Using Detailed Linguistic Structure In Language Modelling

Ruijiang Zhang and Ezra Black and Andrew Finch
ATR Interpreting Telecommunications Laboratories
2-2 Hikarida Seika-cho, Soraku-gun
Kyoto, Japan 619-02
rzhang@itl.atr.co.jp
finch@itl.atr.co.jp
black@itl.atr.co.jp

Abstract
Recently, considerable attention has been accorded to attempts to apply natural language processing techniques to language modelling for speech recognition. Another extension to the standard n-gram technique has been the use of trigger-pair predictors. In the present experiments, we incorporate into language models, information derived from detailed syntactic and semantic parses and taggings. We use a human expert to define the interesting features of the history, and these are formalized as triggers and integrated with a trigram language model using the maximum entropy framework. We select maximum entropy because it provides a convenient method of combining multiple information sources. We employ two different kinds of triggering events: those based on a knowledge of the full parse of the previous sentences in the document, and those based on knowledge of the syntactic/semantic tags to the left of and in the same sentence as the word being predicted. We contrast results obtained using these events plus a baseline n-gram language model, both with the baseline model itself, and with the baseline model plus triggers based on word triggers chosen automatically. Mutual information selects the best trigger pairs from all candidates generated by combining each of these triggering events with every word in the vocabulary. The grammar and tagset used to express linguistic information about English are unusually detailed. The tagset contains some 3,000 syntactic/semantic tags. Using a 200-million-word training set composed of Wall Street Journal and Associated Press newswire text we reduced test-set perplexity by 11.3% as against the baseline model. Further, our method when combined with long-distance word triggers reduced test-set perplexity by 21.7%.

1 Introduction

It appears intuitively that information from earlier sentences in a document ought to help reduce uncertainty as to the identity of the next word at a given point in the document. (Rosenfeld, 1996) and (Lau et al., 1993) demonstrate a significant "word/word trigger-pair" effect. That is, given that certain "triggering" words have already occurred in a document, the probability of occurrence of specific "triggered" words is raised significantly.

The present paper undertakes to demonstrate that semantic/syntactic part-of-speech tags, and parse structure of previous sentences of the document being processed, can add trigger information to a standard n-gram language model, over and above the improvement delivered by word/word triggering along the lines of the work by Rosenfeld and Lau et al.1 We formulate "linguistic-question" triggers which query either: (a) the tags of the words to the left of, and in the same sentence as, the word being predicted; or (b) parse structure and/or tags within any or all of the previous sentences of the document to which the word belongs that is being predicted; or both of (a) and (b) together. Each of these questions then triggers a particular word in the vocabulary, i.e. raises the probability of that word's being the next word of the document.

As the source of both tags and parses in the present experiments, we use a 181,000-word subset of the approximately~1-million-word ATR General English Treebank (Black et al., 1996). This treebank sub-set consists exclusively of text drawn from Associated Press newswire and Wall Street Journal articles. The 181,000 words are partitioned into a training set of 107,000 words and a test set of 14,000 words. We utilize this portion only of the treebank, as opposed to the entire corpus, in order to match the text type of the raw data set used to train our baseline n-gram language model, which is AP and WJS text in roughly the same proportions as in our treebank, and of course not including any portion of our training or test text.

We train (i) a baseline 200-million-word n-gram language model; (ii) a model combining this baseline plus a word/word trigger model trained on a 10-million-word subset of the larger training corpus; and finally (iii) a model combining both (i) and (ii) with linguistic-question triggers trained as just indicated. Performance differences of (i)/(ii)/(iii) are measured, with the result that model (iii) is shown to yield a significant perplexity reduction vis-a-vis models (i) and (ii).

In what follows, Section 2 provides a basic overview
of the language modelling techniques employed; Section 3 discusses and offers examples of the linguistic questions of model (ii); Section 4 describes the language-modelling experiments we performed, and presents our experimental results; and Section 5 discusses our results and indicates future research directions.

