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

By Carlos Troncoso Alarcón

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Associate Professor Hiroshi Shimodaira

March, 2003

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

By Carlos Troncoso Alarcón (110089)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Associate Professor Hiroshi Shimodaira

and approved by
Associate Professor Hiroshi Shimodaira
Professor Masato Akagi
Associate Professor Kentaro Torisawa

February, 2003 (Submitted)

Abstract

In this work, an extension to the trigger-based language model (LM) for large vocabulary continuous speech recognition (LVCSR) is proposed. In this approach, instead of trigger pairs based on the average mutual information measure, related words extracted from a probabilistic thesaurus and document clusters, both created from a text corpus by using EM-based clustering, are used. The probabilistic thesaurus captures syntactic and semantic relations between words, while the document clusters can provide information about the topic of discourse. Short-range dependencies from the baseline bigram + trigram model and long-range dependencies from the extended trigger model are integrated by interpolating the models, and the resulting LM score is used to rescore N-best lists. A little improvement in speech recognition accuracy over both the baseline and the model with only a cache component was obtained for a Japanese newspaper task.

Contents

List of Figures

List of Tables

Acknowledgments

The present research would have been impossible to complete without the help of many people.

First of all, I would like to thank the Japanese Ministry of Education, Culture, Sports, Science and Technology for giving me the opportunity to study in Japan.

I would also like to give many thanks to Professor Hiroshi Shimodaira and Mitsuru Nakai for their constant guidance, advice and support, and Professor Shigeki Sagayama for encouraging and accepting me for coming to Japan.

I am also very grateful to Professor Kentaro Torisawa, for his helpful ideas, support and for constructing the data on which my research is based, and to Shigeki Matsuda, for his bright ideas, C libraries and constant support.

Thanks go also to Professor Masato Akagi for supporting my research.

Many thanks to Hiroki Morimoto and Youichi Dohi, for recording the evaluation data for my research, and to Hiroo Nishiyama, for supervising the recording.

I'd also like to thank Dr. Masafumi Nishimura and his colleagues at IBM Tokyo Research Laboratory, for their feedback and help.

I owe a considerable debt of gratitude to my wife, Remedios García Bonilla, for her encouragement, understanding, and patience during my research.

I am particularly thankful to God, who has helped me in every moment.

Last, but not least, I would also like to take this opportunity to thank all my fellows in the laboratory and many other people at JAIST for their sincere help and cooperation.

Chapter 1

Introduction

1.1 Background

Automatic Speech Recognition (ASR) is typically based on two stochastic models: the acoustic model and the language model (LM). LMs are an important part of ASR systems, because they model the linguistic relations among words in the utterance that is to be recognized.

The most widely used LM in ASR is the n-gram model, where n typically equals 2 (bigram model) or 3 (trigram model). n-grams model the occurrence probability of n consecutive words in the text, and their parameters are estimated from a large text corpus.

n-gram models 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. Nevertheless, it has proved very difficult to outperform these models, mainly due to their simplicity, and many attempts to model longer-range dependencies have resulted in a very little improvement in recognition accuracy.

One of the approaches that tried to cope with this limitation of n-grams is the trigger LM [32]. This model uses a cache component similar to that of the cache-based LM [25], 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.

1.2 Motivation

The drawback of the trigger LM is that its performance is similar to that of the basic cache-based LM, 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 to extract words related to the one being processed. In addition, a further extension is proposed, in which related words from document clusters are also extracted and incorporated into the cache.

The probabilistic thesaurus incorporates to the model syntactic and semantic dependencies between words, while the document clusters can provide information about the topic of discourse. By taking advantage of the different features of these knowledge sources, this approach aims to improve the concept of trigger pairs with stronger word correlations, in order to improve the overall recognition accuracy in a typical speech recognition system.

1.3 Thesis Organization

This thesis is organized as follows. First, chapter 2 presents an introduction to statistical LMs, including those models relevant to this work, as well as the N-best rescoring paradigm. Chapter 3 describes the proposed extension based on a probabilistic thesaurus, and shows the experimental results obtained for it. Chapter 4 proposes a further extension based on document clusters, with its experimental results. Finally, in chapter 5 the conclusions and directions for future works are presented.

Chapter 2

Language Modeling

2.1 Introduction

Language modeling is the attempt to characterize, capture and exploit regularities in natural language [32]. Natural language is extremely difficult to model formally, due to its inherent variability and uncertainty.

There are two main approaches to language modeling: statistical language modeling and knowledge-based language modeling. The statistical approach tries to capture regularities in language from large amounts of text in a process known as *training*. On the other hand, knowledge-based modeling uses a set of linguistic rules coded by experts, as well as domain knowledge, to assess the grammaticality of sentences.

The advantages of statistical language modeling over the knowledge-based approach are:

- Statistical models assign a probability to each possible sentence, while knowledge-based models usually only provide a "yes"/"no" answer to the grammaticality of a sentence. Probabilities convey much more information than such a simple answer. Moreover, spoken language is often ungrammatical.
- Statistical models can be unexpensively built from a great variety of domains, as soon as the training procedure has been implemented.
- Coding linguistic rules by hand can be tedious and sometimes erroneous.
- At runtime, knowledge-based models like parsers are more computationally expensive than statistical models.

Statistical language modeling has also some disadvantages:

- They do not capture the meaning of the text. Therefore, they may assign a high probability to nonsensical sentences.
- Statistical models require large amounts of training data, which are not always available.

- Statistical language modeling often do not make use of linguistic and domain knowledge, which sometimes can be very helpful.

Language modeling is useful in areas like Automatic Speech Recognition (ASR), machine translation and any other application that process natural language with incomplete knowledge. In this work, statistical language modeling is used for ASR.

2.2 Language Models in Automatic Speech Recognition

ASR is typically performed as follows. First, the features of the input speech signal are extracted via a spectral analysis. Then, based on the acoustic and the LM probabilities, the search for the best hypothesis is performed. The result of this search is the output of the ASR system. This paradigm is illustrated in figure 2.1.

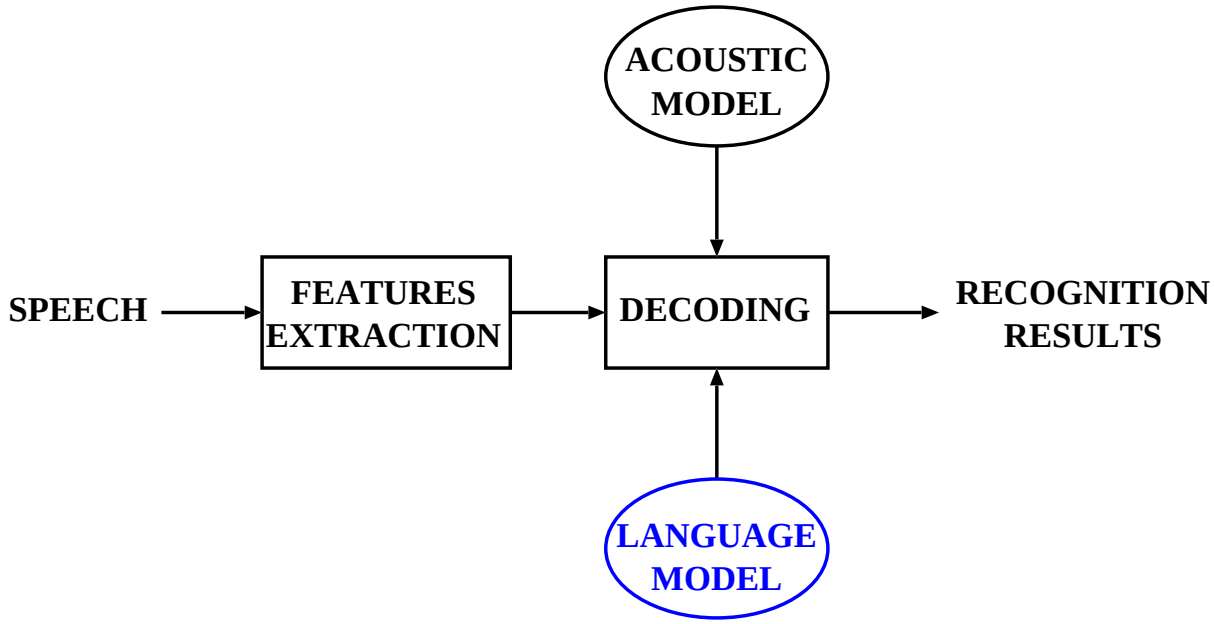

Figure 2.1: The automatic speech recognition paradigm

Probabilistically, the aim is to find the word sequence W that maximizes the probability of a word sequence given the acoustic signal A. Applying the Bayes rule,

$$\arg\max_W P(W|A) = \arg\max_W \frac{P(A|W)P(W)}{P(A)} = \arg\max_W P(A|W)P(W) \qquad (2.1)$$

The calculation of $P(A|W)$ is the role of the acoustic model, whereas the LM is responsible for the computation of $P(W)$.

Let $W = w_1^n \triangleq w_1, w_2, ..., w_n$, where the w_i's are the words that make up the word sequence. $P(W)$ can be decomposed, by using the chain rule, in the following way:

$$P(W) = P(w_1) \prod_{i=2}^{n} P(w_i | w_1^{i-1}) \tag{2.2}$$

Most statistical LMs try to estimate expressions of the form $P(w_i|h)$, where $h \triangleq w_1^{i-1}$ is known as the *history*.

Since the number of possible histories that can precede a given word is very large, it is unfeasible to try to estimate the probability of all of them from the limited corpora that are available. Therefore, some simplification must be applied to the above equation. Usually, the event space is partitioned in equivalence classes depending on some property of the text. For instance, in the trigram model the partition is based on the last two words of the history.

2.3 n-grams

An n-gram [1] is a model that uses the last $n-1$ words of the history as its sole information source. Typically n equals 2 or 3, and they are called *bigram* and *trigram* models, respectively.

As commented in the previous section, n-gram models partition the data into equivalence classes based on the last $n-1$ words of the history. Therefore, the following simplification is made:

$$P(w_i | w_1^{i-1}) = P(w_i | w_{i-n+1}^{i-1}) \tag{2.3}$$

In this way, a bigram estimates $P(w_i|h)$ by $P(w_i|w_{i-1})$, a trigram by $P(w_i|w_{i-2}, w_{i-1})$, and so on.

n-grams are affected by the classic modeling tradeoff between detail and reliability. When n is small, the parameters are reliably estimated from the training data, because the tuples are found easily. However, the modeling power is smaller than for greater values of n. On the other hand, when n is big, the data are insufficient and the estimates become unreliable. Nevertheless, the modeling power is bigger in this case.

The choice of n should depend on the amount of data available. For the sizes of the corpora typically available nowadays, trigrams own the best balance between reliability and detail, although interest is gradually moving towards 4-grams and beyond.

These models are easy to implement and easy to interface to the ASR decoder. They are very powerful and difficult to improve, mainly because of their simplicity. They seem to capture well short-range dependencies. It is for these reasons that they have become the standard LMs in ASR.

Unfortunately, they also have their drawbacks. First, they are unaware of any phenomenon or constraint that is outside their limited scope. Therefore, they may assign high probabilities to nonsensical and even ungrammatical utterances, as long as they satisfy local constraints. In addition, the predictors in n-gram models are defined by their order

in the sentence, not by their linguistic properties. Therefore, histories like "the fireman extinguished the" and "the fireman extinguished quickly the" are very different for a trigram, even though they are very likely to precede the same word.

2.4 Alternatives to n-grams

There are many works in the literature that tried to overcome the mentioned limitations of n-grams. Below is a description of the most interesting approaches classified by the length of the scope they cover.

2.4.1 Short Distance

Class-based n-grams [3] are n-grams whose parameter space has been reduced by clustering the words into classes. The n-grams are then based on these classes, rather than the words themselves.

If it is assumed that each word w belongs to only one class $g(w)$, then this model can take many forms, for example,

$$P(w_i|h) = P(w_i|g(w_{i-2}), g(w_{i-1})) \quad (2.4)$$

$$P(w_i|h) = P(w_i|g(w_{i-2}), w_{i-1}) \quad (2.5)$$

$$P(w_i|h) = P(g(w_i)|g(w_{i-2}), g(w_{i-1}))P(w_i|g(w_i)) \quad (2.6)$$

In practice, it is the last one that is the most used in class-based n-grams.

The clustering method itself can also take many forms.

Firstly, the clustering can be based on the linguistic knowledge. The best known example of this method is clustering by part of speech (POS). POS clustering attempts to capture syntactic dependencies between adjacent words in the text. This approach has several problems, though: some words can belong to more than one POS, POS classifications made by linguists may not be optimal for language modeling, and there are many different schemes for POS classification.

In second place, in clustering by domain knowledge, all words that will behave in a similar fashion are manually grouped together. For example, days of the week, numbers, etc. This approach can be specially helpful when the amount of training data is limited.

Finally, in data-driven clustering, a large amount of data is used to automatically derive classes by statistical means. This is often better than clustering by hand based on one's intuition. However, reliance on data instead of on external knowledge sources can also be problematic. For example, if the amount of training data available is not large enough, the resulting classes may not be reliable. The ideal data-driven clustering would be one supervised by an expert.

