SUB-SENTENCE DISCOURSE MODELS FOR CONVERSATIONAL SPEECH RECOGNITION

Kristine W. Ma, George Zavaliagkos, Marie Meteer

GTE/BBN Technologies
70 Fawcett Street
Cambridge, MA 02138
kma@bbn.com

ABSTRACT

According to discourse theories in linguistics, conversational utterances possess an informational structure that partitions each sentence into two portions: a "given" and "new". In this work, we explore this idea by building sub-sentence discourse language models for conversational speech recognition. The internal sentence structure is captured in statistical language modeling by training multiple n-gram models using the Expectation-Maximization algorithm on the Switchboard corpus. The resulting model contributes to a 30% reduction in language model perplexity and a small gain in word error rate.

1. INTRODUCTION

Numerous researchers have worked on building statistical language models using topic, discourse or dialogue structure for speech recognition and spoken language understanding [4, 5, 8], with a focus on modeling structures that occur at the sentence level or higher. In this work, we aim at building language models that capture dialog structure at the sub-sentence level, to reduce language model perplexity and word error rate for conversational speech recognition on the Switchboard corpus.

This work is motivated by suggestions from discourse theories in linguistics that conversational sentences not only have a syntactic structure, but also an information structure, consisting of two parts: a "given" and "new" [1, 3]. The "given" part typically occurs at the beginning of a sentence where there is less informational content, whereas the "new" tends to occur towards the end where most of the new information is being conveyed.

In Section 2, we present some preliminary data analysis based on the given-new idea on the Switchboard corpus. Section 3 describes our proposed statistical sentence model, as well as the training procedure. Recognition and perplexity results and conclusions are in Section 4 and 5, respectively.

Table 1: Use first strong verb of sentence as pivot in dividing a sentence into a "before" and an "after" part.

BEFORE PIVOT	AFTER PIVOT
That's	a good point.
But, uh, other than that I think maybe it just	depends on how you define honesty.
That's a int-, you know, that's	interesting.

2. PRELIMINARY ANALYSIS

2.1. Background

The concept that a conversational utterance is comprised of two distinct parts was first explored by Meteer in [6] for the purpose of speech recognition. There, the authors worked with linguistic clauses, and divided each utterance into two portions, a "before" and an "after", by pivoting on the first strong verb of each sentence as shown in Table 1. They observed differences in vocabulary frequency and dysfluency distributions, and conducted a set of bi-gram perplexity experiments that suggested the structural differences between the two portions of a sentence.

2.2. Perplexity

In this work, we begin with a similar set of perplexity experiments, except we train and test only on sentences that have a pivot point, use tri-gram models and a slightly larger training set. We also eliminated artifacts due to sentence boundary markers that would bias perplexity results. For example, a 5-word training sentence can be represented as

```
Full:   <b> w1 w2 {w3} w4 w5 <e>
Before: <b> w1 w2 {w3} <e>
After:  <b> {w3} w4 w5 <e>
```

where {w3} is the pivot verb. In testing, we discount the contribution made by the end of sentence boundary marker <e>, and the

begin of sentence boundary marker <b> for the "Before" and the "After" segments, respectively:

```
Full:   <b> wa wb (wc) wd we <e>
Before: <b> wa wb (wc)
After:  (wc) wd we <e>
```

The perplexity results are shown in Table 2. Three trigram language models are trained; one on full sentences, one on the "before" portion only, and one on the "after" portion only. Each of the trigram language models is then tested on a full sentence, a "before", and an "after" test set. The experiment was carried out on the manually annotated and linguistically segmented Switchboard data obtained in the 1995 Language Modeling workshop at Johns Hopkins.

From the table of perplexity numbers, we see that even though having a smaller training set hurts (81 → 86) the "Before Pivot" performance, it is out weighed by the gain (145 → 114) observed on the "After Pivot" performance. This indicates that modeling the "before" and "after" parts of a sentence separately reduces perplexity despite partitioning of the training data, which suggests potential recognition gains.

Table 2: Trigram perplexity on linguistic data.

