set C1.

During the clustering, it is required that all the probabilities involved in calculating similarity be larger than 0. We have no threshold except this constraint.

The outcomes are pairs of phrases.

It is noticed that most of the clustered words and phrases are domain entities.

In our experiments, we merged the induced similar pairs into large clusters. For example, if a is similar to b and b is similar to c, then (a, b, c) are merged into one category. In the end we use the categories to replace those words and phrases in corpus C1 and obtained templates.

Examples of the results are given below.

$ask $toponym $near $wh-word $sevice
[麻烦] $ask $toponym $near 有 $sevice 吗
我 在 $toponym $ask 怎么去 $poi
where:
$ask = 请问 | 问一下| 查询一下 | ...
$toponym = 清华大学 | 知春路 | ...
$sevice = 银行 | 加油站 | 体育馆 |...
$near = 附近 | 周围 | ...
$wh-word = 有没有 | 有什么 | 有哪些 | ...
$poi = 北京饭店 | 国家体育馆 | ...

To evaluate the induction performance, we compare the induced word pairs against manual annotation. We manually annotated each phrase with a tag like $toponym, $poi and so on. If a and b are calculated as a pairs and the annotation is the same, we see that they are correctly induced which is referred to Pangos (2006).

We compute the metrics of precision P, recall R and f-score F_1 as follows:

$$P = \frac{m}{M} \times 100\% \quad (21)$$

where m is the number of correctly induced pairs, and M is the number of induced pairs.

$$R = \frac{n}{N} \times 100\% \quad (22)$$

where n is the number of correctly induced words and phrases, and N is the number of words and phrases in the annotation.

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (23)$$

which is a harmonic mean of P and R.

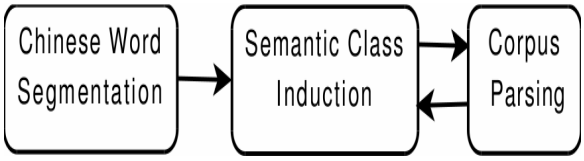

Figure 2. Induction process

The iterate process we adopted is as in Pargellis et al. (2004). In the first iteration, we calculated the similarity and use the largest similarity pairs to generate large classes which can be called semantic generalizer. Then we use these semantic classes to replace the corpus, and obtained new corpus just as the example presented above. Then we duplicate this process for the second iteration and so on.

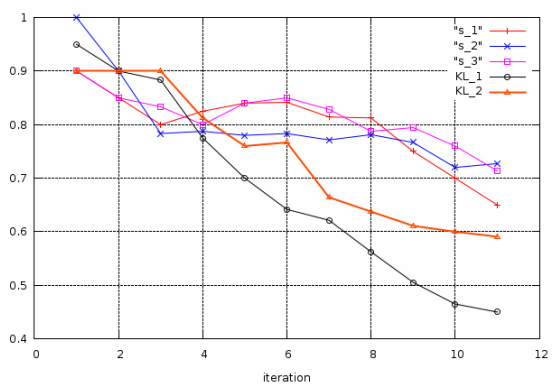

Figure 3. Precision according to iterations induced by KL and S1 similarity measure