Class-based n-grams have advantages over the basic n-grams. Since the possible number of histories is reduced, the model becomes more compact. Therefore, it could be expanded to include more context. For example, a class-based 4-gram model might be approximately

the same size as a trigram. In addition, since the number of classes is generally smaller than the size of the vocabulary, the data sparsity is reduced, and even if a word n-gram is not found in the training data, the equivalent class-based n-gram is likely to have been seen. For this reason, these models have been very helpful in situations where the training data available were limited.

The disadvantage of these models is that they lose some of the semantic information that word n-grams, however, capture. This can be partially overcome by constructing LMs that incorporate information from both word and class-based n-grams. A more important drawback of class-based n-grams is that they don't solve the locality problem of n-grams.

2.4.2 Intermediate Distance

Long-distance n-grams [14] attempt to capture the dependencies between the predicted word and $n-1$-grams that are some distance back in the history. For instance, a distance-2 trigram predicts w_i based on (w_{i-3}, w_{i-2}). Distance-1 n-grams are consequently the conventional n-grams themselves.

These models have very serious limitations. Even though they capture dependencies between words that are separated by distance d, they cannot use different values of d at the same time during training, therefore, they unnecessarily fragment the training data. In other words, they do not pay attention to the nature of the text in order to decide an appropriate value for d, but they simply skip the words that are nearer than d words back in the history.

2.4.3 Long Distance

Mixture-based language models [5, 15] are composed of several LMs, each of which is specific to a particular topic or sublanguage. The probability distributions from these component LMs are linearly interpolated to form the global LM probability. The interpolation weights reflect, at each moment, which component sublanguage is currently being recognized.

Let $M_1, M_2, ..., M_k$ be the component LMs. The overall LM probability is then

$$P(w_i|w_1^{i-1}) = \sum_{j=1}^{k} \lambda_j P_{M_j}(w_i|w_1^{i-1}) \tag{2.7}$$

where the λ_j's are the interpolation weights, with values such that

$$\sum_{j=1}^{k} \lambda_j = 1 \tag{2.8}$$

Usually, the first step when creating a mixture-based LM is the clustering: the training data has to be partitioned in homogeneous components. This can be done automatically,

with some iterative clustering algorithm, or manually, according to the topic, style of text, etc.

The number of clusters in which the training data should be partitioned is a delicate matter. A number too small will result in a model incapable of discerning between topics or linguistic styles in detail. Too large a number will lead to a bunch of undertrained models with poor probability estimates. It is common that one of the components be the whole training data, in order to smooth the estimates and avoid data fragmentation.

The next step is typically to construct an n-gram model for each of the constituents. Then, the interpolation weights can be calculated by using the expectation maximization (EM) algorithm [10] in such a way as to maximize the likelihood of some held-out data.

These LMs are theoretically very attractive and represent a sound approach to LM adaptation. However, they have not significantly improved speech recognition accuracy so far.

Inside this category are also the **cache-based language model** and the **trigger language model**, which are presented in the next section.

2.5 Language Models Relevant to this Research

From the various alternatives to n-grams presented above, the present research is particularly based on the trigger LM, which in turn is based on the cache-based LM.

Both models are presented in this section.

2.5.1 The Cache-Based Language Model

The cache-based LM [25, 26] is based on the observation that a word that has appeared recently in a document has a high probability of reappearing.

A cache memory similar to that of computers is used to store the words of recent appearance. The word probabilities are estimated from their recent frequency of use. If a candidate word is in the cache, its probability is raised.

Typically, a cache-based component is linearly interpolated with an n-gram LM:

$$P(w_i|w_1^{i-1}) = \lambda P_{cache}(w_i|w_1^{i-1}) + (1-\lambda)P_{n-gram}(w_i|wi-n+1_{i-1}) \tag{2.9}$$

Usually, a cache of the last K words is maintained, and the cache-based probability of a word is computed as the unigram probability of the word within the cache, that is,

$$P_{cache}(w_i|w_1^{i-1}) = \frac{N_{cache}(w_i)}{K} \tag{2.10}$$

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

Figure 2.2 shows the outline of the cache-based model.

The original cache-based model was interpolated with a class-based trigram based on the POS, and a cache of size 200 was maintained for each POS. The interpolation weights were calculated individually for each POS.

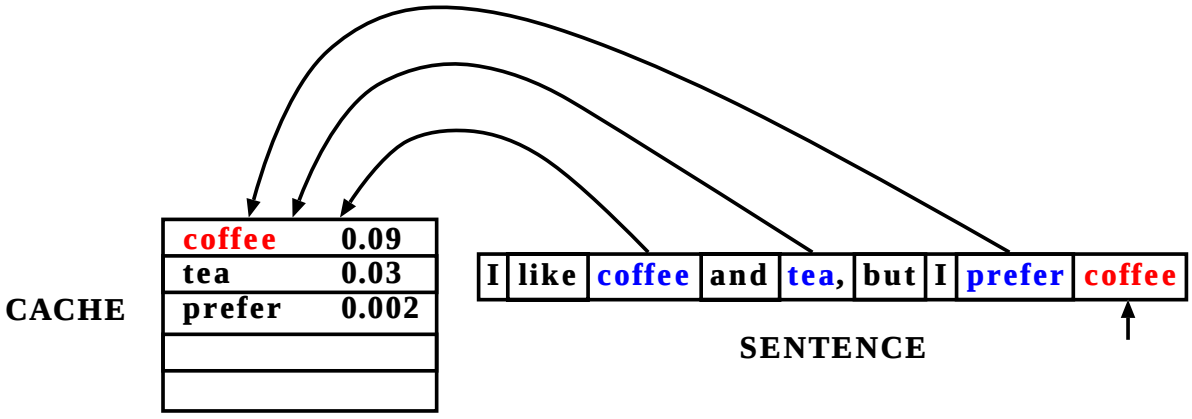

Figure 2.2: Cache-based language model

Several extensions have been proposed to this LM, being the most obvious the addition of the cache-based component to a word-based trigram, rather than a class-based model [15].

The cache need not be limited to containing single words. Instead, recent bigrams and trigrams can also be incorporated to the cache and their probabilities boosted [18]. This approach has the problem that the probabilities of n-grams in the cache cannot be reliably estimated due to the insufficient information contained in several hundred words back.

Another extension used the idea that the more recent words are more influential in predicting forthcoming words than those in the more distant past [5]. With this in mind, an exponentially decaying cache was constructed. This is a cache in which the probability of the words inside the cache decay exponentially with the distance from the word being predicted.

The cache-based LM significantly reduces the perplexity of standard LMs, and some of the extensions mentioned above contributed to a further improvement in terms of perplexity. However, the same does not apply to recognition accuracy, which has not been noticeably improved by this model so far.

2.5.2 The Trigger Language Model

The trigger model [32, 33], like the cache-based model, also uses a cache memory of recent words. However, contrary to the original cache-based model, only "rare" words are incorporated to the cache. A word is defined as rare relative to a threshold of static unigram frequency.

In order to extract information from the document history, a basic information bearing element called *trigger pair* is used. If a word a is semantically well correlated with another word b, then $(a \rightarrow b)$ is called a trigger pair, with a being the trigger and b the triggered word. When a occurs in the cache, it triggers b, and the model will predict a heightened probability not only for a, but also for b.

The trigger pairs are created from a big text corpus by using the average mutual information measure:

$$\begin{aligned} I(a;b) &= P(a,b)\log\frac{P(b|a)}{P(b)} + P(a,\bar{b})\log\frac{P(\bar{b}|a)}{P(\bar{b})} \\ &+ P(\bar{a},b)\log\frac{P(b|\bar{a})}{P(b)} + P(\bar{a},\bar{b})\log\frac{P(\bar{b}|\bar{a})}{P(\bar{b})} \end{aligned} \tag{2.11}$$

The model is formulated as a constraint of a maximum entropy (ME) framework [9, 16] in which n-grams, long-distance n-grams and so on can also take part as constraints of the model.

The outline of this model is depicted in figure 2.3.

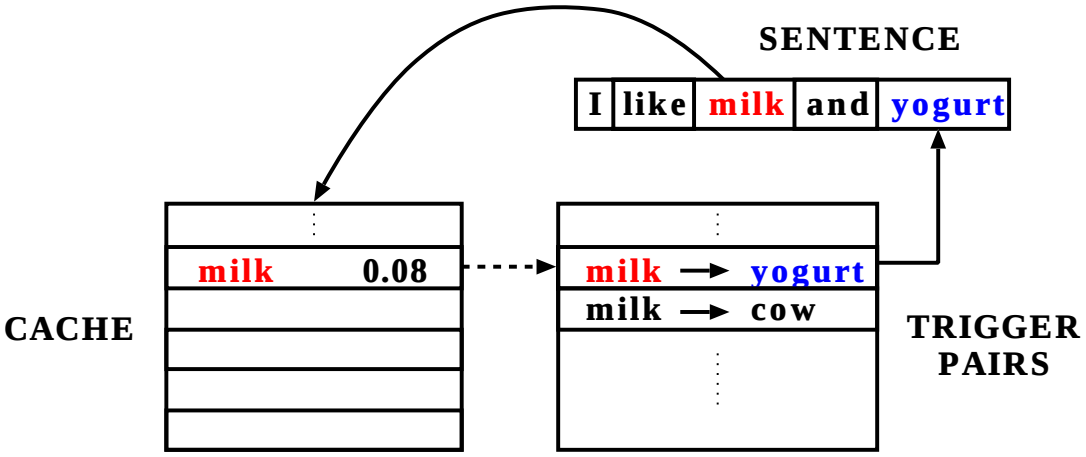

Figure 2.3: Trigger language model

The drawback of trigger pairs is that far more information is contained in the *self-triggers*, that is, words that trigger themselves, than in any others; even the non-self-triggers tend to be triggers with the same stem (e.g. abuse, abused, abusing). Therefore, the improvement over the basic cache-based model is small.

2.6 N-best Rescoring

Most LMs that try to overcome the limitation of n-grams use a standard trigram or bigram-based speech recognizer to output the N-best list, that is, the N most likely hypotheses. Then, based on a combination of the scores provided by the speech recognizer and the new scores assigned by an alternative (generally more complex) LM, they perform a rescoring of the N-best list, reordering the hypotheses and proposing the most likely hypothesis as the output of the whole recognition process.

This process is called *N-best rescoring*, and it is widely used in language modeling for ASR, because of its easy implementation and fast evaluation.

In this work, N-best rescoring is also used.

2.7 Summary

An introduction to language modeling, with the two main approaches to language modeling and their pros and cons, has been presented in this chapter. The application of LMs to ASR, n-grams as the standard LMs and some alternatives to them have been discussed, with special emphasis given to the models relevant to this thesis. Finally, the N-best rescoring paradigm has been introduced.

In the next chapter, the proposed approach in detail is presented.

Chapter 3

Extension Based on a Probabilistic Thesaurus

3.1 Probabilistic Thesaurus

The probabilistic thesaurus [31, 43] consists of sets of words and related postposition + word pairs clustered in semantic classes, with their probability distributions (e.g. *densha* (train), *basu* (bus),... $\leftrightarrow$ *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 (see Appendix A).

This thesaurus was automatically created from a large text corpus, namely, five and nine years of two Japanese newspapers, by using a statistical parser and EM algorithm-based clustering. Figure 3.1 illustrates this procedure.

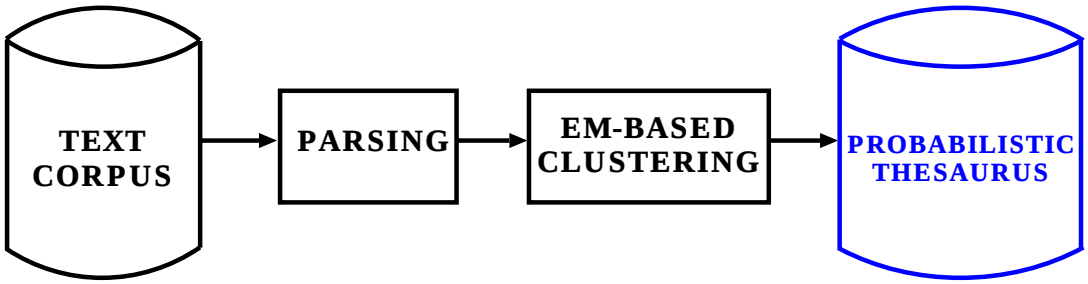

Figure 3.1: Construction of the probabilistic thesaurus

This method used triples of the form $< r, rel, l >$ as learning data, where r was a related word, l a leading word and rel the relationship between r and l. rel could be a postposition, a relative clause marker or an empty marker. The relative clause marker refers to the relation between the head verb of a given relative clause and its head noun. The empty marker expresses the relation between words related to each other without the help of a postposition.

Each triple was divided in two items, $< r, rel >$ and l. The probability that the triple occurred was defined as follows:

$$P(< r, rel, l >) \triangleq \sum_{a \in A} P(< r, rel > |a) P(l|a) P(a) \tag{3.1}$$

where a denoted a class of the occurrences and A the set of all classes. The number of classes was fixed a-priori.