2 The Language Model (LM)

2.1 ME Model

Our language model is a maximum entropy (ME) model of the following form:

\[
P(u|h) = \sum_{k=1}^{K} \alpha_k f_k(u, w) P_l(u|h_l) 
\]

where:
- \( u \) is the word we are predicting;
- \( h \) is the history of \( u \);
- \( \alpha_k \) is a normalization coefficient;
- \( K \) is the number of triggers;
- \( \alpha_k (k = 0, 1, \ldots, K) \) is the weight of trigger \( f_k \);
- \( f_k (i = 0, 1, \ldots, K) \) are trigger functions, \( f_k \in \{0, 1\} \);
- \( P_l(u|h_l) \) is the base language model.

In our experiments we use as base language models both a conventional bigram model and the extension of this model with long history word triggers. The improved iterative scaling technique (Della Pietra et al., 1997) is used to train the parameters in the ME model.

2.2 Trigger selection

The linguistic-question information is embodied in our model in the form of “triggers”. A trigger pair \( \langle q, w \rangle \) consists of a triggering question \( q \) together with a triggered word \( w \). The number of possible triggers is the product of the number of questions with the number of words in the vocabulary. This gives rise to too many features from which to build an ME model in a reasonable time. We therefore select only those trigger pairs which can be expected to provide the most benefit to the model. We use mutual information (MI) to select the most useful trigger pairs (for more details, see (Rosenfeld, 1996)). That is, we use the following formula to gauge a feature’s usefulness to the model:

\[
MI(q, w) = P(q, w) \log \frac{P(q|w)}{P(q)} + P(q) \log \frac{P(q)}{P(q|w)}
\]

where:
- \( w \) is the word we are predicting;
- \( q \) is a triggering feature (e.g. the answer to a linguistic question).

In the final trigger set, we use only those trigger pairs having the highest mutual information.

3 Linguistic Information

The experiments reported here consist in adding “linguistic-question constraints”\(^2\) to a baseline n-gram language model. To understand the linguistic questions used, one needs some familiarity with the ATR General English Treebank and the ATR General English Grammar and Tagset. For detailed presentations, see (Black et al., 1998; Black et al., 1997; Black et al., 1996). Briefly, however, each verb, noun, adjective and adverb in the ATR tagset includes a semantic label, chosen from 42 noun/adjective/adverb categories and 29 verb/interpersonal act categories. These semantic categories are intended for any “Standard-American-English” text, in any domain. Sample categories include: “physical attribute” (nouns/adjectives/adverbs), “alter” (verbs/verbal), “interpersonal act” (nouns/adjectives/adverbs/verbs/verbs), “organic” (proper nouns), and “zip code” (numericals). The semantic categorization is, of course, in addition to an extensive syntactic classification, involving some 165 basic syntactic tags.

The ATR English Grammar is unrestricted in its coverage, and particularly detailed and comprehensive, vis-a-vis other existing grammars. For instance, complete syntactic and semantic analysis is performed on all nominal compounds. Again, see the above-cited references for details.

Each parse of the ATR Treebank was entered by hand by a professional expert in parsing and tagging with the ATR English Grammar (Black et al., 1996). This Treebank is used as training data for an unrestricted-coverage parser of English (Black et al., 1997).

One can get a feel for the type of linguistic-question triggers we defined via Table 1, which shows three triggers with high mutual information with the word “Mrs.”, and three for “added”. The trigger with the highest mutual information with the word “Mrs.” among all linguistic-question triggers does not ask either about tags or parse structure, but simply makes good use, over raw text, of our “Question Language”, the flexible language for formulating grammar-based and lexically-based questions about Treebank text, which we normally use to compose contextual questions about text which we are parsing with our probabilistic parser.\(^3\) Specifically, the question, defined over raw text, determines which reference has been made to a female, within the last 12 sentences of the current document.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mutual Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>“was”</td>
<td>0.62</td>
</tr>
<tr>
<td>“word”</td>
<td>0.58</td>
</tr>
<tr>
<td>“word”</td>
<td>0.54</td>
</tr>
</tbody>
</table>

A question which asks about tags is question 2a of Table 1. It queries the semantic portion of tags

\(^2\)as well as “word/word triggers”\(^3\)For details, see (Black et al., 1997).
within the entire history of the document, and determines whether tags have frequently occurred which label nouns, adjectives or adverbs of saying, writing, objecting, or other verbal activities. A "yes" answer to this question turns out to raise the probability of the word "Mrs." as the next word of a document.

