Speech Recognition across Multiple Regions

DT2118: André Algotsson, 870319-1638

Abstract

This paper investigates two state-of-the-art language models (LM) for Automatic Speech Recognition (ASR). The first model uses *geographical metadata* in order to create “regional models”, which are then used to tune the vocabulary distributions of the LM. The second model uses *Substate Gaussian Mixture Models* for phoneme-based recognition, and shows that it can quickly be adapted to work well on a language where resources are sparse. The techniques are introduced and the results are discussed.

Introduction

This paper investigates two state-of-the-art language models (LM) for Automatic Speech Recognition (ASR).

Creating good language models is of vital importance when it comes to recognize speech at an advanced level. The ultimate goal is to create a full-featured ASR that can understand any standard speech, regardless of speaker and topic.

Both of the investigated models attempt to find ways to extend previous models, so that they can more easily cope with the large variety of speakers and situations that exist in the real world.

The first model uses geographical data in order to make the speech recognizer regionally aware. Both dialectal differences and city-specific vocabulary etc. are modelled. By doing this, the ASR gets more prior information about words that are likely to be spoken, which in some cases can enhance the overall performance.

The second model manages to capture features that can be reused on a different set of languages. This is achieved using *Substate Gaussian Mixture Models* – an enhanced version of the standard GMM that permits a more flexible use of the mixtures.

Both of the techniques are in their infancy (as of 2011), but have already shown some interesting improvements over other methods.

Investigation

Model 1: Local LM using Geographical Metadata

Experimental Setup

With the advent of modern cellphone technology, it has become increasingly easy to retrieve detailed geographical data from phone calls. Knowing the location of a speaker enables the speech recognition system to fine-tune its parameters and thus adapt to the situation.

This has been explored in [1] in the area of business calls. It starts by assuming that the location of a speaker will be correlated to what company names that are likely to be mentioned. For instance, knowing that a speaker is calling from a certain city X makes it reasonable to assume that he or she will be likely to refer to a business that also is located in X.

The model was given information about various businesses and their locations across the USA, together with their corresponding states.

Voice data was then collected anonymously from business query phone calls. Dialectal variance among speakers was modelled by partitioning the voice data into six different regions.

Since the voice data was not extensive enough to capture the full span of possible queries, the local language model was trained with an additional set of 100M typed web queries.

The Model

The local language model constists of a mixture of pdf's which span over regions of different size. Together they form a weighted sum that is used to determine the probability of a given word: p(x|locality), see *Figure 1*.

By using mixtures in this way, the whole country can be covered even when data from some specific regions is sparse.

$$P(x|location) = w_l p_l(x) + w_{l1} p_{l1}(x) + w_{l2} p_{l2}(x) + w_u p_u(x)$$

Figure 1 – Visualizing the mixture of pdf's, each mixture capturing a certain region of the country.

The *local* language model is retrieved by interpolating the transcribed voice data with 3-grams that have been extracted from the 100M web queries.

The recognizer is then trained using HMM. A variety of different methods was used for interpreting the regional data, in order to get a good base for comparison.

The results show a 2.2 % absolute improvement, or 7 % relative error reduction in comparison to the global model.

Further details about the model and the results can be found in [1].

Discussion

This experiment has shown that it is possible to get improvements in recognition performance using information about locality. By providing regionalized information on what words are likely to be spoken and how the speakers are likely to sound, the Word Accuracy was visibly enhanced.

The idea of using this kind of information can prove to be a useful tool to be aware of when designing future applications.

However, designing language models that thrive on locality will often cause a loss of generality. The ultimate goal for any system would naturally be to capture any word, regardless of the context where it was spoken. These kinds of specialized local models will naturally be of little use when there are little regional differences between the utterances (e.g. querying a nation-wide business).

In the cases where the method is applicable, a good local language model will enable the ASR to make more educated guesses about the speakers, and to predict what words are likely to show up.

The experiments that were carried out in [1] were using very specialized data (i.e. company names). Further areas of research could include finding more types of regional features to learn. A more powerful regional model could contain deeper knowledge about how humans differ across regions. That kind of knowledge could perhaps be retrieved from social networks and blogs, which contain a wealth of data describing human behaviour.

It is possible that there will be a stronger connection between language models and broader human behavioural models in the future, since speech cannot completely be separated from its speaker.

Model 2: SGMM-based Multilingual Acoustic Modelling

Introduction to Multilingual Modelling and SGMM

When dealing with *multiple languages*, one has to model the phonetic differences between the languages, and use it during training.

The naïve approach would be to model and train each language separately, but this becomes infeasible for obscure languages where digital training materials are sparse and time is limited. This especially relates to intelligence agencies, whose attention quickly can shift towards a completely unexplored area of the world. [2]

Instead of assuming a "universal phone set" that is shared among all languages of the world, a new approach has been suggested which is an extension of the normal Gaussian Mixture Models (GMM).