The EM-based clustering method estimates $P(< r, rel > |a)$, $P(l|a)$ and $P(a)$ for each related word r, leading word l, relationship rel and class a. Unfortunately, this estimation is not straightforward, because the class a is not observed in the training data.

The following iterative algorithm was used to estimate the probabilities. First, a statistical parser was used to obtain a set of parse trees that capture the co-occurrence relations observed in the corpus. Then, the following list was created:

$$Q = \{< r_0, rel_0, l_0 >, < r_1, rel_1, l_1 >, ..., < r_n, rel_n, l_n >\} \tag{3.2}$$

The likelihood that Q is observed was calculated by the following formula:

$$\prod_{< r_i, rel_i, l_i > \in Q} P(< r, rel, l >) = \prod_{< r_i, rel_i, l_i > \in Q} \left\{ \sum_{a \in A} P(< r_i, rel_i > |a) P(l_i|a) P(a) \right\} \tag{3.3}$$

The EM algorithm maximized the above probability by adjusting the parameters $\{P(< r, rel > |a) | r \in R, rel \in Rel, a \in A\} \cup \{P(l|a) | l \in L, a \in A\} \cup \{P(a) | a \in A\}$ iteratively, where R is the set of all related words, Rel the set of all relations and L the set of all leading words. The iteration continued until convergence or near-convergence of the likelihood.

The probabilities at the j-th iteration step were calculated as follows:

$$P_j(a| < r, rel >, l) = \frac{P_j(a) P_j(< r, rel > |a) P_j(l|a)}{\sum_{a' \in A} P_j(a') P_j(< r, rel > |a') P_j(l|a')} \tag{3.4}$$

Based on the above formula, the probabilities at step $j+1$ are computed in the following way:

$$P_{j+1}(a) = \frac{1}{|L|} \sum_{< r_i, rel_i, l_i > \in Q} P_j(a| < r, rel >, l) \tag{3.5}$$

$$P_{j+1}(< r, rel > |a) = \frac{\sum_{< r, rel, l_i > \in Q} P_j(a| < r, rel >, l_i)}{\sum_{< r_i, rel_i, l_i > \in Q} P_j(a| < r_i, rel_i >, l_i)} \tag{3.6}$$

$$P_{j+1}(l|a) = \frac{\sum_{< r_i, rel_i, l > \in Q} P_j(a| < r_i, rel_i >, l)}{\sum_{< r_i, rel_i, l_i > \in Q} P_j(a| < r_i, rel_i >, l_i)} \tag{3.7}$$

The probabilities of the last iteration are the output of the entire learning process.

The probabilistic thesaurus captures the syntactic and semantic relations between strongly correlated words better than trigger pairs.

In the proposed approach, for each word that is added to the cache, the most likely leading words and the most likely related words, without the postposition, from the most likely classes related to that word in the probabilistic thesaurus are also added to the cache. In this way, these words are incorporated into the cache component of the proposed model, so that they can help improve the predictors.

If a word is a verbalized noun + *suru* or an inflected form of a verb, only the base form is used. For example, from *benkyou suru* (to study) only *benkyou* (study) is used, and from *tsukawareru* (to be used) only the base form *tsukau* (to use) is employed. By applying this generalization, the algorithm can compare the base form of the verbs in the cache with that of the verbs in the hypotheses during the N-best rescoring procedure, and thus, the prediction power of the verbs is raised. For example, in the sentences *terebi wo miru* (I watch TV) and *terebi wo mita* (I watched TV), it seems reasonable that the correlation between *terebi* (TV) and *miru* (to watch) should be used in both cases. Furthermore, when looking up in the probabilistic thesaurus, it is also desirable to use the base form of verbs, as we do when we look up a word in a dictionary.

Figure 3.2 shows the outline of the proposed model.

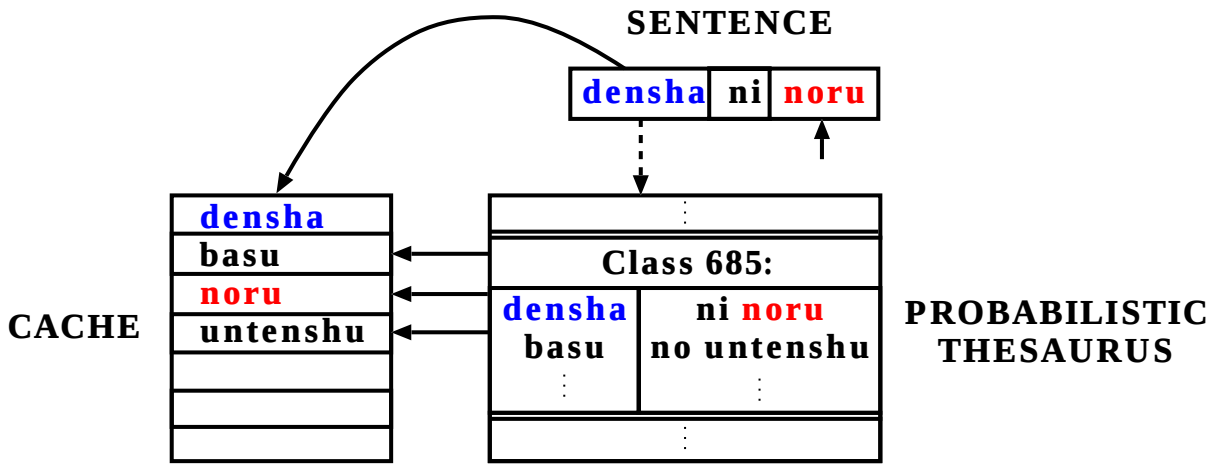

Figure 3.2: Outline of the extension based on a probabilistic thesaurus

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 can be found 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. It reflects 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) $\leftrightarrow$ *toushu* (pitcher)), etc.

3.2 Methodology

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

The rescoring algorithm proceeds as follows.

1. Read 1-best sentence
2. Add to the cache all words in the sentence
3. For each word in the sentence, add to the cache the most significant related words from the probabilistic thesaurus
4. For each hypothesis in the N-best:
   - Calculate the score of the extended cache component
   - Calculate the score of the proposed LM
   - Calculate the total score (acoustic model + proposed LM)
5. Output the hypothesis with the highest total score

Here, by "significant" I mean "with the highest probability".

The total score is the score of the acoustic model output by the speech recognizer times the score of the proposed 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,

$$S(W) = S_{extended}(W)^{\lambda} S_{baseline}(W)^{1-\lambda} \tag{3.8}$$

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.

We define the score of the extended cache component as the normalized product of the cache score for all the words in the sentence. Since the length of the sentences within the N-best is variable, the score needs to be normalized. I propose the following way:

$$S_{extended}(W) = \prod_{i=1}^{n} (S_{cache}(w_i))^{\frac{m}{n}} \tag{3.9}$$

where w_i are the words that compose W, n is the length (number of words) of W and m is the average length of the N-best sentences.

For every word, the cache score 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:

$$S_{cache}(w_i) = \begin{cases} \frac{N_{cache}(w_i)}{Cache\ Size} & N_{cache}(w_i) \neq 0 \\ \varepsilon & \text{otherwise} \end{cases} \tag{3.10}$$

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

3.3 Experimental Environment

Experiments with two different test data sets of the same 71 sentences from two different male speakers were conducted. These test data consisted of an article about education from the Japanese Yomiuri Shimbun newspaper (see Appendix B). These data were not used to create the probabilistic thesaurus or the document clusters.

The ASR system Julius 3.1 [21] was used to output the N-best hypotheses that the model rescores, where N was set to 100. This system performs a two-pass (forward-backward) search using a back-off bigram and a back-off trigram model in the respective passes, with a cut-off threshold of 1 for both models. These models were trained from 75 months (01/1991-09/1994, 01/1995-06/1997) of the Japanese Mainichi Shimbun newspaper. A vocabulary of 21322 words was used.

The recognition accuracy for this baseline model was 89.72% for test set 1 and 85.10% for test set 2. The average recognition accuracy of the baseline model is thus 87.41%.

The maximum recognition accuracy that can be attained by choosing the best hypothesis from the N-best each time is 93.54% for test set 1 and 89.16% for test set 2. Therefore, the average maximum attainable accuracy is 91.35%.

The value of ε in equation 3.10 was set to 10^{-30}.

Five years (1991-1995) of the Japanese Mainichi Shimbun newspaper and nine years (1990-1998) of the Japanese Nihon Keizai Shimbun newspaper were used to construct the probabilistic thesaurus.

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. Therefore, for every word that is added to the cache, 50 related words are also added, and consequently, the cache size for the proposed model is 51 times the size of that for the model with only a cache component.

The previous parameters were empirically tuned.

The speech recognition accuracy for the model with only the cache component and the extended trigger model based on the probabilistic thesaurus 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.

The experimental environment is summarized in table 3.1.

3.4 Experimental Results

Many experiments were carried out in this research. Some were used to tune the model parameters, like those described in the previous section, others used a-priori assumptions,

ASR system	Julius 3.1
Vocabulary size	21322 words
N (number of output hypotheses)	100
Test data	Source: Yomiuri Shimbun Number of sentences: 71 Topic: Education
Number of speakers	2 male speakers
Average baseline accuracy	87.41%
Average maximum attainable accuracy	91.35%
Number of significant classes	5
Number of significant leading words	5
Number of significant related words	5
Base cache size	5, 10, 25, 50, 100, 250, 500
ε	10^{-30}

Table 3.1: Experimental environment for the extension based on the probabilistic thesaurus

which turned out to be erroneous, and others tried different data or different methods of combining the model parameters.

In this section, the basis for these experiments and the results obtained are described.

3.4.1 Final Model

The proposed model was tested against a model that only uses the cache component, that is, it does not add to the cache any related words.

The maximum recognition accuracy of the proposed model was obtained for a cache size of 25 for both data sets, and the average of the results of the experiments from the two sets for this cache size is shown in figure 3.3.

It can be observed that, for certain values of λ, the extended trigger model based on the probabilistic thesaurus has a higher recognition accuracy than both the baseline and the model with the cache component alone. An improvement of 0.31% (absolute) over the baseline was obtained, which represents a 7.9% of the total possible improvement.

This improvement, although not very significative, may mean that the related words extracted from the probabilistic thesaurus constitute indeed a useful external source for the LM. In the next chapter, an experiment will try to confirm the usefulness of these related words.

3.4.2 Stop List

It is not hard to realize that function words, like Japanese postpositions or auxiliary verbs, are somewhat uniformly distributed all over a given text. On the other hand, the frequency of appearance of content words usually depends on linguistic properties of the

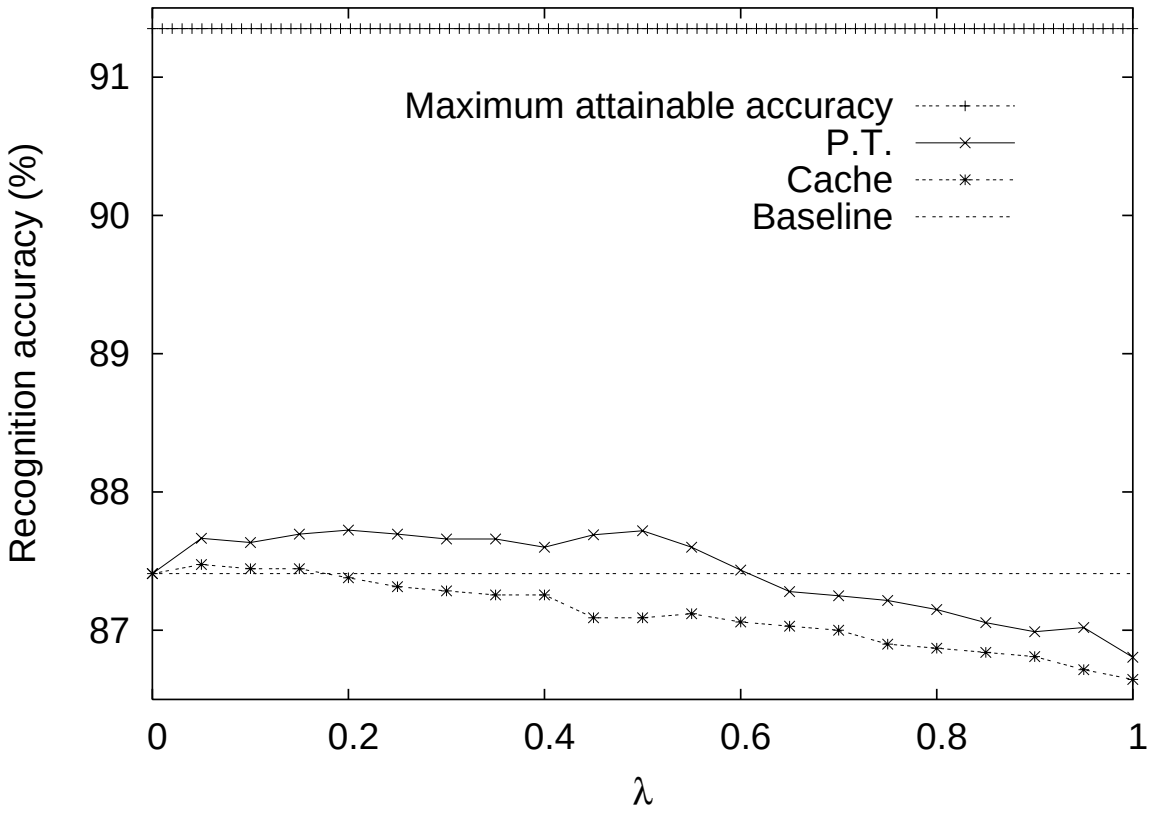

Figure 3.3: Speech recognition accuracy of the extension based on a probabilistic thesaurus, for different values of λ and a base cache size equal to 25

text like the topic of discourse, the type of text, etc. Consider, for example, the word "pitcher". It is obvious that it will appear with a greater frequency in a sports article than in one about religion.