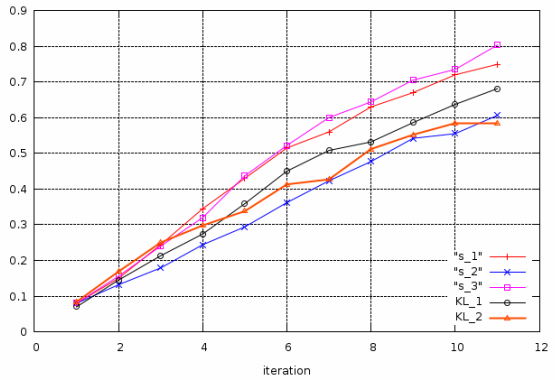

Figure 4. Recall according to iterations induced by KL and S1 similarity measure

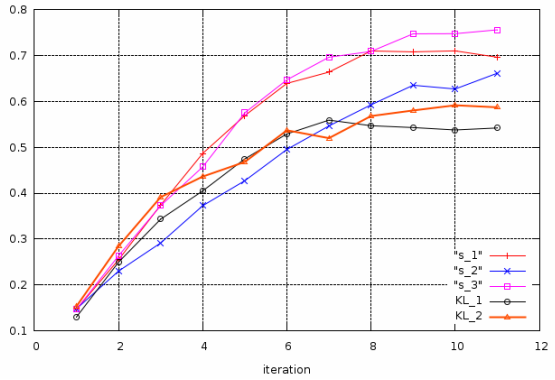

Figure 5. F1 according to iterations induced by KL and S1 similarity measure

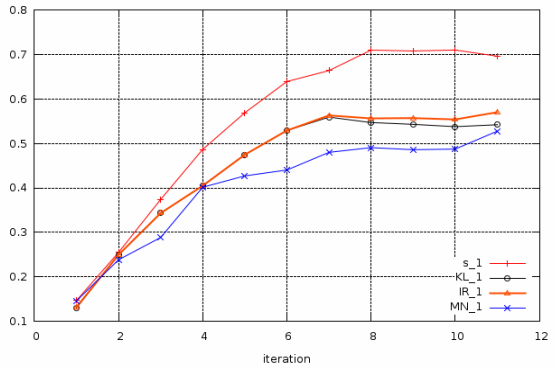

Figure 6. F1 according to iterations induced by all bigram similarity measure

From figures (Figure 3-6), we can see that clustering with our new co-occurrence probability based similarity measures outperforms that with the widely used relative entropy based distance measure consistently for both bigram and trigram contexts. This confirms the effectiveness of our new and simple measure. Regarding the context size, the results from using the bigram context outperforms that from using the trigram context in precision. But recall and F_1 drops a lot. This is due to that larger contexts bring more constraints. The context size effect holds for both types of similarity measures. And the best performance is achieved with the similarity measure S_3. It is based on S_1 and takes both left and right contexts into account at the same time.

3.3 Corpus Generation

Since the number of the domain entities (terminals) we can collect from the dialogues is very limited, we have to expand those variables (non-terminals) in the induced templates with domain information from the application database and relevant web sites. For example, we used all the words and phrases in the toponym cluster, e.g., ``清华大学 | 知春路 | ...", to replace $toponym in the templates above. Then we generated a large collection of artificial data which has a good coverage in both the utterance structures (the way people speak) and the domain entities. This resulted in the generated corpus A in Table 1. In generation we used the semantic classes and structures induced with S_3 and manually corrected some obvious errors. In the generated data, there are 14,205 utterances and 365,576 Chinese characters.:

3.4 Language Model Adaptation

There are some language model adaptation (LMA) work oriented to the dialogue systems e.g. Wang et al(2006), Hakkani-Tür et al.(2006), Bellegarda(2004). So far major effort has been spent on adaptation for large vocabulary speech recognition or transcription tasks. But recently there have been a few studies that are oriented toward dialogue systems, e.g. Wang et al(2006), Hakkani-Tür et al.(2006). In our experiments,

three trigram language models were built, each trained separately on the large text collection (T), on the WOZ data (H) and on the artificially generated data (A). These trigram models were then combined through model interpolation as follows: We used the linear interpolation to adapt language model. The formula is shown as follows. T is the out-of-domain data, H is the humane-to-humane dialogues, and A is the corpus generated by grammars

$$\begin{aligned} P(w_i \mid w_{i-1}w_{i-2}) &= \lambda_T P_T(w_i \mid w_{i-1}w_{i-2}) \\ &+ \lambda_H P_H(w_i \mid w_{i-1}w_{i-2}) \\ &+ \lambda_A P_A(w_i \mid w_{i-1}w_{i-2}) \end{aligned} \quad (24)$$

where $0 < \lambda_T, \lambda_H, \lambda_A < 1$ and $\lambda_T + \lambda_H + \lambda_A = 1$.

The weights were determined empirically on the held-out data (C3 in Table 1}).

All the language models were built with the Stolcke(2002)'s {SRILM} toolkit.

Why we did not use the C corpus directly is that it does't have a good covering on the domain-entities and other users usually say utterances similar to C in structures but different domain entities. So we use the good covering generated data to make LMA.

We evaluated the different language models with both intrinsic and extrinsic metrics. For intrinsic evaluation, we computed the perplexity. For extrinsic evaluation, we ran speech recognition experiments on the test data C2 and calculated the character error rate (CER).

We can see that corpus A is useful to make model adaptation and it is closer to the in-domain data than the human-human data for human-computer dialogues. By using these generated sentences, our domain-specific Chinese speech recognition have a growth from 85.2% to 91.4%.

λ_T, λ_H, λ_A	1, 0, 0	0.2, 0.8, 0	0.2, 0, 0.8	0.2, 0.4, 0.4
PP	984	95.4	33.6	23.3
CER(%)	32.3	14.8	10.7	9.0

Table 2. perplexity and character error rate according to model interpolation

The optimized weights (0.2,0.4,0.4) is obtained from the develop sets C3. From Table 2, we can see that language models built using additional dialogue related data, either human-human/WOZ dialogues or data generated from human-computer dialogues, shows significant improvement in both perplexity and speech recognition performance over the one built with the general text data only. For the two dialogue related data, the generated data is better than the WOZ data or closer to the test data, since perplexity further drops from 103.5 to 38.1 and CER drops from 14.8 to 10.7. This confirms our conjecture that human-human WOZ dialogue data is near in-domain and not very proper for human-computer dialogues. Therefore, to effectively improve language modeling for human-computer dialogues, we need more in-domain data, even if it is generated or artificial. The best language model is obtained through interpolation of both language models from dialogue related data with the one from general text data. This may be because there is still some mismatch between data sets C1 (for induction and generation) and C2 (for test).

