A Comparison of Rule–Based and Statistical Methods for Semantic Language Modeling and Confidence Measurement

Ruhi Sarikaya **Yuqing Gao** **Michael Picheny**

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
{sarikaya,yuqing,picheny}@us.ibm.com

Abstract

This paper presents a comparison of a rule-based and a statistical semantic information modeling technique. For the rule–based method we employ Embedded Grammar (EG) tagging and for the statistical method we use a previously proposed Semantic Structured Language Modeling (SSLM) technique. Both EG and SSLM achieve around **15%** relative improvement in speech recognition performance over the baseline dialog state–based trigram language model in a financial transaction domain. Combining EG and SSLM using linear interpolation results in further improvement. We also use the features obtained from EG and SSLM for confidence measurement. Word level confidence measurement experiments using EG and SSLM–based semantic features combined with posterior probability show over **20%** relative improvement in correct acceptance rate (CA) at 5% false alarm (FA) rate over the posterior probability based feature. In both language model rescoring and confidence measurement experiments SSLM outperforms EG by a small margin.

1 Introduction

There are two main approaches for semantic information modeling: rule–based (or grammar–based) and statistical. For spoken dialog systems, grammar and statistical methods occupy the opposite sides of the spectrum in terms of the assumptions they make on users and the "completeness" of utterances. In general, grammar–based approaches expect sophisticated users, who can form detailed, grammatical and complete utterances. On the other side of the spectrum, statistical methods treat speech as an inherently incomplete process, since users in general do not know the system coverage and also they may not always form grammatical sentences (i.e., spontaneous speech). Both methods have advantages and disadvantages. Statistical methods require significant amount of annotated data for reliable information modeling. They usually do not need *a priori* information about the task, which makes them portable to other tasks as long as there is annotated data for those domains. However, statistical methods suffer from poor generalizations when data is insufficient. On the other hand, grammar–based methods do not need annotated training data, but require major effort by experts to hand–code the *a priori* information into the system. Grammar–based methods for language modeling are attractive alternatives to statistical models in domains that lack extensive speech corpora (Jurafsky, 1995).

We introduced a set of statistical language modeling techniques that use semantic analysis for spoken dialog systems (Erdogan, 2002). The motivation was to incorporate the semantic information from the semantic parse tree into language modeling. The SSLM uses varying levels of lexical and semantic information using maximum entropy (ME) modeling.

Semantic information can also be used for confidence measurement. Since the speech recognition output is always subject to some level of uncertainty, it is essential to employ a measure that indicates the reliability (of the correctness) of hypothesized words. There are a number of overlapping speech recognition based features that were exploited in many studies (San-Segundo, 2001; Zhang, 2001; Pao, 1999). For domain independent large vocabulary speech recognition systems, posterior probability based on a word graph was shown to be the single most useful confidence feature (Wessel, 2000). In many, if not all, of the previous studies the way semantic information was incorporated into decision process is rather *ad hoc*. For example in (Pao, 1999), semantic weights assigned to words are based on heuristics. Likewise, in (Carpenter, 2001) semantic features such as "uncovered word percentage", "gap number", "slot number", etc. were generated experimentally in an effort to incorporate semantic information into confidence metric. We proposed two methods to obtain semantic information from the parser output to incorporate into the posterior probability (Sarikaya, 2004; Sarikaya, 2003). In this study, we compare and combine the grammar and statistical methods for language modeling and the features obtained from them for confidence measurement.

The rest of the paper is organized as follows: in Sec-

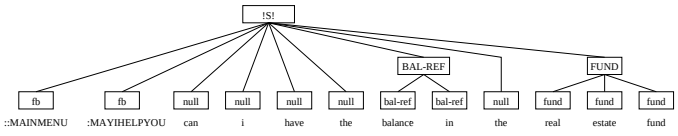

Figure 1: An example of a semantic parse tree.

tion 2, we describe EG tagging. The ME based SSLM is presented in Section 3 followed by experimental results in Section 4. Section 5 summarizes the findings.

2 Embedded Grammar Tagging

One of the main issues with spoken dialog systems is the lack of sufficient carefully transcribed domain specific training data. Since the collection of this kind of data is a time and financial bottleneck, some researchers choose to write hand–crafted EG in some subset of the context–free grammar (CFG) with the corresponding semantic labeling.