Initially in this research, the assumption that content words would have a greater impact in the cache than function words was made. Therefore, a stop list, that is, a list of function words that are not incorporated into the cache nor used to look for related words, was constructed.

The stop list was created from a large text corpus of Japanese newspapers by inserting in the list all the words with a frequency higher than or equal to 260. This threshold was fixed by hand to meet the requirement that no content words, except those that frequently appear in Japanese newspapers like *nihon* (Japan), *keizai* (economy) or *beikoku* (America), appeared in the list. The list contains 92 words.

Experiments refuted this hypothesis, as shown in figure 3.4, where it can be seen that the model with an empty stop list performed much better than the one with the non-empty list.

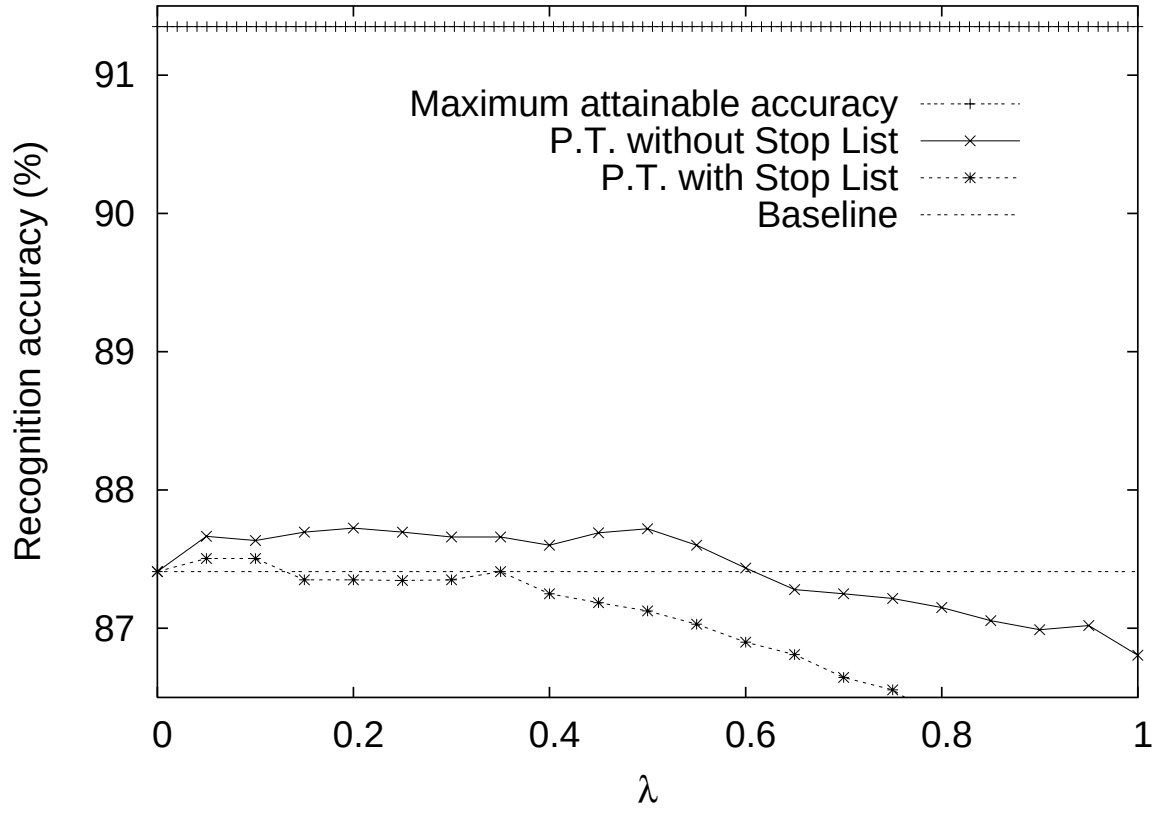

Figure 3.4: Speech recognition accuracy of the model with the stop list vs. the model without the stop list, for different values of λ and a base cache size equal to 25

The possible reason for this counterexample can be that the absence of function words

in the cache makes the cache component assign a very low probability to function words in the hypothesis, therefore, sentences with fewer function words, like erroneous sentences with less function words than normal, are assigned higher probabilities, and consequently the recognition accuracy drops.

3.4.3 Independent Components

In the original trigger LM, the words triggered by those in the cache are not incorporated into the cache, as opposed to the proposed approach. In the original model, the trigger pairs are stored apart from the cache, and the words that are in the cache are looked up in them.

On the other hand, in the proposed model, the words that are added to the cache are looked up in the probabilistic thesaurus, and the related words are inserted in the cache.

Two different experiments were performed, in which analogously to the original trigger model, the related words were not incorporated into the cache.

In the first experiment, for all the words in the cache, a list of related words was generated and stored apart. Then, the cache score and the trigger score were computed separately and interpolated with the baseline score by means of the following formula:

$$S(W) = S_{cache}(W)^{\lambda_1} S_{related}(W)^{\lambda_2} S_{baseline}(W)^{\lambda_3} \tag{3.11}$$

The cache score was computed in the same way as in equation 3.10.

The score of related words was defined as follows:

$$S_{related}(W) = \prod_{i=1}^{n} (S_{related}(w_i))^{\frac{m}{n}} \tag{3.12}$$

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

$$S_{related}(w_i) = \begin{cases} \frac{N_{related}(w_i)}{Cache\ Size} & N_{related}(w_i) \neq 0 \\ \varepsilon & \text{otherwise} \end{cases} \tag{3.13}$$

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

This model did not perform better than the final one for any of the cache sizes tried.

In the second experiment, the words with more recent appearance were stored both in the cache and in a word buffer smaller than the cache. Then, instead of creating the list of related words from the cache, it was created by looking up the words in the buffer in the thesaurus. In this way, the size of the cache is made independent of the number of words that take part in the thesaurus lookup.

The scores were computed as in the experiment above.

Analogously to the previous experiment, this approach did not help to improve the recognition accuracy.

There can be several reasons for the underperformance of these two approaches. One possible cause can be in the interpolation scheme: the interpolation weights might not be optimal, or the magnitude of the scores could be very different. Another possible problem

can be the score of the related words itself. It has been defined exactly in the same way as the cache score, but it might not be useful to use the unigram probability of the related words. The usage of the probability distributions of the words in the probabilistic thesaurus seems more reasonable. These matters will be dealt with in future works.

Using the same word buffer as that of the last mentioned approach, another experiment was tried, where the model components were no longer independent, but the only related words that were added to the cache were those triggered by the words in the buffer. Thus, in this approach the cache size is also independent of the number of words that are looked up, but all the words end up together inside the cache.

Different sizes for the cache and the word buffer were tried. The sizes that achieved the best results were 1250 for the buffer and 500 for the cache. Observe that 1250 is 25 times 50, that is, almost the same size as that of the cache of the final model. The results for these sizes are presented in figure 3.5.

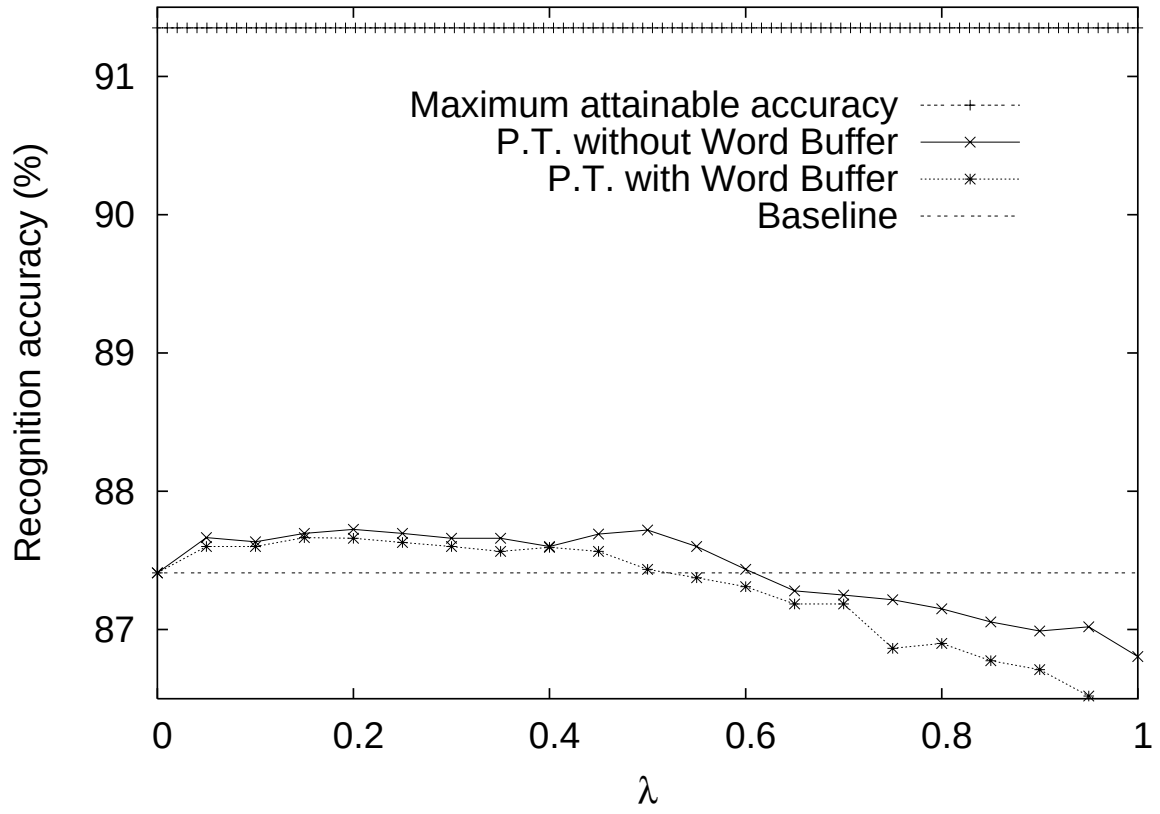

Figure 3.5: Speech recognition accuracy of the model with a buffer, for different values of λ, a base cache size equal to 500 and a buffer size equal to 1250

As it can be seen, this approach did not outperform the proposed model. As the parameters were made closer to those of the final model, the recognition accuracy was improved. This means that the final model behaves better than this approach.

3.4.4 Words in 1-best Sentence

As mentioned in section 3.2, the proposed model uses the words in the 1-best sentence for the cache component, and the related words from the probabilistic thesaurus are based on them too.

The 1-best sentence normally contains errors. If it didn't, it would not be necessary to improve the LM. It may then seem unappropriate to use the words in that sentence for the cache component, because a misrecognized word in the cache may contribute to incur the same error again, since the probability of the erroneous word may be raised over that of the correct one.

With this in mind, a modification to the algorithm was made. In it, instead of using the words in the 1-best sentence, the words that appeared more than 10 times within the 20-best sentences were used. This figures were optimized empirically.

The results of this experiment are shown in figure 3.6.