The main drawback of GMMs in this application is that unrelated Gaussian mixtures tend to overlap, making it hard to model the phones satisfactorily. Instead, a new approach called *Substate Gaussian Mixture Models* (SGMM) has been proposed, which effectively "turns off" Gaussians at irrelevant locations, see *Figure 2*. [3]

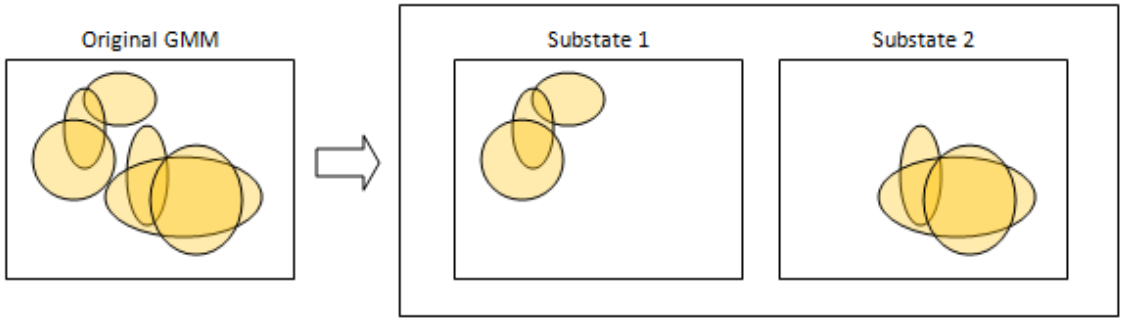

Figure 2 – A simple Gaussian Mixture model, split up into two subspaces. The mixtures in Substate 1 now have a limited influence on the mixtures in Substate 2 and vice-versa.

This means that every significant phone cluster will have its own dedicated set of Gaussians. These Gaussians define a *subspace* for each phone cluster. A given state is then defined as a *mixture of the substate distributions.*

Note that there are two levels of mixtures in the SGMM: One level of Gaussian distributions, and one level of substate distributions (which contain the Gaussians).

The parameters of the SGMM are trained using EM. Further theory can be found in [2] and [3].

Performance of SGMM

It has been shown that an SGMM that has been thoroughly trained on one or more languages easily can be adapted to another language where training data is very limited. The authors of [2] report an experiment where an English speech recognition system was trained using only 1 hr of English data. As would be expected, the Word Error Rate (WER) was very poor (~68 %) using standard approaches. However, by first training the SGMM on a Spanish and a German corpora, and then adapt it to the 1 hour English data, a very large WER reduction of 10.9 % could be noted, see *Table 1*.

Table 1 – Comparison between English recognition performance for a very limited data set

Model	Training Data	WER [%]
Conventional models	1 hr English	70.5
SGMM	1 hr English	67.6
SGMM	Spanish+German, and 1 hr English	59.6

Discussion

The performance of the SGMM does look very promising, as it suggests that it is possible to identify *general features* that are shared between many languages. In the ideal case, one may be able to train the model extensively on a bigger set of major languages where resources are easy to find (e.g. English, German, Japanese etc.), and achieve a “universal language” model. This unified model could then possibly be adapted to smaller languages using only a minimum of training. However, it should be remembered that the results presented in [2] are still at a very modest Word Error Rate of 58 %. Future research will hopefully help to improve the performance further, and it will be interesting to see if the cross-lingual effect still will be visible as the WER goes down.

This experiment found some similarities (in the SGMM sense) between Spanish, German and English. Another useful area for further investigation would be to find other “language clusters” that tend to match each other.

Conclusions

Two novel approaches to language modelling have been investigated. The first model makes use of external knowledge (i.e. regional data), enabling it to make well-informed decisions about what topics are likely to come up and hence biasing the probabilities towards the vocabulary of the topic.

As research progresses, it is likely that the regional models can be made more powerful, and consequently enable a real-time reduction of the hyotheses space for the ASR.

The second model attacks the language-modelling problem from a different angle. Instead of dividing the language model into a number of regions, it tries to find features that are common for *all* regions and languages. More specificially, it finds a number of subspaces that encapsulate the most important dimensions of the data, and the initial experiments suggest that these dimensions may even be shared between different languages.

A future system could possibly use these shared dimensions to recognize speech on languages where the amount of training data is minimal.

There is still plenty of research left before any definite conclusions can be drawn, but the future does indeed look promising.

References

[1]Use of geographical meta-data in ASR language and acoustic models
Bocchieri, E.; Caseiro, D., AT&T Res., Florham Park, NJ, USA
Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference, page 5118-5121.

[2] Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models
Agarwal, M; Akyazi, P; Feng, K; et al.
2010 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, : 4334-4337

[3] Subspace Gaussian Mixture Models for Speech Recognition
Daniel Povey; Lukas burget et al.
2010, Submitted to: ICASSP