In this study, semantic concepts for a financial transaction domain are represented by EGs. Seventeen semantic concepts are determined and the corresponding EG rules are written. The rules are written in standard Backus–Naur Form (BNF) and compiled into a stochastic recursive transition network. In order to achieve appropriate associations and minimize the number of specific rules, concept spotting is performed. The EG searches for the phrase patterns corresponding to concepts in user utterances and may generate numerous slot–filling signals. The system decision is based on maximum word coverage. For example, the following word sequence is part of a request from the system to sell "thirty thousand dollars" of a fund:

```
fb_TARGETFUND sell <AMOUNT> thirty thousand
              dollars </AMOUNT> of ...
```

where "fb_TARGETFUND" is the dialog state feedback and "AMOUNT" is one of the concepts defined by an EG. The LM probability of starting grammar <AMOUNT> is $p(AMOUNT|sell, fb\_TARGETFUND)$. The LM probability of "thirty" is the probability of "thirty" given it is the first state of the <AMOUNT> grammar, $p(thirty|s_1\{AMOUNT\})$. The LM probability of < /AMOUNT> is the grammar completeness probability. The trigram language model treats the EG as a single token of context: $p(of|AMOUNT, sell)$. The seventeen grammars used include: {AGE, AMOUNT, DATE, DURATION, FUND, LOANTYPE, MARKET, MTYPE, NUMBER, ORD, PERCENT, PIN, PLAN, SHARES, SSN, SYSTEM, WITHDRAWAL}. Optimizing grammars is an iterative process. It took several months to complete the process since adding new data requires evaluating the grammars against the new data and redesigning them.

3 Semantic Structured Language Modeling Using Maximum Entropy Method

The purpose of semantic analysis is to model relationships between semantically associated words. A set of words can form a semantic unit, such as a concept. Depending upon the depth of analysis, relationships between semantic units can be modeled as well. The semantic analysis used here is based on statistical parsing. The decision tree based statistical parser uses training data to assign probabilities to each node and extension in a parse tree. A parse tree is represented as a connected, single–rooted graph with feature values at each node. An example of a parse tree in the financial transaction domain is shown in Fig. 1. As seen in the figure, each word is assigned a tag and certain tags are grouped under a label to form a concept.

The ME method is a flexible modeling framework that allows the combination of multiple overlapping information sources. In natural language processing, ME has been widely employed in statistical language modeling (Chen, 2000). The ME modeling matches the feature expectations exactly while making as few assumptions as possible in the model. The multiple information sources are combined as follows:

$$P(o|h) = \frac{e^{\sum_i \lambda_i f_i(o,h)}}{\sum_{o'} e^{\sum_i \lambda_i f_i(o',h)}}, \qquad (1)$$

where o is the current word, f_i are the feature indicators that are activated for a certain history, and h represents the history which may include previous words as well as tags and labels that can be used in predicting the current word.

We used the ME to model semantic and syntactic information in a sentence (Erdogan, 2002). We computed the joint probability of a word sequence and a parse tree: $P(W, C)$. Although this joint probability can be decomposed in two ways, as $P(W|C)P(C)$ and $P(C|W)P(W)$, we built a direct ME model (Erdogan, 2002). The first step in building the ME model is to represent a parse tree as a sequence of words, tags, and labels. Converting the parse tree into a text sequence that is composed of of labels, words and tags allows us to group the semantically related words and even semantically related concepts. For example, the parse tree given in Fig. 1 can be converted into the text format as follows:

```
{!S! :MAINMENU_fb :MAYIHELPYOU_fb can_null
i_null have_null the_null {BAL-REF
balance_bal-ref in_bal-ref BAL-REF} the_null
{FUND real_fund estate_fund fund_fund FUND} !S!}
```

The MELM2 is one of the SSLMs proposed in (Erdogan, 2002), which employed 7 types of questions about the current token in a sentence. Any word, tag or label in the text representation above is considered as a

token. In addition to regular n–gram questions, four more questions are used regarding the semantic structure of the sentence. These questions are (1) current active parent label (L_i), (2) L_i and number of words to the left since starting the current concept (N_i), (3) L_i, N_i and previous word token, (4) the previous completed constituent (O_i) and number of words to the left since completing O_i. The history given in Eq. 1 consists of answers to these questions.

The language model score for a given word in MELM2 model is conditioned not only on the previous words but also on the labels and the relative coverage of these labels over words. The SSLM presents an effective statistical method to combine word sequences with semantic parse tree. Therefore we can use the SSLM score as a feature for confidence measurement.