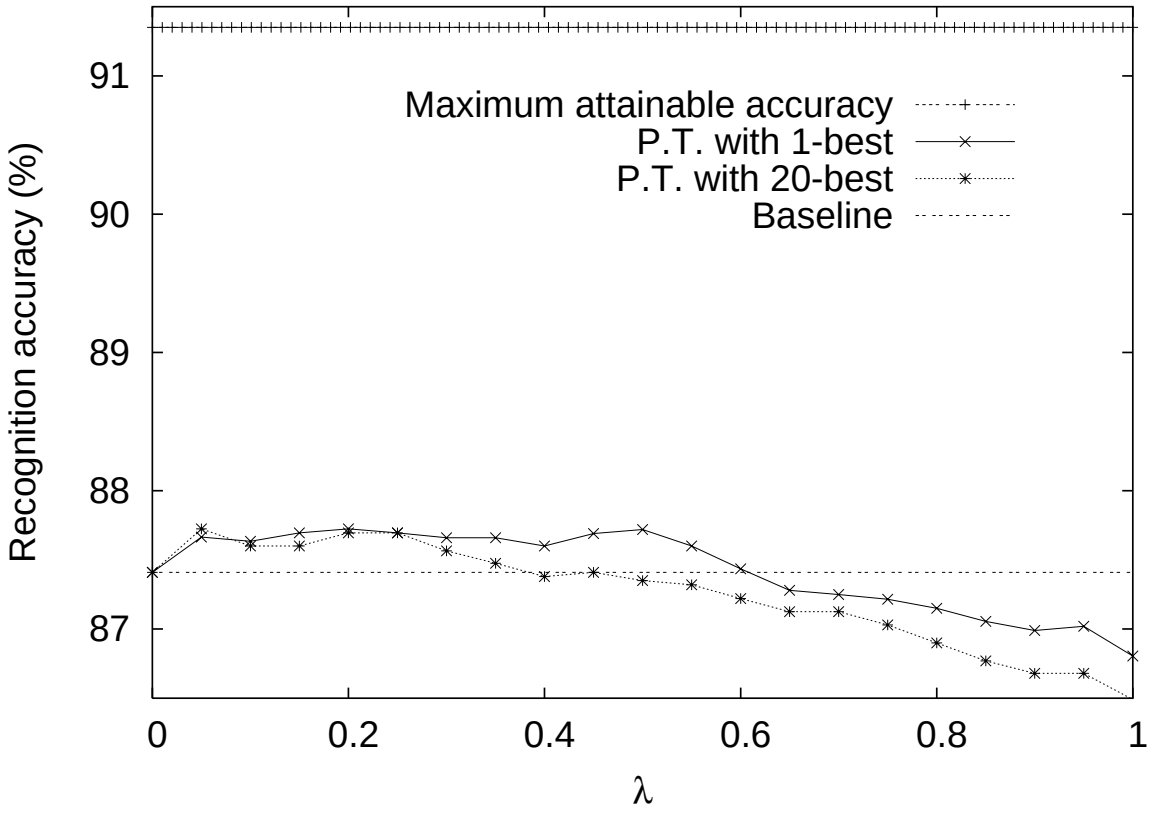

Figure 3.6: Speech recognition accuracy of the model that uses the 1-best vs. the model that uses the 20-best, for different values of λ and a base cache size equal to 25

It didn't come as a surprise the fact that the described modification did not improve the results, because there are several works with similar results [7].

3.5 Summary

In this chapter, an extension to the trigger LM based on a probabilistic thesaurus has been presented. The features of the probabilistic thesaurus, the main differences with respect to the original trigger LM and the methodology of the proposed approach have been discussed. Finally, the experimental environment used in this research, as well as several different experiments, with their corresponding results, have been detailed.

In chapter 4, an additional extension to the approach that has been discussed here is presented.

Chapter 4

Further Extension Based on Document Clusters

4.1 Document Clusters

The document clusters [13] 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,... $\leftrightarrow$ *densha* (train), *eki* (station), *sen* (line),...).

They were also created by means of EM-based clustering from a text corpus different to that used for the probabilistic thesaurus, in this case, five years of a different Japanese newspaper.

The algorithm used for creating these data was the same as in the probabilistic thesaurus, except that, in this case, the method used pairs of the form $< d, w >$, where d denoted a document and w a word.

Then, the probability was defined as

$$P(< d, w >) \triangleq \sum_{a \in A} P(d|a)P(w|a)P(a) \tag{4.1}$$

where a denoted a class of the occurrences and A the set of all classes.

The document clusters can specify the words that are likely to denote major topics in a set of similar documents.

Like in the approach presented in the previous chapter, 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 incorporated into the cache if they are not already in it.

Figure 4.1 shows the outline of the proposed model.

The main differences between the probabilistic thesaurus and the document clusters are the following.

The probabilistic thesaurus captures syntactic and semantic relations between correlated pairs of words, while the document clusters capture topic constraints, such as word

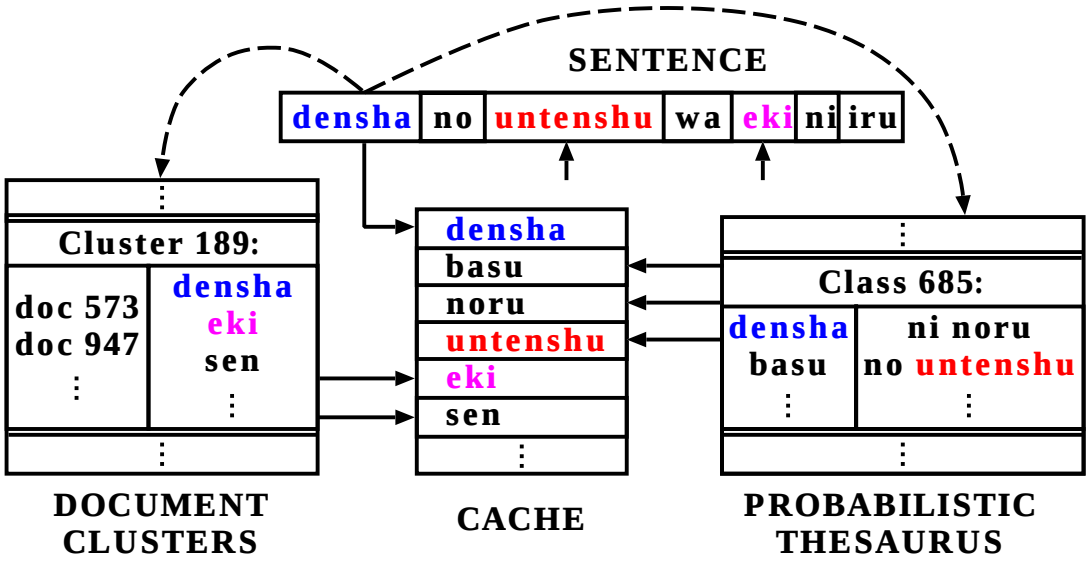

Figure 4.1: Outline of the further extension based on document clusters

choice and co-occurrence patterns.

The leading words set in the probabilistic thesaurus is associated to the related words set through a postposition, and the words in the former set are very likely to be found in the text followed by the corresponding words in the latter set. However, the related words in the document clusters are not syntactically related to each other, so they simply constitute a set of semantically related words, which may well belong to the same topic of discourse.

Consider as an example, that of figure 4.1. *densha* (train) and *basu* (bus) can be easily found in the text preceding *ni noru* (to get on) or *no untenshu* (driver), like in *densha no untenshu* (the train driver) or *basu ni noru* (to get on the bus). However, *eki* (station), although strongly associated to *densha* as well, is not usually followed by *ni noru* (to get on). Therefore, *eki* will probably not appear in class 685 of the probabilistic thesaurus.

Consequently, the two knowledge sources can be complementary to each other and provide different features to the LM.

4.2 Methodology

The rescoring algorithm proceeds as follows.

1. Read 1-best sentence
2. Add to the cache all words in the sentence
3. For each word in the sentence, add to the cache the most significant related words from the probabilistic thesaurus

4. For each word in the sentence, add to the cache the most significant related words from the document clusters, if they are not already in the cache

5. For each hypothesis in the N-best:

   - Calculate the score of the extended cache component
   - Calculate the score of the proposed LM
   - Calculate the total score (acoustic model + proposed LM)

6. Output the hypothesis with the highest total score

The scores in this further extension are calculated exactly in the same way as in the previous chapter.

4.3 Experimental Environment

The environment of the experiments that were conducted to test this model is almost the same as that of the previous chapter. Only the new parameters are presented here.

The document clusters were created from five years (1996-2000) of the Japanese Yomiuri Shimbun newspaper.

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 extended trigger model based on both 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.

The experimental environment is summarized in table 4.1.

4.4 Experimental Results

4.4.1 Final Model

The proposed extension was tested against the approach in the previous chapter and the model with only the cache component.

The average results of the experiments from the two sets, for a base cache size equal to 25, are shown in figure 4.2.

As it was also shown in the previous chapter, it can be observed that the extended trigger model based on the probabilistic thesaurus has a higher accuracy than both the baseline and the model with only the cache component. Furthermore, the extended trigger model based on both the probabilistic thesaurus and the document clusters has even a higher accuracy than the one based only on the probabilistic thesaurus. An improvement

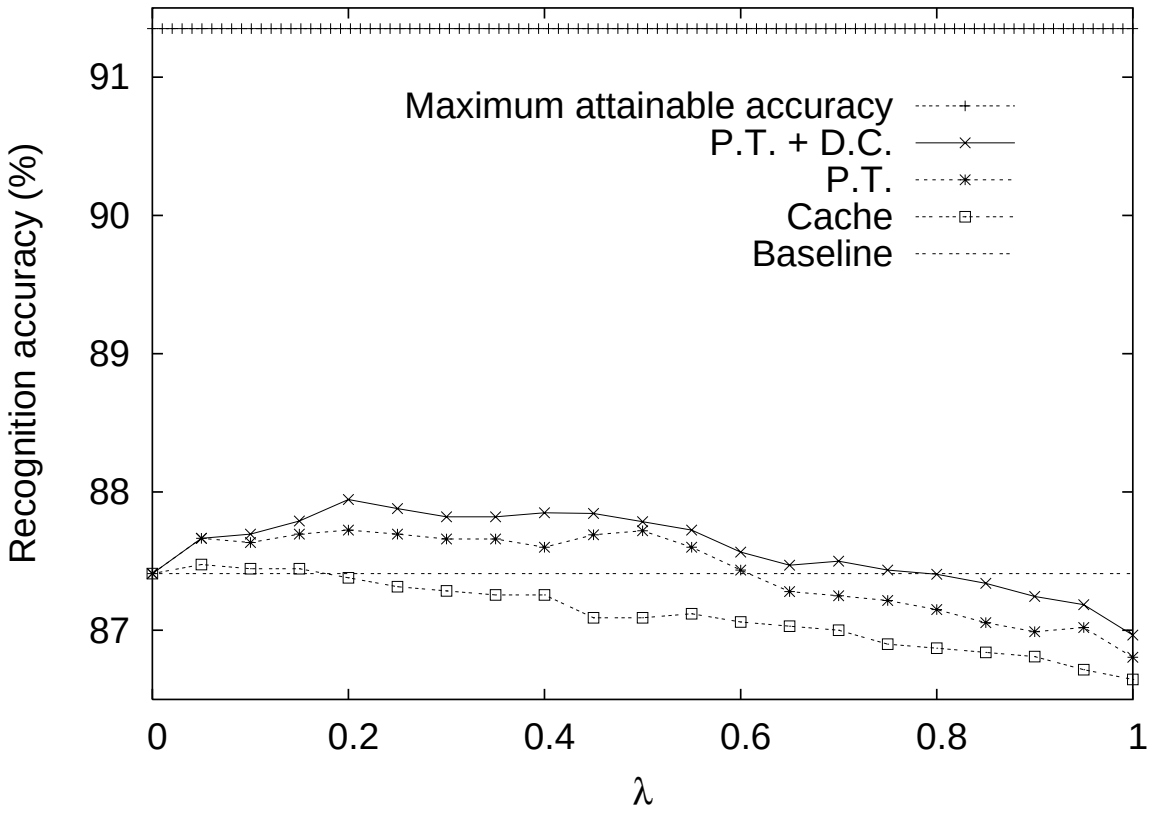

Figure 4.2: Speech recognition accuracy of the further extension based on document clusters, for different values of λ and a base cache size equal to 25

ASR system	Julius 3.1
Vocabulary size	21322 words
N (number of output hypotheses)	100
Test data	Source: Yomiuri Shimbun Number of sentences: 71 Topic: Education
Number of speakers	2 male speakers
Average baseline accuracy	87.41%
Average maximum attainable accuracy	91.35%
Number of significant classes (P.T.)	5
Number of significant leading words (P.T.)	5
Number of significant related words (P.T.)	5
Number of significant clusters (D.C.)	1
Number of significant related words (D.C.)	5
Base cache size	5, 10, 25, 50, 100, 250, 500
ε	10^{-30}

Table 4.1: Experimental environment for the further extension based on the document clusters

of 0.53% (absolute) over the baseline was obtained, which represents a 13.5% of the total possible improvement.

This additional improvement over the approach proposed in the previous chapter seems to prove that the additional related words extracted from the document clusters also contribute to improve the predictors in the LM. In the experimental results section this will be discussed again.

4.4.2 Cache Size

As it was previously commented, experiments with different cache sizes were performed. Specifically, sizes of 5, 10, 25, 50, 100, 250 and 500 were tried. Remember that these sizes are later multiplied by the number of related words that are incorporated into the cache for each word that enters the cache (51 in the probabilistic thesaurus case and 56 if it is the model based on both the thesaurus and the document clusters).

Figure 4.3 shows the maximum recognition accuracy of the extension based on the probabilistic thesaurus and of the further extension based on the document clusters, for the different sizes of the cache. In figure 4.4 the speech recognition accuracy of both models for a fixed λ equal to 0.2 for the different cache sizes is illustrated.

In both cases, the higher recognition accuracy was achieved for a base cache of size equal to 25. This is why the cache size in the final models were set to this value.