Finally, question 3b queries the complex parse structure of previous sentences of the document. The question tests whether frequently in the history of the document, sentences occurred with a human subject and a main verb of verbal activity, e.g. "Mr. Smith stated..." In addition, it tests the current sentence to see whether a human subject has just been received, and a verb now appears to be likely to occur. The expectation, thus, is that a verb of saying will now occur. This expectation turns out to be realized for the verb "added", as there is a relatively high correlation between a "yes" for this question and the occurrence of the word "added".

4 The Experiments

4.1 Experimental Procedure

We used the well-known trigram LM as the base LM for our experiments. This model was selected because it represents a respectable language model which most readers will be familiar with. The ME framework was used to build the derivative models since it provides a principled manner in which to integrate the diverse sources of information needed for these experiments.

In all models built for these experiments we use a word vocabulary of 20001 (the 20000 most frequent words plus a token for words not in the vocabulary). We used a corpus of newspaper text drawn from 1987-1996 Wall Street Journal and Associated Press Newswire in equal proportion. Certain types of words were mapped to generic tokens representing the class of word. These were: words representing time of day (e.g. 12:21), dates (e.g. 11/02/94), price expressions (e.g. $100) and year expressions (e.g. 1970-1999). The mapping was done using simple regular-expression pattern matching. The substitutions were implemented to assist the trigram model, which is unable to ask questions about the internal structure of words and cannot be expected to form useful n-grams from this class of words. The linguistic questions, however, being able to query the word's internal structure, were more effective on the raw words themselves and were used in that way. The vocabulary, and therefore the words being predicted, was constructed from data in which these tokens had been mapped.

The training set used to train the linguistic question-based triggers for all experiments was approximately 167,000 words of hand-labelled and parsed ATR treebank, drawn from Wall Street Journal and Associated Press texts. The test set consisted of 14,000 words of hand-labelled and parsed ATR treebank, again drawn in the same proportion from Wall Street Journal and Associated Press. We measure the test set perplexity (PP) to gauge the quality of the models produced.

<table>
<thead>
<tr>
<th>Model</th>
<th>Tri20M.k4</th>
<th>Tri100M.k4</th>
<th>Tri200M.k8</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigram</td>
<td>20001</td>
<td>20001</td>
<td>20001</td>
</tr>
<tr>
<td>bigram</td>
<td>395663</td>
<td>1230040</td>
<td>1204727</td>
</tr>
<tr>
<td>trigram</td>
<td>527782</td>
<td>2724346</td>
<td>2492309</td>
</tr>
</tbody>
</table>

Table 2: Trigram model size varying dataset size

<table>
<thead>
<tr>
<th>Model</th>
<th>Base PP</th>
<th>Base+Q's</th>
<th>Change(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tri20M.k4</td>
<td>155.0</td>
<td>142.7</td>
<td>6.7</td>
</tr>
<tr>
<td>Tri100M.k4</td>
<td>117.8</td>
<td>110.0</td>
<td>6.6</td>
</tr>
<tr>
<td>Tri200M.k8</td>
<td>108.0</td>
<td>101.0</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Table 3: Effect of varying dataset size

4.2 Effect of Dataset Size

In this experiment we used base trigram models of three differing sizes. The three models: Tri20M.k4 (k4 = cutoff of 4), Tri100M.k4 and Tri200M.k8 were built from 20M, 100M and 200M words of training data, respectively. Table 2 shows the number of n-grams we used in our models. Table 3 shows the reduction in perplexity. Note that here we used 33000 question-based triggers and the question set size from which the triggers were produced was 396.

In Table 3, "Base" is the perplexity of the base trigram model before any ME training. "Base + Q's" is the perplexity of the full ME model after training. "Change" is the perplexity reduction resulting from using our question triggers.

Notice that increasing the quality of the underlying trigram LM has little effect on the change in perplexity resulting from adding the information from linguistic questions. This indicates that the additional information will be useful to any trigram LM and that simply improving the LM by adding more data is no substitute for this information.