4 Experimental Results and Discussions

The experiments are conducted on a financial transaction task. The SSLM used 28.3K semantically annotated sentences (105K words) as training data. The ME–based SSLM is trained with the improved iterative scaling algorithm using fuzzy smoothing (Erdogan, 2002; Chen, 2000). The acoustic data of the SSLM training data is used as confidence measurement training data. The confidence measurement test data consists of 3152 sentences amounting to 11.4K words. The confidence training and test data have 27.9% and 28.1% word error rates (WER), respectively. The speech recognition acoustic models are trained using generic telephony data. A dialog state–based trigram language model (DS–3gr) with deleted interpolation is used for the speech recognition to obtain the baseline WER and generate an N–best list. The baseline DS–3gr used a separate 194K sentences as training and additional 10K sentences as held–out data from the financial domain.

The N–best list contains an average of 34 alternative hypotheses per sentence with an oracle WER of 16.2%. The SSLM and EG are used to rescore the N–best list hypothesis. Table 1 shows the baseline DS–3gr, EG and the ME–based SSLM results. The EG achieved a 15.3% relative improvement over the baseline language model. The SSLM resulted in 15.7% improvement. These improvements are due to the inclusion of new semantic information that was not part of the original speech recognition system. Even though individual improvements are similar, linearly interpolating EG with SSLM led to further improvement. The overall improvement compared to baseline is 18.9%. The interpolation weight used for EG and SSLM is 0.5. The results indicate that EG models local semantics in a sentence and SSLM models overall semantic structure of a sentence. Combining them can improve the semantic modeling of the sentence.

Language Model Rescoring	
LM	WER (%)
DS–3gr	28.1
EG	23.8
SSLM	23.7
EG + SSLM	22.8

Table 1: Word error rates for the baseline, EG and SSLM–based language models.

The posterior probabilities are based on the sausages which are obtained from the word graph (Mangu, 1999). A sausage is a simplified word graph with a specific topology. The goal in this conversion is to minimize the WER rather than the sentence error rate. The technique is named as "sausage" since the visual representation of this graph looks like a sausage in its literal sense. The word graph is converted into a sequence of confusion sets along time. Each confusion set consists of a group of words, which are competing hypotheses for a certain time interval. The posterior probability for each word is obtained by summing the probabilities of all the paths going over that word.

A sausage is generated for each sentence in the confidence training and test data. The best path from the sausage is hypothesized as the speech recognition output. Each word is labeled as correct ("1") or incorrect ("0") after aligning the hypothesis with the reference transcript. All recognition hypotheses are parsed using the statistical semantic parser. Each sentence is scored with both EG and SSLM to assign semantic probabilities to each word. The corresponding semantic features are extracted for all the words in the sentence. All of the positive (correct recognition) and negative (misrecognition) examples are pooled in two sets. A decision tree is built using the respective features. The decision tree used the raw scores of each feature. The tree is grown by partitioning the data recursively at each node until either the node becomes homogeneous or contains too few observations (≤ 500).

Receiver operating characteristic (ROC) curve is one of the commonly used tools for confidence measurement performance. The most interesting part of the ROC curve for dialog systems is where the false acceptance rate is low and correct acceptance rate is high, because one needs to accept as many correct words as possible at low False Acceptance (FA) rates. The FA and CA are calculated using the following formula:

$$\begin{aligned} \text{FA} &= \frac{\text{\# of falsely accepted words}}{\text{Total \# of negative examples}} \times 100 \\ \text{CA} &= \frac{\text{\# of correctly accepted words}}{\text{Total \# of positive examples}} \times 100 \end{aligned} \quad (2)$$

Here, *EG* refers to EG language model score, *SSLM* refers to the semantic language model score, and *post* refers to posterior probability for a given word. Based on the individual feature performances *post* outperformed both *EG* and *SSLM* for almost all of the FA

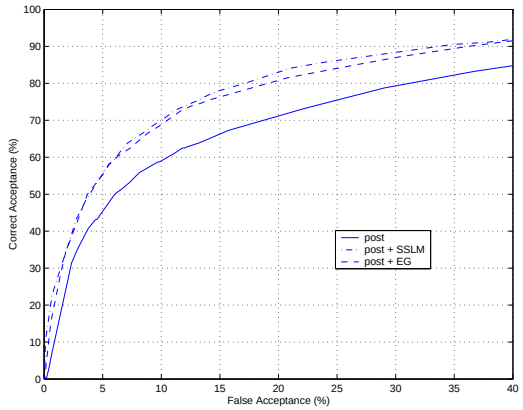

Figure 2: ROC for combination of posterior probability with the EG and SSLM–based features.