		Testing Condition		
Training Condition	Train Set Size	Full Sentence	Before Pivot	After Pivot
Full-LM	1345k words	99	81	145
Full-LM	673k	106	86	135
Before-LM	695k	-	86	-
After-LM	770k	-	-	114

2.3. Word Recognition

We further tested this idea by building a two-state sentence model (Figure 1), where the "before" state represents the "given," or earlier part of a linguistic sentence, while the "after" state models the "new," or latter part of a linguistic sentence. We trained and tested this model via n-best re-scoring on linguistic data and obtained a 0.4% reduction in word error rate. This trial confirms that one could indeed take advantage of the internal given-new sentence structure for the purpose of speech recognition.

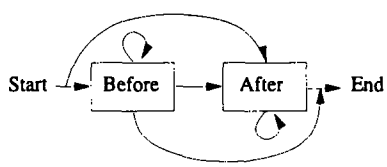

Figure 1: Model for linguistic sentence.

This proposed model is ideal for representing linguistically segmented data with complete clauses. It fails, however, to model acoustically segmented speech data, which we need to use for training and testing our speech recognizer. In acoustic segmentation, speech waveform is segmented according to simple acoustic events such as filled pauses, non-speech elements and turn boundaries.

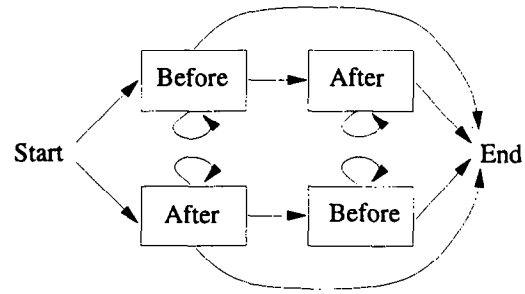

Figure 2: A Four State Sentence Model.

Generating acoustic segmentation is inexpensive and fully automatic, however, obtaining good linguistic segmentation boundaries is still an ongoing research topic [2, 7]. Rather than relying on errorful linguistic segmentations, we chose to extend our 2-state model to handle arbitrary utterances not necessarily occurring in the rigid "given" followed by "new" structure. We developed a training and testing procedure that automatically extracts the "given" and "new" part of the sentence structure from acoustically segmented data. In the following sections, we experiment with three sentence models, each allowing a higher degree of flexibility in informational structure.

3. FLEXIBLE SENTENCE MODELS

3.1. A Four State Sentence Model

In order to work with acoustically segmented data, we adapted a model that allows the following utterance structure:

```
B-A: (it's a it's) (a fairly large community)
A-B: (raise those children) (and and now ...)
A only: (supposed to be a great baby-sitter)
B only: (you know so that's that's that's)
```

Such a flexible multi-state model is shown in Figure 2, and is trained as shown in Figure 3. To begin, we use some linguistically segmented sentences to bootstrap our automatic training procedure. As in Meteer [6], we partition each linguistic sentence into a "before" and an "after" portion by pivoting on the first strong verb, then use each subset to train a "before" and an "after" n-gram language model, respectively. We use each n-gram to compute the emission probabilities of its corresponding states in Figure 2. The arcs in the diagram simply depict allowable state transitions; they have no probability mass. Once the initial models have been trained, they are re-estimated via the Expectation-Maximization (EM) algorithm using the acoustic segments. Let $w_1^T = \{w_1, w_2, ..., w_T\}$ be the utterance string. Following standard forward-backward computation for HMM, the language model weight for each n-tuple, $w_{t-n+1}^t = \{w_{t-n+1}, ..., w_t\}$, and for each state, $k \in \{Before, After\}$, is updated as,

$$c_k^i(w_{t-n+1}^t) = P(s_t = k \mid w_1^T, \theta^{i-1}) = \frac{\alpha_t(k)\beta_t(k)}{\sum_j \alpha_T(j)}$$

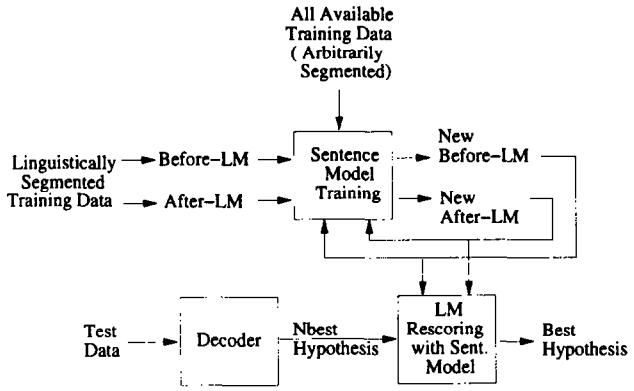

Figure 3: System Overview.