4.3 Effect of Adding Word Triggers

In this experiment we measure the effect of using long-range word triggers on our corpus together with the effect of combining these with our question-based triggers. 33067 long history word triggers are chosen by mutual information from 200 million words of data. Due to the prohibitively long training times needed to train models using word triggers we restricted the training set for the ME training to 10M words. The base language model was trained on the full 200M word corpus. We then used the ME model built by adding word-triggers to the base model as the base model for a second ME model which incorporated our question-based triggers. We found this approach effective in dealing with the large number of triggers involved. The number of question-based triggers used was 110,000 and the question set size from which the triggers were produced was 6,659. The results are shown in Table 4.
Table 1: Selected triggers from top-20-highest-MI linguistic-question triggers for the words “Mrs.” and “added”

<table>
<thead>
<tr>
<th>#</th>
<th>Question Description</th>
<th>MI (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>Any reference to a female within the last 12 sents of doc</td>
<td>0.001210</td>
</tr>
<tr>
<td>2a</td>
<td>Many nouns, adj or adv of verbal action (e.g. statement) within last 100 sents</td>
<td>0.000803</td>
</tr>
<tr>
<td>3a</td>
<td>Many nouns, adj or adv of helping (e.g. assistance) within last 100 sents</td>
<td>0.000737</td>
</tr>
<tr>
<td>1b</td>
<td>Any subject pronoun to the immediate left.</td>
<td>0.000579</td>
</tr>
<tr>
<td>2b</td>
<td>Subject of current sentence is a person and verb is likely</td>
<td>0.000407</td>
</tr>
<tr>
<td>3b</td>
<td>Many recent sents had person subjects and “saying” main verbs AND</td>
<td>0.000314</td>
</tr>
</tbody>
</table>

Table 4: The effect of combining the models

<table>
<thead>
<tr>
<th>Model</th>
<th>PP</th>
<th>Change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base ('Tri200M.18)</td>
<td>108.0</td>
<td>-</td>
</tr>
<tr>
<td>Base + WTM</td>
<td>94.4</td>
<td>12.6</td>
</tr>
<tr>
<td>Base + Q’s</td>
<td>95.8</td>
<td>11.3</td>
</tr>
<tr>
<td>Base + WTM + Q’s</td>
<td>84.6</td>
<td>21.7</td>
</tr>
</tbody>
</table>

5 Discussion

The maximum entropy framework adopted for these experiments virtually guarantees that models which utilize more information will perform as well as or better than models which do not include this extra information. Therefore, it comes as no surprise that all models improve upon the baseline model, since every model effectively includes the baseline model as a component. The experiments presented here have focused on showing that that we can glean useful information from the parse structure and part-of-speech tags in the history of the word being predicted. Our main result is that this information is useful, and is of similar magnitude to that provided by the long-range word triggers used by (Rosenfeld, 1996). Moreover, when these triggers are used in conjunction with a model incorporating long-range word triggers, almost all of the perplexity gain is inherited by the new model. This indicates that the information we are providing is largely new and complementary. This is in line with our intuition, given the nature of the questions we ask. Furthermore, we obtained this gain from a very small 167,000 word training corpus (as opposed to the 10 million word corpus used to train the long-range word triggers). It is reasonable to expect significant improvement on domains where more data is available to train from.

This work is a first attempt at exploiting the parse structure in the extraneous history to assist a language model. A major practical concern is that the predictions are being made from correctly analysed text rather than the output of a parsing device. Our intention in this paper was to show that there is useful information in the parses in the history. In further research, we intend to incorporate a real parsing device.

When a real parser is used, the system (including the grammarian writing the questions) will need to overcome the errors made by the parser/tagger. However, one point in favour of this approach is that if we train from the output of the parser (one way to learn to predict from only the reliable parts of the parse), we will have a much larger corpus from which to train the question-based component of the LM. Additionally, although we are currently able to ask quite sophisticated questions of the structure of parses in the history, we feel that we can realize considerable gain by further developing the language we are using to ask these questions, and thereby improving their expressive power.

6 Acknowledgments

The authors wish to thank J. Lafferty, Y. Sagisaka, T. Shalit, and S. Shirai for their helpful roles in the research reported above.

References