rates. For example at 5% FA rate the CA rates are 45.4%, 42.5% and 32.3% for *post*, *SSLM* and *EG*, respectively. In Fig. 2, we present the ROC curve for feature performances of the *post*, combined with *EG*, and *SSLM*. Combining *SSLM* with the *post* improved the CA rate significantly. This is because of the fact that *SSLM* brings complementary information to speech recognition based *post*. Similar observations hold for combining *EG* with *post*. When *EG* and *SSLM* are combined individually with *post* the difference in performance decreases significantly. Even though *SSLM+post* is slightly better than *EG+post*, the difference is not significant. We extracted the CA rates at 5% and 10% FA rates from the ROC curve and presented them in Table 2. At 5% FA rate, the improvements in CA rate over posterior probability are 21.8% and 22.2% for *EG+post* and *SSLM+post*, respectively. At 10% FA rate the respective improvements are 17.0% and 18.8%. The confidence measurement results indicate that *SSLM* outperforms *EG* by a small margin. Combination of all three features (*post*, *EG* and *SSLM*) did not provide further improvement.

5 Conclusions

We compared a rule–based embedded grammar tagging and the statistical semantic structured language modeling for modeling semantic information in a sentence. These two techniques were compared in language model rescoring and confidence measurement experiments. The N–best list rescoring with the EG and SSLM showed over 15% relative improvement over dialog state–based trigram language model. Linear interpolation of EG and SSLM resulted in 18.9% improvement in WER. For confidence measurement, combining the EG and SSLM–based features with speech recognition based posterior probability features provided an improvement of over 20% in correct acceptance at 5% false acceptance rate over posterior probability. In both language model rescoring and confidence measurement experiments, SSLM slightly outperformed EG.

Performance of EG and SSLM–based Features.(%)		
	5% FA	10% FA
Posterior (Post)	45.4	58.9
EG	32.3	45.7
SSLM	42.5	55.2
Post + EG	55.3	68.9
Post + SSLM	55.5	70.0
EG + SSLM	42.5	54.7
Post + EG + SSLM	55.6	70.2

Table 2: Correct Acceptance (CA) rates at 5% and 10% False Acceptance (FA) rates.

Acknowledgments

The authors would like to thank Mike Monkowski for designing the embedded grammars, and Hakan Erdogan and Mark Epstein for fruitful discussions.

References

S.F. Chen and R. Rosenfeld. 2000. *A Survey of Smoothing Techniques for Maximum Entropy Models.* IEEE Trans. Speech Audio Proc. (SAP), 19(3):37–50.

R. Sarikaya, Y. Gao, M. Picheny and H. Erdogan. *Semantic Confidence Measurement for Spoken Dialog Systems.* IEEE Trans. SAP, to appear.

D. Jurafsky *et al.* 1995. *Using a Stochastic Context-Free Grammar As a Language Model for Speech Recognition.* Inter. Conf. on Acous. Speech Signal Process (ICASSP).

R. Sarikaya, Y. Gao, and M. Picheny. 2003. *Word Level Confidence Measurement Using Semantic Features.* ICASSP.

R. San-Segundo, B. Pellom, K. Hacioglu, and W. Ward. 2001. *Confidence Measures for Spoken Dialog Systems.* ICASSP.

P. Carpenter *et al.* 2001. *Is This Conversation on Track.* European Speech Technology Conference (Eurospeech).

D. M. Magerman. 1994. *Natural Language Parsing As Statistical Pattern Recognition.* Ph.D. Thesis, Stanford University.

C. Chelba and F. Jelinek. 1999. *Recognition Performance of a Structured Language Modeling.* Eurospeech.

F. Wessel, K. Macherey and H. Ney. 2000. *A Comparison of Word Graph and N-best list based Confidence Measures.* ICASSP.

R. Zhang and A. Rudnicky. 2001. *Word Level Confidence Annotation Using Combination of Features.* Eurospeech.

L. Mangu, E. Brill and A. Stolcke. 1999. *Finding Consensus Among Words: Lattice–based Word Error Minimization.* Eurospeech.

C. Pao, P. Schmid and J. Glass. 1999. *Confidence Scoring for Speech Understanding Systems.* Inter. Conf. on Spoken Language Process (ICSLP).

H. Erdogan, R. Sarikaya, Y. Gao and M. Picheny. 2002. *Semantic Structured Language Models.* ICSLP.