where,

$$\alpha_t(k) = \sum_j \alpha_{t-1}(j) P(w_t, s_t = k \mid w_{t-n}^{t-1}, s_{t-1} = j, \theta^{i-1})$$
$$\approx \sum_j \alpha_{t-1}(j) P_k^{i-1}(w_t \mid w_{t-n}^{t-1})$$

$$\beta_t(k) = \sum_j \beta_{t+1}(j) P(w_{t+1}, s_{t+1} = j \mid w_{t-n+1}^{t}, s_t = k, \theta^{i-1})$$
$$\approx \sum_j \beta_{t+1}(j) P_j^{i-1}(w_{t+1} \mid w_{t-n+1}^{t})$$

where $s, j \in \{Before, After\}$ are the state variables, θ^{i-1} is the sentence model and $P_M^{i-1}(.)$ is the n-gram probability evaluated with n-gram $M \in \{Before, After\}$ obtained from EM iteration $i-1$.

Using EM allows us to automatically extract the "given" and "new" part of the sentence structure from arbitrarily segmented data and re-estimate a new "before" and "after" n-gram model in a maximum likelihood manner. Furthermore, the two resulting n-gram models are smoothed in the sense that each n-gram model was re-estimated using all the training data rather than from a subset of the data. Hence, the "before" model is updated with a small, but non-zero, weight for n-tuple that are unlikely to occur in the "before" part of a sentence. In addition, using EM training, we smooth over the assumption of pivoting at the first strong verb. This training step can be repeated as many times as needed. We found that the system usually converges within 3 iterations.

In testing, we use the resulting sentence model to re-score the n-best hypothesis of the decoder output by computing the likelihood score of each sentence hypothesis given the model, which is simply

$$P(w_1^T | \theta^i) = \sum_j \alpha_T(j).$$

3.2. A Six State Sentence Model

The four-state sentence model attempts to represent speech segments of the form `before`, `before-after`, `after`, and `after-before`. It is, therefore, inadequate for modeling compound sentences such as:

```
We found this one area that doesn't have
mosquitos they just don't have them and
it's just wonderful ...
```

In the six-state sentence model, we simply extended our existing topology for one extra level of flexibility by adding two more states so as to also model segments of the form `before-after-before` and `after-before-after`.

3.3. Unconstrained Sentence Model

Finally, we developed an unconstrained sentence model that allows an unlimited number of state transitions, as shown in Figure 4, with parameter p and q as the state transition penalty.

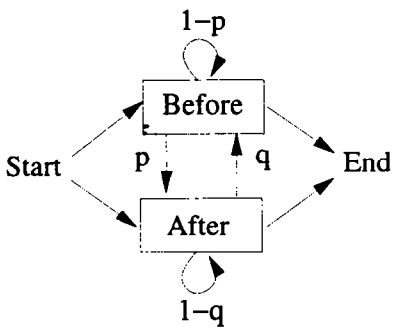

Figure 4: Unconstrained Sentence Model.

4. EXPERIMENTAL RESULTS

To build the initial "before" and "after" n-gram models, we use 1.3 million words of linguistically segmented Switchboard data obtained from the 1995 Language Modeling workshop at Johns Hopkins. For subsequent EM iterations, we use all the available Switchboard and Callhome English training data. This training set is automatically segmented by NIST, and comprises more than 3 million words. We use trigram as our building blocks for all the sentence models. Our test set is the development set defined by NIST for the Spring 1997 Large Vocabulary Speech Recognition Evaluation. It consists of 7 Switchboard and 7 Callhome English conversations for a total of 13k words.