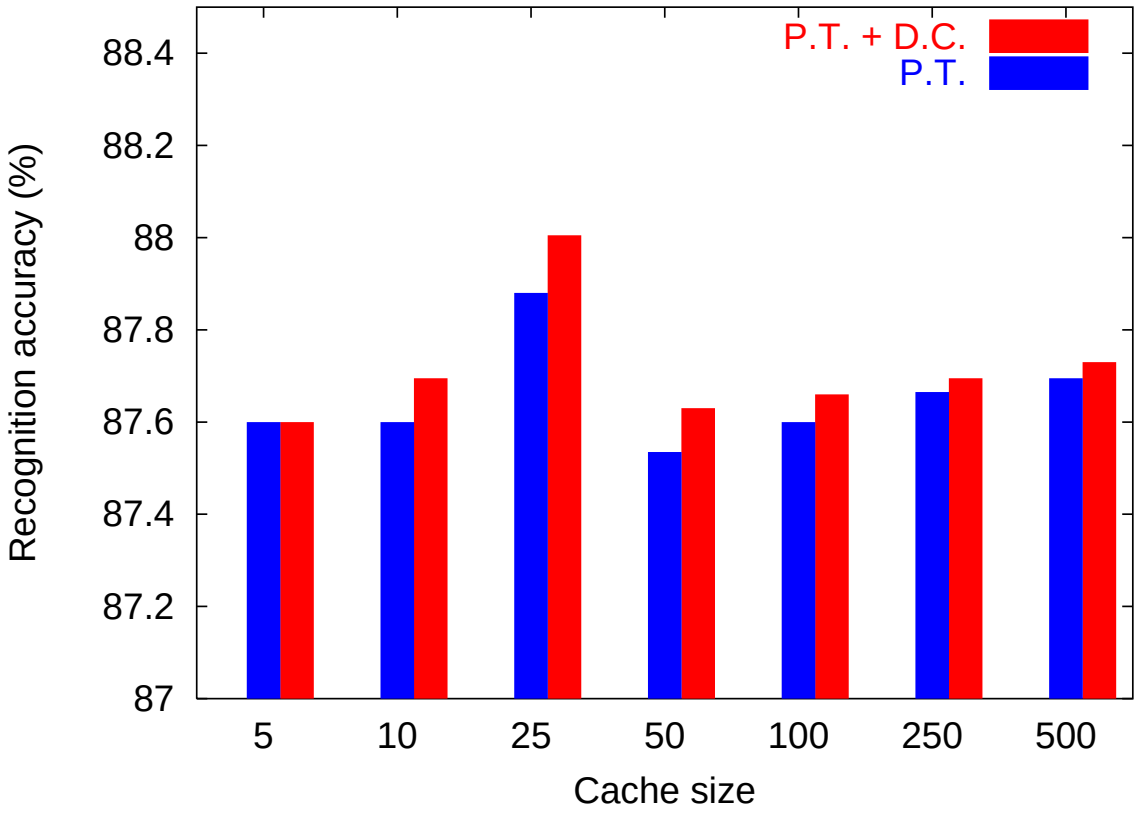

Figure 4.3: Maximum speech recognition accuracy of the two proposed extensions, for different values of the cache size

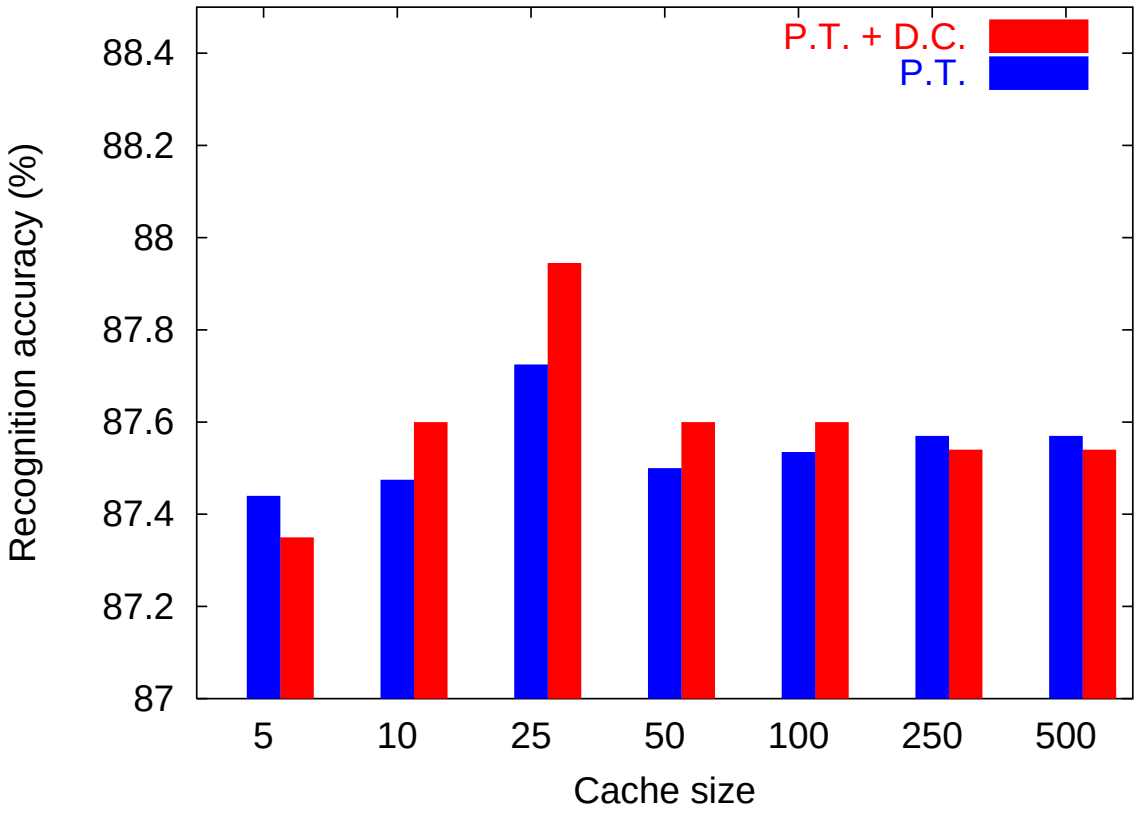

Figure 4.4: Speech recognition accuracy for λ equal to 0.2 of the two proposed extensions, for different values of the cache size

4.4.3 Extension Based Solely on Document Clusters

For the purpose of comparison, a model based solely on the document clusters was constructed. The model is analogous to the one based only on the probabilistic thesaurus, that is, the related words that are incorporated into the cache are only the ones found in the document clusters.

This time, the number of significant clusters was set to 1, and the number of related words extracted from each cluster was 20. Therefore, for each word that enters the cache, 20 related words are also added, and thus the base size of the cache is multiplied by 21.

The average results are illustrated in figure 4.5.

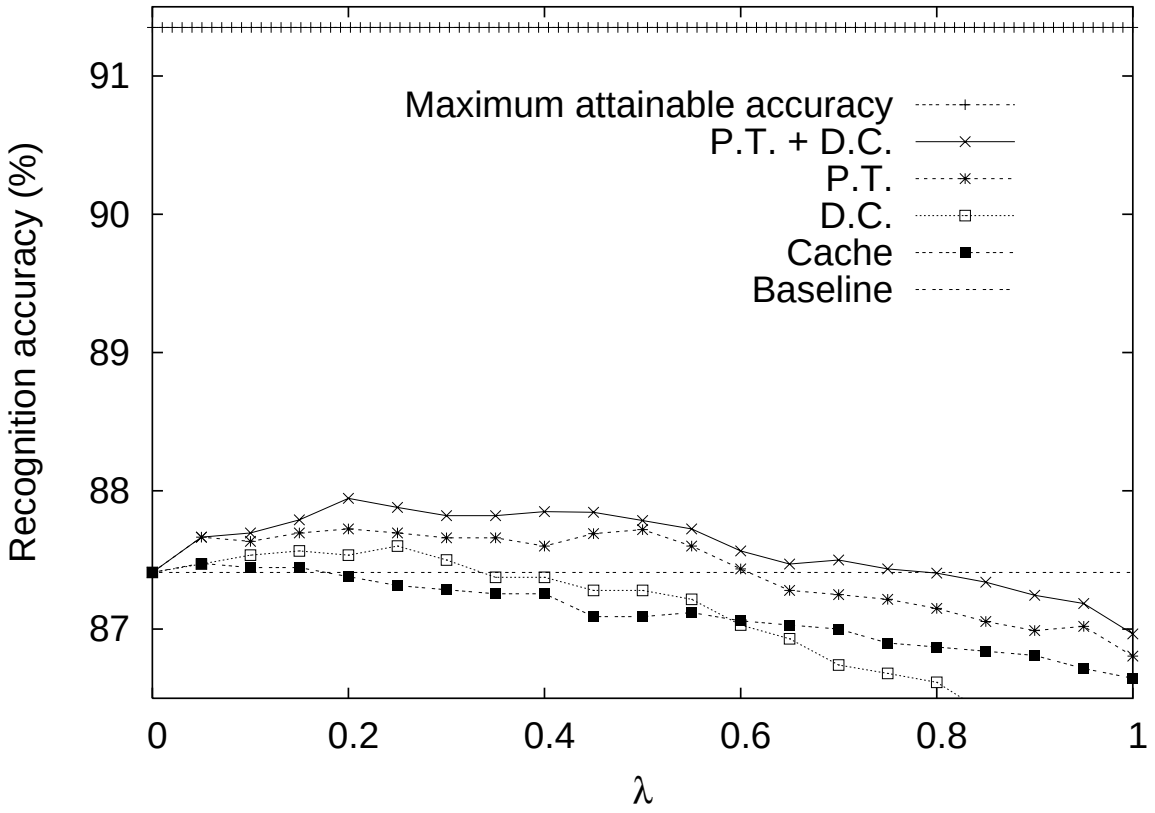

Figure 4.5: Speech recognition accuracy of the extension based solely on document clusters, for different values of λ and a base cache size equal to 25

As it can be seen, the model based on the document clusters alone performs also slightly better than the model with only the cache-based component, but worse than the model based on the probabilistic thesaurus and the one based on the two knowledge sources. This may mean that the document clusters have a less significative effect in the model than the probabilistic thesaurus.

In the next section, the usefulness of each of the model components is analyzed in more detail.

4.4.4 Assessing the Usefulness of the Related Words

In order to assess the usefulness of the words that are extracted from both the probabilistic thesaurus and the document clusters, some experiments were performed.

The idea is the following. Every word that is incorporated into the cache is looked up in both knowledge sources, as usual. However, when the most significant classes are found, instead of using them for extracting the most significant related words, different classes are used.

In these experiments, the previous class in number was used, except for class 0, in which case the last class was used (2499 for the probabilistic thesaurus and 299 for the document clusters). Of course, the previous class may also be related to the word being searched, by chance. Therefore, this is not a very formal experiment, in the sense that it cannot be stated that the different classes that are used are unrelated to the words that are being searched. Anyway, the results are sufficient to assess that the related words that are being used are indeed useful, that is, they contribute to the model by providing additional constraints such as semantic and syntactic dependencies between words and topic information.

Three experiments were carried out. In the first one, the classes in the probabilistic thesaurus were the ones made different. In the second one, the clusters in the document clusters were the object of modification. The third experiment changed both classes and clusters in the two knowledge sources.

The results can be seen in figure 4.6.

As it has been already mentioned, the experiment seems to confirm the hypothesis that the two knowledge sources employed in this research successfully contribute to improve the predictors in the baseline LM.

In the previous section, the hypothesis that the document clusters have a less significant effect in the model than the probabilistic thesaurus was formulated. According to the results shown in the figure above, it seems that this hypothesis is confirmed, because using erroneous clusters for extracting the related words does not affect the recognition accuracy so much as does the modification of the classes in the probabilistic thesaurus.

4.5 Results Analysis

In this section, the output of the model is compared with that of the baseline system, in order to see the actual source of the improvement.

Cases where the proposed model helped improve the correctness of sentences and cases where some sentences were replaced by less correct counterparts were found.

Here, some examples are presented to illustrate both cases.

The improvement was found to be due to two possible sources: the cache-based component alone and the related words extracted from the two knowledge sources used in this research.

