Title	確率的シソーラスと文書クラスタに基づいたトリガー言語モデルの拡張による音声認識
Author(s)	Troncoso Alarcon, Carlos
Citation	
Issue Date	2003-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1653
Rights	
Description	Supervisor:下平 博, 情報科学研究科, 修士

An Extension to the Trigger Language Model Based on a Probabilistic Thesaurus and Document Clusters for Automatic Speech Recognition

Carlos Troncoso Alarcón (110089)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 14, 2003

Keywords: language model, speech recognition, probabilistic thesaurus, trigger language model, EM algorithm.

1 Introduction

The most widely used language model (LM) in automatic speech recognition (ASR) is the n-gram model. n-grams are very powerful in modeling dependencies between words that are adjacent or very near to each other within the text. However, they fail in modeling long-range dependencies between words, because they rely on a past word history limited to $n-1$ words.

One of the approaches that tried to cope with this limitation of n-grams is the trigger LM. This model uses a cache component similar to that of the cache-based LM, in which the most recent "rare" words are stored. In addition, a set of semantically related pairs of words called *trigger pairs*, constructed from a large text corpus by using the average mutual information measure, is also used. For every word in the cache, the model will predict a heightened probability not only for it, but also for all the words related to it through a trigger pair.

The drawback of the trigger LM is that its performance is very similar to that of the basic cache-based model, because most of the best triggers are the so-called *self-triggers* or triggers with the same root.

It seems reasonable to think that if the correlations between words were improved, we could have trigger pairs with a more significant effect in the overall system performance.

In this work, an extension of the trigger LM is proposed, in which, instead of trigger pairs, a probabilistic thesaurus of related pairs of words is used. In addition, a further extension is proposed, in which related words from document clusters are also extracted and incorporated into the cache.

2 Proposed Approach

2.1 Concept

The proposed approach uses two different information sources for extracting words related to the one currently being processed. These sources are a probabilistic thesaurus and document clusters.

The probabilistic thesaurus consists of words and related postposition + word pairs clustered in semantic classes, with their probability distributions (e.g. *densha* (train), *basu* (bus),... ↔ *ni noru* (to get on), *no untenshu* (driver),...). Each class is divided in two sets: a "leading words" set, i.e. words semantically related to each other, and a related words set, i.e. words related to the leading words set through a postposition. It was created automatically from a large text corpus by using a statistical parser and expectation maximization (EM) algorithm-based clustering. It captures the syntactic and semantic dependencies between strongly correlated words better than trigger pairs.

The document clusters consist of clusters of documents with similar contents along with words that are likely to appear in these documents, with their probability distributions (e.g. document 573, document 947,... ↔ *densha* (train), *eki* (station), *sen* (line),...). They also were created by means of EM-based clustering from the same text corpus, namely, five years of Japanese newspapers. They can specify the words that are likely to denote major topics in a set of similar documents.

For each word that is added to the cache, the most likely leading words and related words, without the postposition, from the most likely classes for that word in the probabilistic thesaurus are also added to the cache.

In addition, the most likely words from the most likely clusters for that word in the document clusters are also added to the cache if they are not already in it.

The main differences between the trigger LM and the proposed approach are the following.

First, the models use different data. The trigger pairs are pairs of well-correlated words that appear in similar contexts (e.g. education → academic). On the other hand, the probabilistic thesaurus groups pairs of words syntactically related through a postposition in semantic classes, while the document clusters are words and documents divided in semantic clusters. Both reflect different uses of words (e.g. *Daiei* can be the name of a department store or the name of a baseball team).

In addition, the proposed model should model better the syntactic and semantic relations between strongly correlated nouns and verbs (e.g. *biiru* (beer) ↔ *nomu* (to drink)), pairs of nouns (e.g. *Kyojin* (Giants) ↔ *toushu* (pitcher)), etc.

2.2 Methodology

The proposed approach rescores the N-best hypotheses output by an ASR system using the scores provided by the new LM.

The score of the proposed LM is the interpolation between the score of the extended cache component and the baseline LM score output by the speech recognizer, that is,

$$S(W) = S_{extended}(W)^{\lambda} S_{baseline}(W)^{1-\lambda} \tag{1}$$

where λ is the interpolation weight and W is the sentence being processed. In this way, one can take advantage of the short-range dependencies modeled by the baseline model and add the longer-range dependencies that the proposed model captures.

The score of the extended cache component is the normalized product of the cache score for all the words in the sentence:

$$S_{extended}(W) = \prod_{i=1}^{n} (S_{cache}(w_i))^{\frac{m}{n}} \tag{2}$$

where n is the length of W and m is the average length of the N-best sentences.

The cache score of a word is defined as the unigram probability inside the cache if the word belongs to the cache, and a value close to 0, ε, otherwise, as follows:

$$S_{cache}(w_i) = \begin{cases} \frac{N_{cache}(w_i)}{Cache\ Size} & N_{cache}(w_i) \neq 0 \\ \varepsilon & \text{otherwise} \end{cases} \tag{3}$$

where $N_{cache}(w)$ is the number of times w appears in the cache.

3 Experimental Results

Experiments with two different test sets of the same 71 sentences from two different male speakers were conducted. The test data consisted of an article about education from the Japanese Yomiuri Shimbun newspaper.

The ASR system Julius 3.1 was used to output the N-best hypotheses that the model rescores, where N was set to 100. This system uses a bigram LM for the first decoding pass and a trigram model for the second pass.

The maximum recognition accuracy that can be attained by choosing the best hypothesis from the N-best each time is 91.35%.

The number of significant classes from the 2500 in the probabilistic thesaurus was 5, and the number of significant leading words and significant related words for each class were also 5 each. The number of significant clusters from the 300 document clusters was 1, and the number of significant words for each cluster was 5. Therefore, for every word that is added to the cache, 55 related words are also added, and consequently, the cache size for the proposed model is 56 times the size of that for the standard cache-based model.

The speech recognition accuracy for the model with only the cache-based component and the extended trigger model based on the probabilistic thesaurus and the document clusters was computed for values of λ from 0 to 1 incremented by 0.05, and base cache sizes equal to 5, 10, 25, 50, 100, 250 and 500.

An absolute improvement of 0.53% over the baseline was obtained for a base cache size equal to 25, which represents a 13.5% of the total possible improvement.

4 Conclusion

In this research, an extension to the trigger LM has been proposed. Contrary to the original trigger LM, the proposed approach is based on two different knowledge sources, namely, a probabilistic thesaurus and document clusters. The former captures syntactic as well as semantic dependencies between words in the text, while the latter provides information about the current topic of discourse.

Experiments demonstrated that the related words that are extracted from the two knowledge sources successfully incorporate to the model constraints that help in the prediction process.