The results are shown in Table 3. The baseline system yields a 35.94% word error rate. This is simply the error rate of the 1-best hypothesis from the decoder. The baseline perplexity is from the decoding language model, a monolithic trigram trained on the same 3 million words of acoustically segmented data. The 4-state sentence model gives a significant drop in trigram perplexity, and a slight improvement on word recognition error rate via n-best rescoring. Note that in Table 3, "iter0" is the initial sentence model with no EM training, "iter1" is after 1 iteration of EM training, and so forth. All the perplexity numbers reported in this paper are from the third EM iteration and are based on the likelihood score of the best word path (as opposed to the sum of all paths) through

the sentence model. This is done so that the perplexity numbers across different sentence models are compatible with each other and to that of the baseline trigram model.

The 6-state sentence model provided an additional drop in grammar perplexity but not a significant change in word recognition error rate. The additional gain of allowing a more flexible sentence topology suggests that we should reformulate our sentence model to allow an unlimited number of state transitions, maybe at the expense of over generalization. This leads to our final unconstrained sentence model. Thus far, the resulting performance did not make a significant difference whether the state transition probabilities, p and q, are updated via EM or are fixed at 0.5. This new model gives us an additional gain in perplexity, but not so much in word recognition error rate.

Overall, we obtained a slight improvement of 0.26% in word recognition error rate in using our proposed sentence models for n-best re-scoring. From Table 3, we see that half of the improvement can be obtained without running EM training, but in order to compute the likelihood score of the initial sentence model given the test utterances, an underlining system still needs to be implemented.

The three sentence models give similar performance in terms of word recognition error rate. In examining the Viterbi state alignment of the sentence string with the models, however, we found that the unconstrained model was able to represent compound sentences with fairly accurate state transitions, and thus it is our preferred system. Since our sentence model operates on the decoder output, its current performance may be hampered by the high baseline word recognition error rate. Therefore we believe that it will become more effective as the underlying performance of the decoder improves. Another source of sub-optimality is that our models were trained using all the available data, which includes a substantial amount of back-channel utterances that dilute the "before" and "after" n-grams. To separate back-channel phrases, we could simply add an extra back-channel single-state path to our current model, then train via EM, or we could rely on using discourse prediction algorithms that in addition use acoustic cues.

Table 3: Word error rate from N-best re-scoring using sentence model. Iter refers to EM training iterations.

Model	Iter0	Iter1	Iter2	Iter3	PP
Baseline	35.94				156
4-state Sent. Model	35.81	35.82	35.78	35.77	121
6-state Sent. Model	35.81	35.81	35.77	35.71	115
Unconstrained Model	35.85	35.81	35.74	35.68	108

5. CONCLUSION

We have presented the motivation, the system implementation, and experimental results for building language models based on modeling internal sentence structure. According to linguistic discourse theories, conversational utterances contain a given-new informational structure. This theory is supported by statistical data analysis and perplexity experiments carried out on the linguistically segmented Switchboard data. To take advantage of this internal sentence structure for conversational speech recognition, we proposed a training procedure that automatically extracts the "given" and "new" parts of the sentence from acoustically segmented speech utterances, and experimented with a few sentence models. Our results show that our given-new language model is a better fit to Switchboard conversational speech than a generic language model, as indicated by the decrease in both perplexity and word recognition error rate.

6. REFERENCES

[1] Clark and Haviland, "Comprehension and the Given-new Contract" in Freedle (ed.) Discourse Production and Comprehension, Ablex Publishing Corporation, New Jersey, 1977.

[2] M. Gavalda, K. Zechner, G. Aist, "High Performance Segmentation of Spontaneous Speech Using Part of Speech and Trigger Word Information," Applied Natural Language Processing, 1997.

[3] M. A. K. Haliday and R. Hasan, Cohesion in English, Longman, London, 1976.

[4] Daniel Jurafsky and et.al. "Switchboard Discourse Language Modeling Project," Johns Hopkins Summer Workshop, 1997.

[5] R. Kneser and V. Steinbiss, "On the Dynamic Adaptation of Stochastic Language Models," ICASSP 1993.

[6] M. Meteer and R. Iyer, "Modeling Conversational Speech for Speech Recognition," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, May 1996.

[7] A. Stolcke and E. Shriberg, "Automatic Linguistic Segmentation of Conversational Speech," ICSLP 1996.

[8] Y. Wang and A. Waibel, "Statistical Analysis of Dialogue Structure," EUROSPEECH 1997.