Consider the following sentence from the evaluation data:

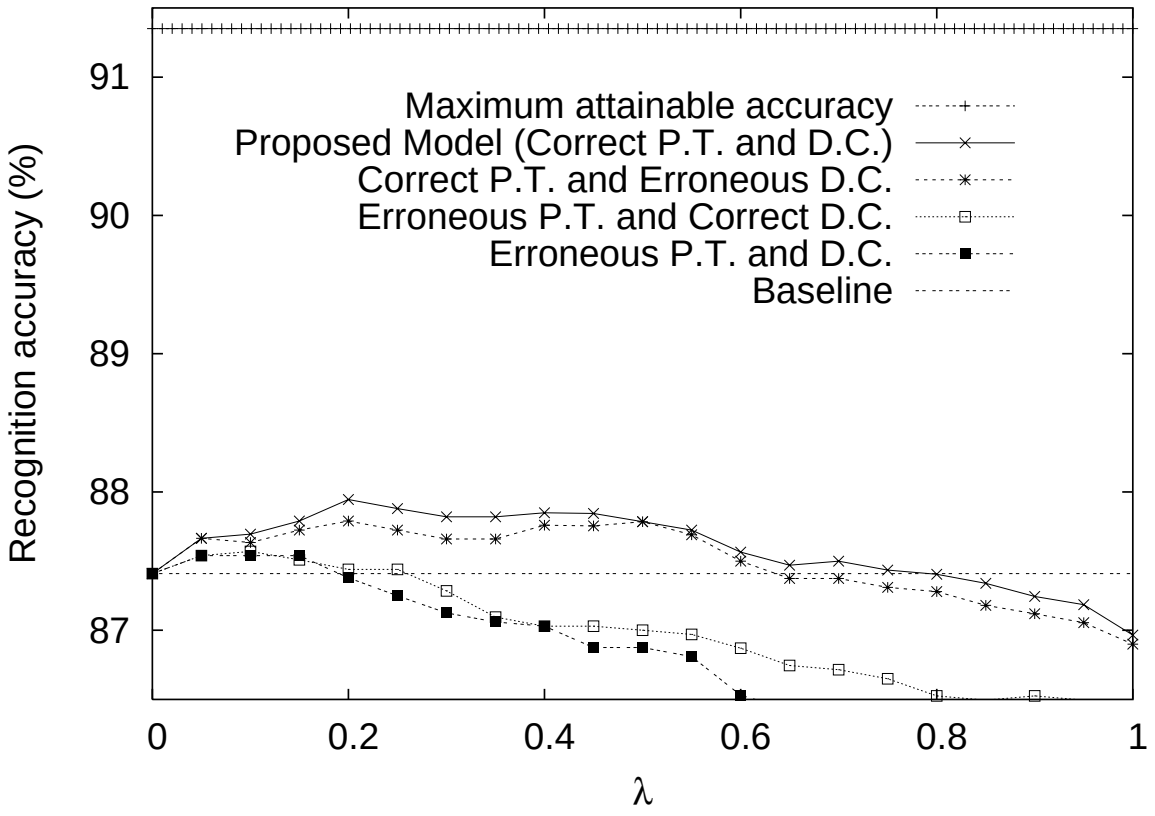

Figure 4.6: Speech recognition accuracy of the further extension based on document clusters with erroneous classes, for different values of λ and a base cache size equal to 25

その後、同省は高校教科書で指導要領を超えた記述を容認するなど転換の実質化を図ってきた。

The 1-best hypothesis output by the baseline model for test set 1 is the following:

その 後 同省 は **故郷 を** 教科書 で 指導 要領 を 超え た 記述 を容認 する など **返還** の 実質 化 を **はかっ** て き た

where the erroneous words are in boldface.

The output of the proposed model for the same sentence is as follows:

その 後 同省 は **故郷 を** 教科書 で 指導 要領 を 超え た 記述 を容認 する など **を 変換** の 実質 化 を 図っ て き た

As it can be seen, the word "図る" (*hakaru*: to plan) was correctly replaced in the sentence above. When the origin of this replacement was investigated, the fact that the suffix "化" (*ka*: similar to "-zation") induced the addition to the cache of the word 図る from the probabilistic thesaurus was discovered.

A similar example is found in the following sentence:

四年前の二・二倍で、年々増加している。

The 1-best hypothesis and the output of the proposed model are showed below.

四 年 前 の 二 **点 に 描い て** 年々 増加 し て いる

四 年 前 の 二 **点 に** 倍 で 年々 増加 し て いる

The word "倍" (*bai*: times) was successfully incorporated with the proposed model. This word also was extracted from the probabilistic thesaurus and inserted in the cache when the word "昨年" (*sakunen*: last year) from the previous sentence was looked up (see Appendix B).

An example where the cache component alone was sufficient for improving the accuracy is in the following sentence:

国立大の改革を促すスピードは急速に早まっており、六月には「民間的経営手法の導入」、「国立大の再編・統合」などを盛り込んだ文部科学省の「大学の構造改革の方針」が打ち出され、波紋を呼んだ。

The corresponding 1-best sentence and proposed model output are

国立 大 の **開花 コーナー が** スピード は 急速 に **ハイ 余っ**て おり 六月 に は 民間 **定期 型** 手法 の 導入 国立 大 の 再編統合 など を 盛り込ん だ 文部 **各省** の 大学 の 構造 改革 の 方針が 打ち出さ れ 波紋 を 呼ん だ

国立 大 の 改革 **大 な が** スピード は 急速 に **ハイ 余っ** ており 六月 に は 民間 **定期 型** 手法 の 導入 国立 大 の 再編 統合など を 盛り込ん だ 文部 **各省** の 大学 の 構造 改革 の 方針 が 打ち出さ れ 波紋 を 呼ん だ

In this case, the word "改革" (*kaikaku*: reform) appears in the 1-best hypothesis and, therefore, it is incorporated to the cache by the cache component. In this way, its probability is raised and it is correctly recognized by the proposed model.

In contrast to these successful examples, an example where the proposed model performed worse than the baseline is presented below.

母親一人が育児を担っており、周囲の協力がなく、社会の子育て支援システムが不十分だからが専門家の一致した見方だ。

The corresponding pair of sentences from the outputs of the baseline and the proposed model are, respectively,

母親 一 人 が 育児 を 担っ て おり 周囲 の 協力 が なく 社会 の 子育て**信 氏 船** が 不十分 だ から が 専門 家 の 一致 し た 見方 だ

反 応援 一 人 が 育児 を 担っ て おり 周囲 の 協力 が なく 社会の 子育て **信 氏 船** が 不十分 だ から が 専門 家 の 一致 し た 見方 だ

In this case, the word "協力" (*kyouryoku*: cooperation) induced the addition to the cache of the word "応援" (*ouen*: help, aid). Therefore, the probability of the sentence above was increased and became the output of the proposed model.

4.6 Summary

An additional extension to the model described in the previous chapter has been proposed in this one. The document clusters have been described first, followed by the explanation of the methodology employed in this extension. Finally, several experiments have been discussed, and their results illustrated.

The conclusions and directions for future works are presented in the next chapter.

Chapter 5

Conclusions and Future Works

5.1 Conclusions

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 first source captures syntactic as well as semantic dependencies between words in the text, while the latter provides information about the current topic of discourse.

An overall absolute improvement of 0.53% in speech recognition accuracy over the baseline was achieved, which represents a 13.5% of the total possible improvement that could be attained if the model found the best hypotheses from the N-best every time.

The possible reasons for this small degree of improvement are the following:

- The hit rate for the related words, that is, the number of words in the current sentence that can be found within the related words divided by the number of words in the sentence, is small (17.7%).
- The 1-best sentence, from where the words that are inserted in the cache and looked up in the two knowledge sources are extracted, sometimes has errors.
- The scores are unnormalized and apply to sentences instead of to words.
- Homophones (i.e. the same words written with different Japanese character sets) and word separation (i.e. compound words with their components written together or separately) can be found in the N-best list, and they are considered different by the program that computes the recognition accuracy.

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.

5.2 Future Works

Instead of using unnormalized scores for the parameters of the model, I plan to use normalized probabilities, and incorporate into the model the information of the probability distributions of both the probabilistic thesaurus and the document clusters.

In addition, I want to use a threshold for the significance of words and classes, so that all words with a probability that is over the significant word threshold, and all the classes with a probability greater than the significant class threshold are considered significant. Then, all the related words can be ordered based on their likelihoods and the M most likely related words would be the ones used for the cache.

The proposed model currently calculates the scores of its components treating the sentence as a unit. That is, for each sentence, its extended cache score and its baseline score are interpolated to form the score of the proposed LM. Alternatively, I want to use the word as a unit, therefore, for each word, its extended cache probability and its baseline probability will be interpolated to form the overall LM probability. This probability can then be used to calculate the perplexity of the model.

It is also my intention to construct a set of Japanese trigger pairs based on the average mutual information measure, and a standard trigger model based on the maximum entropy framework, in order to perform a fair comparison between the two models.

Most of the ASR systems based on adaptive LMs perform the recognition using a standard bigram or trigram LM, and then the output N-best hypotheses are rescored based on the new LM probabilities. Thus, the accuracy of the system output is subject to the reliability of the N-best hypotheses. I also want to incorporate the proposed LM into the ASR decoder in order to take advantage of its features before generating the N-best hypotheses.

Appendix A

Example of Classes in Probabilistic Thesaurus and Document Clusters

Class 121			
Leading word	Probability	Postposition:Related word	Probability
野球 (baseball)	2.051230e-01	を:する (to practice)	2.977070e-02
サッカー (soccer)	1.484840e-01	の:大会 (tournament)	2.959080e-02
ゴルフ (golf)	1.107540e-01	の:選手権 (champion)	2.548160e-02
テニス (tennis)	6.380350e-02	を:やる (to practice)	2.162680e-02
ラグビー (rugby)	5.198480e-02	を:始める (to begin)	2.094190e-02

Class 1505			
Leading word	Probability	Postposition:Related word	Probability
コンピューター (computer)	5.418790e-01	を:使う (to use)	8.794030e-02
パソコン (personal computer)	1.313480e-01	に:よる (by means of)	6.597580e-02
ホストコンピューター (host computer)	3.430110e-02	を:利用:する (to use)	2.223500e-02
ロボット (robot)	2.738840e-02	に:入力:する (to input)	1.735560e-02
コンピューターグラフィックス (computer graphics)	2.080980e-02	で:管理:する (to manage)	1.686100e-02

Class 2451			
Leading word	Probability	Postposition:Related word	Probability
大学 (university)	8.501500e-01	の:教授 (professor)	5.221000e-02
短大 (junior college)	3.983830e-02	:大学 (university)	1.695420e-02
大学院 (graduate school)	3.615060e-02	を:卒業:する (to graduate)	1.600620e-02
学部 (department)	1.045720e-02	の:ら (ungrammatical!)	1.448950e-02
大学校 (college)	6.064700e-03	の:者 (person)	1.425850e-02

Table A.1: Examples of classes from the probabilistic thesaurus

Cluster 181	
Related word	Probability
大会 (tournament)	6.192060e-02
選手 (athlete)	5.735580e-02
する (to practice)	4.798500e-02
チーム (team)	4.705690e-02
出場 (participation)	4.567510e-02

Cluster 60	
Related word	Probability
する (to do)	5.971700e-02
インターネット (internet)	5.513400e-02
パソコン (personal computer)	4.575600e-02
コンピューター (computer)	3.225630e-02
できる (to be able)	2.703840e-02

Cluster 112	
Related word	Probability
大学 (university)	2.087000e-01
する (to do)	5.160510e-02
学部 (department)	4.072550e-02
英語 (English)	3.849120e-02
資格 (qualifications)	3.103240e-02

Table A.2: Examples of clusters from the document clusters

Appendix B

Evaluation Data

様々な面で教育が問われ続けた一年だった。
深刻な議論も、衝撃的な事件もあった。
あの問題、あの事件は何を問い掛けたのか。
そしてそれは今、どうなっているのか。
今年の教育を振り返りつつ、来年の課題を探った。
学力問題に明け暮れた一年だった。
本紙は一月五日朝刊で、文部省が「ゆとり教育」を抜本的に見直す方針であることを特報した。
同省の方向転換は学校現場に大きな衝撃を与え、「一・五読売報道ショック」との言葉も生まれた。
同月二十四日、同省の小野元之次官は都道府県教育長協議会で「ゆとりはゆるみではない」と発言。
学校で基礎的な学習がおろそかにされている傾向もあると注意を促した。
指導要領は最低基準で、指導要領の範囲を超えた授業もできるとした。
その後、同省は高校教科書で指導要領を超えた記述を容認するなど転換の実質化を図ってきた。
東京都が都立高校四校を「進学指導重点校」に指定するなど、学力向上に向けた自治体の動きも目立った。
この転換は、子どもの学力が低下しているとの指摘が各方面から上がったことを受けたものだ。
来年四月から小中学校では、教科内容や授業時間を大幅に削減した新学習指導要領が実施されるが、その影響への懸念もあった。
こうした見直しに、「知識の詰めこみに戻るのか」という反発もある。
しかし、様々な調査や論議で分かってきたのは、日本の子どもたちの学習意欲の乏しさや家庭での学習時間の少なさだった。
「考える力」と基礎学力の関係も論議になった。
受験のための偏差値序列でしかとらえられてこなかった学力の内実や水準が、初めて本格的に論議された一年だった。
来年は新指導要領の効果が厳正に評価されることになりそうだ。
「開放」との両立探る。
学校開放が進むなか、六月八日に大阪教育大付属池田小で児童殺傷事件が起き、学校の安