And some of the missing bits in C1 appeared in the WOZ data (corpus A).

4 Related Works

The most relevant work to ours is done by Wang et al. (2006), who generated in-domain data through out-of-domain data transformation. First some artificial sentences are generated through parsing and reconstructing out-of-domain data and the illegal ones are filtered out. Then the synthetic corpus is sampled to achieve a desired probability distribution, based on either simulated dialogues or semantic information extracted from development data. But we used a different approach in producing more in-domain data. First semantic classes and structures are induced from limited human-computer dialogues. Then large amount of artificial in-domain corpus is generated with the induced semantic classes and patterns augmented with domain entities.

The main difference between the two works lies in how the data is generated and how the generated data helped.

5 Conclusions and Future Work

In this paper, we described our work on generating in-domain corpus using the auto-induced semantic classes and structures for language model adaptation in a Chinese voice search dialogue system. In inducing semantic classes we proposed a novel co-occurrence probability based similarity measure. Our experiments show that the simple co-occurrence probability based similarity measure is effective for semantic clustering which is used in our experiment. For interpolation based language model adaptation, the data generated using the induced semantic classes and structures enhanced with domain entities helped a lot for human-computer dialogues. Despite that we dealt with the language of Chinese, we believe that that approaches we employed are language independent and can be applied to other languages as well.

In our experiment we noticed that the performance of semantic clustering was affected quite a lot by the noises in the data. For future work, we would like to investigate how to further improve the robustness of semantic clustering in noisy spoken language. The semantic structures induced above are very shallow. We would like to investigate how to find deep semantics and relations in the data.

Acknowledgement

This work is partially supported by The National Science & Technology Pillar Program (2008BAI50B03), National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014).

References

Arai, K. J. H., Wright, G. Riccardi, and Gorin, A. L. "Grammar fragment acquisition using syntactic and semantic clustering," *Speech Communication*, vol. 27, iss. 1, pp. 43–62, 1999

Bellegarda, J. R. Statistical language model adaptation: review and perspectives, *Speech Communication*, vol. 42, iss. 1, pp. 93–108, 2004

Cover, T. M. and Thomas, J. A., *Elements of Information Theory*. Wiley-Interscience, 2006

Dagan, I., Lee, L. and Pereira, F. C. N. "Similarity-Based Models of Word Cooccurrence Probabilities," *Machine Learning*, 1999

Gorin, A. L. "On automated language acquisition," *Acoustical Society of America Journal*, vol. 97, pp. 3441–3461, 1995

Hakkani-Tür, D. Z., Riccardi, G. and Tur, G. An active approach to spoken language processing, *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 3, iss. 3, pp. 1–31, 2006

Lin, D. "An information-theoretic definition of similarity," in *Proc. ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, 1998

Meng, H. M. and Siu, K.-C. "Semiautomatic Acquisition of Semantic Structures for Understanding Domain-Specific Natural Language Queries," *IEEE Trans. Knowl. Data Eng.* 2002

Pargellis, A. N., Fosler-Lussier, E., Fosler-Lussier, Lee, C.-H., Potamianos, A. and Tsai, A. "Auto-induced semantic classes," *Speech Communication*, vol. 43, iss. 3, pp. 183–203, 2004

Pangos, A Combining statistical similarity measures for automatic induction of semantic classes, 2005

Pangos, A., Iosif, E. and Tegos, A. Unsupervised combination of metrics for semantic class induction, *SLT 2006*, 2006

Pangos, A. and Iosif, E., A Soft-Clustering Algorithm for Automatic Induction of Semantic Classes, *interspeech07*, 2007

Stolcke, A. SRILM – an extensible language modeling toolkit, in *Proc. ICSLP*, 2002

Wang, C. Chung, G. and Seneff, S. Automatic induction of language model data for a spoken dialogue system, *Language Resources and Evaluation*, vol. 40, iss. 1, pp. 25–46, 2006

Wang, Y.-Y. and Dong Yu, E. A., An introduction to voice search, *Signal Processing Magazine, IEEE*, vol. 25, iss. 3, pp. 28–38, 2008

Weeds, J., Weir, D. and McCarthy, D. "Characterising measures of lexical distributional similarity," in *Proc. in Proc. COLING '04*, 2004,