全確保が問題になった。
東京都は、緊急通報システム「学校１１０番」を、今年度末までに公私立の小中学校や幼稚園など約五千七百施設に設置する。
校内の非常ボタンを押すと、警視庁通信指令本部に自動通報され、警官が急行する。
地域が子どもを守る動きも広がっている。
事件後から川崎市内の市立小学校百十四校では、ＰＴＡや住民らによる防犯パトロールが続く。
市教委は当初、毎日一人八百円の謝礼を支払う形を取った。
「子どものためという気持ちに水を差す」との声が参加者から上がり、今パトロールは保護者らの自主性に任せる形だ。
同市幸区の夢見ヶ崎小学校では、授業や課外活動を支援する教育ボランティアらが協力。
渡辺則雄さんは手が空いた時間に、校内の様子がよく見えるよう校門や校庭の周囲の植え込みを刈り、携帯型の防犯ブザーを持って見回る。
「地域の人たちが学校に関心を持てば、児童の安全は保てると思う」と言う。
「地域交流棟」がある新潟県聖籠町立聖籠中学校では、教職員と生徒は名札、来校者は「お客様」プレートの着用を徹底。
坂口真生校長は「学校を閉ざすのではなく、地域住民が多く学校に出入りすることで不審者を排除したい」と話す。
「開かれた学校」と安全確保の両立が課題となっている。
私立、ＡＯ入試で学生確保図る。
国立はセンター試験科目増やす。
私立大では、入学者が定員に満たない定員割れの大学が今春、全体の約三割を超えた。
書類審査や面接で選抜するＡＯ入試が急速な広がりを見せているが、受験生の「青田買い」に利用しているのでは、との批判も出てきている。
一方、大学入試センター試験を実施する国立大学の八割にあたる七十五校は、二〇〇四年度入試から、同試験で「五教科七科目」の受験を義務づける方針を打ち出した。
受験生確保策もあって、国立大でも入試科目を減らす傾向が続いており、受験生が入試に必要な科目しか勉強しない弊害や、新入生の学力が低下していることへの懸念が強まっていた。
すでに、「生徒には安易にＡＯで進学先を決めないように促している」という高校も増えている。
新入生の「学力水準」確保のため大学入試の科目を増やす傾向は、私立大にも影響を与えそうだ。
今年は国立の山形大をはじめ、過去の入試ミスの発覚が相次いだ年でもあった。
ミスを知りながら隠ぺいしていた富山大のケースは悪質だった。
私立大も含め、大学の組織の在り方が改めて問われた。
国立大の改革を促すスピードは急速に早まっており、六月には「民間的経営手法の導入」、「国立大の再編・統合」などを盛り込んだ文部科学省の「大学の構造改革の方針」が打ち出され、波紋を呼んだ。
実施のための具体策が問われる。
昨年十一月、児童虐待防止法が施行された。
しかし、その後も、虐待事件は後を絶たない。
今年七月名古屋市で、三十二歳の母親が七歳の長女を虐待して死なせ逮捕された。

似た事件が頻発している。
児童相談所が「家族からの虐待」を理由に児童福祉施設に保護した子どもは、昨年度は二千五百二十七人。
四年前の二・二倍で、年々増加している。
犯罪に至らなくても、「幼い我が子をたたいてしまった」、「暴言を浴びせてしまった」などと、自分を責めて悩み、各種の相談施設に訴える母親が増えている。
虐待と日常の小さな暴力との境目はなく、虐待事件の“ 予備軍 ”は無数に存在するといわれる。
背景には、強まる母親の育児ストレスがあるとされる。
子育てに疲れ果て、「子どもと二人だけでいたくない」と訴える人は少なくない。
育児文化研究所の丹羽洋子所長は「母親たちの閉そく感はますます強まっている」と話す。
母親一人が育児を担っており、周囲の協力がなく、社会の子育て支援システムが不十分だからが専門家の一致した見方だ。
虐待事件は、低下した家庭や地域の教育力や育児力をどのように高めていくかを突きつけている。
働く女性が増え、多様になった働き方に合わせた社会のシステムがしっかり構築されない限り、また、母親たちの「孤独な育児」に多くの人の理解が進まない限り、今後も悲惨な事件はますます増えると見られる。
国の調査結果などによると、不登校や引きこもりをめぐる状況は深刻さを増している。
一方、国がこれらを社会問題として認識し、対策に重い腰を上げたことを評価する声もある。
文部科学省が八月に発表した調査によると、昨年度の不登校児童・生徒は小中学生は十三万四千人と過去最多。
特に中学ではクラスに一人は不登校の生徒がいる計算だ。
こうした状況を「高度経済成長期以来の長期的な社会現象」とするのは、社会的ひきこもりなどの著書のある精神科医の斎藤環さんだ。
「働かなければ生きていけないという実感がなく、なぜ働くのかを問い始めた」引きこもり世代の意識の変化を見て取り、今の日本に成熟のモデルがないことを指摘する。
「本当の意味でのかっこいい大人がいない、早く大人になりたいと思わせる要素がないため、若者の意識が成熟拒否へと向かってしまう」。
教育の現場での具体的な対策として、「朝の十分間読書」運動の広がりや、ディベートなどコミュニケーション技術を伸ばす取り組みに期待するという。
また「社会に世間体に変わる価値基準が出てきてほしい」とも話す。
不登校も引きこもりも、偏見の目で見るのでなく、社会全体の問題として家族や学校が認識することが必要だ。

Bibliography

[1] Lalit Bahl, Frederick Jelinek, Robert L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, volume PAMI–5, number 2, pages 179–190, 1983.

[2] Brigitte Bigi, Armelle Brun, Jean-Paul Haton, Kamel Smaïli, Imed Zitoumi, "A Comparative Study of Topic Identification on Newspaper and E-mail," Proceedings of the 8th International Symposium on String Processing and Information Retrieval, 2001.

[3] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jeniffer C. Lai, Robert L. Mercer, "Class-Based n-gram Models of Natural Language," Computational Linguistics, volume 18, number 4, pages 467–479, 1992.

[4] Stanley F. Chen, Kristie Seymore, Ronald Rosenfeld, "Topic Adaptation for Language Modeling Using Unnormalized Exponential Models," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 681–684, 1998.

[5] Philip R. Clarkson, Anthony J. Robinson, "Language Model Adaptation using Mixtures and an Exponentially Decaying Cache," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume II, pages 799–802, 1997.

[6] Philip R. Clarkson, Ronald Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit," Proceedings of the European Conference on Speech Communication and Technology, 1997.

[7] Philip R. Clarkson, "The Applicability of Adaptive Language Modelling for the Broadcast News Task," Proceedings of the International Conference on Spoken Language Processing, volume 5, pages 1699–1702, 1998.

[8] Philip R. Clarkson, "Adaptation of Statistical Language Models for Automatic Speech Recognition," Ph.D. Thesis, 1999.

[9] Stephen Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, Salim Roukos, "Adaptive Language Modeling Using Minimum Discriminant Estimation," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume I, pages 633–636, 1992.

[10] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, volume 39, number 1, pages 1–38, 1977.

[11] Marco Ferreti, Giulio Maltese, Stefano Scarci, "Language Model and Acoustic Model Information in Probabilistic Speech Recognition," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1989.

[12] Radu Florian, David Yarowsky, "Dynamic Nonlocal Language Modeling via Hierarchical Topic-Based Adaptation," Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 167–174, 1999.

[13] Thomas Hofmann, Jan Puzicha, "Unsupervised Learning from Dyadic Data," Technical Report TR–98–042, 1998.

[14] Xuedong Huang, Fileno Alleva, Hsiao-wuen Hon, Mei-Yuh Hwang, Kai-Fu Lee, Ronald Rosenfeld, "The SPHINX-II Speech Recognition System: An Overview," Computer Speech and Language, volume 2, pages 137–148, 1993.

[15] Rukmini Iyer, Mari Ostendorf, "Modeling Long Distance Dependence in Language: Topic Mixtures vs. Dynamic Cache Models," IEEE Transactions on Speech and Audio Processing, volume 7, number 1, pages 30–39, 1999.

[16] Edwin T. Jaynes, "Information Theory and Statistical Mechanics," Physical Review, volume 106, pages 620–630, 1957.

[17] Frederick Jelinek, "Self-Organized Language Modeling for Speech Recognition," Readings in Speech Recognition, pages 450–506, 1990.

[18] Frederick Jelinek, Bernard Merialdo, Salim Roukos, M. Strauss, "A Dynamic Language Model for Speech Recognition," Proceedings of the DARPA Workshop on Speech and Natural Language, pages 293–295, 1991.

[19] Information-Technology Promotion Agency, Kyoto University, Nara Institute of Science and Technology, "Multipurpose Large Vocabulary Continuous Speech Recognition Engine Julius rev. 3.2," http://julius.sourceforge.jp/3.3/Julius-3.2-book-e.pdf

[20] Slava M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," IEEE Transactions on Acoustics, Speech and Signal Processing, volume ASSP–35, number 3, pages 400–401, 1987.

[21] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, K. Shikano, "Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition," Proceedings of the International Conference on Spoken Language Processing, volume 4, pages 476–479, 2000.

[22] Sanjeev Khudanpur, Jun Wu, "A Maximum Entropy Language Model Integrating N-Grams and Topic Dependencies for Conversational Speech Recognition," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume I, pages 553–556, 1999.

[23] Sanjeev Khudanpur, Jun Wu, "Maximum Entropy Techniques for Exploiting Syntactic, Semantic and Collocational Dependencies in Language Modeling," Computer Speech and Language, volume 14, number 4, pages 355–372, 2000.

[24] Roland Kuhn, "Speech Recognition and the Frequency of Recently Used Words: A Modified Markov Model for Natural Language," Proceedings of the International Conference on Computational Linguistics, pages 348–350, 1988.

[25] Roland Kuhn, Renato De Mori, "A Cache-Based Natural Language Model for Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, volume PAMI–12, number 6, pages 570–583, 1990.

[26] Roland Kuhn, Renato De Mori, "Corrections to A Cache-Based Natural Language Model for Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, volume PAMI–14, number 6, pages 691–692, 1992.

[27] Sadao Kurohashi, Manabu Ori, "Nonlocal Language Modeling Based on Context Co-occurrence Vectors," Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 80–86, 2000.

[28] Raymond Lau, Ronanld Rosenfeld, Salim Roukos, "Trigger-Based Language Models: A Maximum Entropy Approach," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume II, pages 45–48, 1993.

[29] Paul Placeway, Richard Schwartz, Pascale Fung, Long Nguyen, "The Estimation of Powerful Language Models from Small and Large Corpora," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume II, pages 33–36, 1993.

[30] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition," Prentice Hall, 1993.

[31] Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, Franz Beil, "Inducing a Semantically Annotated Lexicon via EM-Based Clustering," Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 104–111, 1999.

[32] Ronald Rosenfeld, "Adaptive Statistical Language Modeling: A Maximum Entropy Approach," Ph.D. Thesis CMU–CS–94–138, 1994.

[33] Ronald Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," Computer Speech and Language, volume 10, pages 187–228, 1996.

[34] Richard Schwartz, Yen-Lu Chow, "The N-best Algorithm: Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 81–84, 1990.

[35] Satoshi Sekine, John Sterling, Ralph Grishman, "NYU/BBN 1994 CSR Evaluation," Proceedings of the Spoken Language Systems Technology Workshop, pages 148–152, 1995.

[36] Satoshi Sekine, "Modeling Topic Coherence for Speech Recognition," Proceedings of the International Conference on Computational Linguistics, pages 913–918, 1996.

[37] Satoshi Sekine, Ralph Grishman, "NYU Language Modeling Experiments for the 1995 CSR Evaluation," Proceedings of the Spoken Language Systems Technology Workshop, pp 123–128, 1996.

[38] Satoshi Sekine, Andrew Borthwick, Ralph Grishman, "NYU Language Modeling Experiments for the 1996 CSR Evaluation," Proceedings of the DARPA Speech Recognition Workshop, 1997.

[39] Kristie Seymore, Ronald Rosenfeld, "Using Story Topics for Language Model Adaptation," Proceedings of the European Conference on Speech Communication and Technology, 1997.

[40] Kristie Seymore, Ronald Rosenfeld, "Large-Scale Topic Detection and Language Model Adaptation," Technical Report CMU–CS–97–152, 1997.

[41] Kristie Seymore, Stanley F. Chen, Ronald Rosenfeld, "Nonlinear Interpolation of Topic Models for Language Model Adaptation," Proceedings of the International Conference on Spoken Language Processing, volume 6, pages 2503–2506, 1998.

[42] Kentaro Torisawa, "A Nearly Unsupervised Learning Method for Automatic Paraphrasing of Japanese Noun Phrases," Proceedings of the Workshop on Automatic Paraphrasing: Theories and Applications, pages 63–72, 2001.

[43] Kentaro Torisawa, "An Unsupervised Method for Canonicalization of Japanese Postpositions," Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, pages 211–218, 2001.

[44] Carlos Troncoso, Shigeki Matsuda, Mitsuru Nakai, Hiroshi Shimodaira, Kentaro Torisawa, "An Extension to the Trigger Language Model Based on a Probabilistic Thesaurus," Proceedings of the Spring Meeting of the Acoustical Society of Japan, volume I, pages 141–142, 2003.

[45] Jun Wu, Sanjeev Khudanpur, "Combining Nonlocal, Syntactic and N-Gram Dependencies in Language Modeling," Proceedings of the European Conference on Speech Communication and Technology, volume 5, pages 2179–2182, 1999.

[46] Jun Wu, Sanjeev Khudanpur, "Syntactic Heads in Statistical Language Modeling," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 1699–1702, 2000.

[47] Jun Wu, Sanjeev Khudanpur, "Building a Topic-Dependent Maximum Entropy Model for Very Large Corpora," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2002.