

Retrieving a user language model from an unsupervised document map

Mikko Kurimo and Krista Lagus

Neural Networks Research Centre, Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Finland, Email: `Mikko.Kurimo@hut.fi`

An extended abstract for the “Machine Learning Meets the User Interface” workshop. The targeted discussion topics are the following ones (1. is the main topic):

1. “*User modeling and personalization*”
2. “*Computerized support for meetings*”

This work presents a method to automatically retrieve a statistical language model focused on the specific topic and style of the speech situation at hand.

The retrieval is based on a feature vector representation [2] of a sample text or an approximative intermediate speech transcription. This so-called document vector will be compared to an index of pre-trained language models and the best models are retrieved. The index is not just a list of language models but itself a smooth topological representation of language topics and styles found in all processed training material. The index is organized in a fully unsupervised manner by applying the Self-organizing map algorithm [1] to the document vectors obtained from all training documents [2]. The self-organization drags similar documents close to each other and gradually forces a local ordering in the map.

In addition to smoothing the document reference vectors with their neighbors, the organized structure of the index can be utilized in other ways, as well. First, if the index is very large, the search for the best-matching reference can be boosted by performing a local search instead of a global one. In this way we can start the search based on any hint, such as the best match in the previous search, and proceed to search to its neighbors by following the topological structure defined by the index map. Another and even more important advantage of the ordered index is in the construction of the language models. While the set of documents pointed by a single reference in the index may sometimes be too small for creating a proper statistical language model, it is easy to increase the training set of relevant documents by including the sets from the neighboring indices, as well.

The focused language model is formed by interpolating the prototype language models retrieved from the index. Because the topic and style of the language sample can seldom be exactly specified, or it may contain several topics, or even fall between the modeled topics, the index can be set to point to language model prototypes trained on different amounts of neighboring indices. The interpolation of the probabilities that the retrieved language models provide may also include a general model for the whole data as suggested in [3]. This ensures an adequate model for such rare events that may have been missed by the small topical language models. The suitability of the chosen overlapping language models can be evaluated by measuring the modeling accuracy by perplexity or word entropy on the sample text or the initial speech transcription.

An example application of this language model structure is a multipurpose speech recognition tool that adapts to the user and speaking situation by operating in several passes over the speech data. In addition to model adaptation, the multipass recognition approach provides efficient means to enhance the recognition speed. The first pass can be performed with approximative general models and the second pass can concentrate on

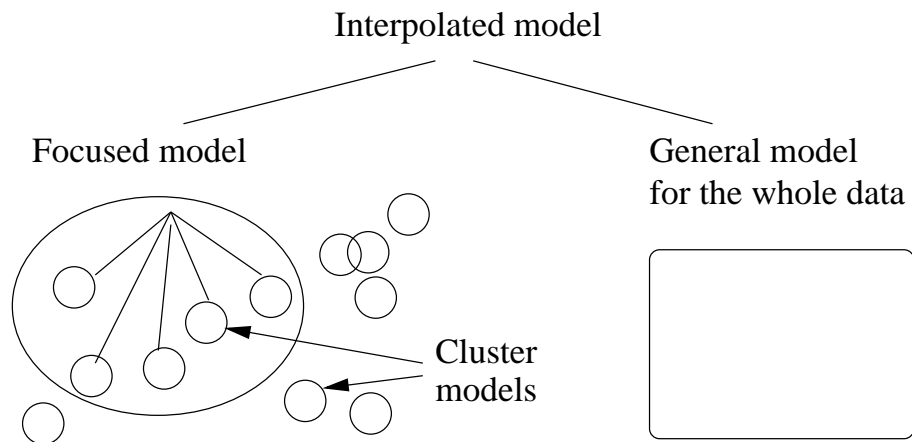


Figure 1: A focusing language model obtained as an interpolation between topical cluster models and a general model.

retrieving better models and evaluating them on a pruned set of transcription hypothesis, so that the additional computational overhead will be negligible [3]. The application of focusing language models is naturally not limited to interaction with online speech. Typical prototype applications include also the transcription of a large variety of different historical speech recordings [4], and recordings of different speakers in a meeting [5] to create a searchable and browsable audio index of the content.

The important aspect in the suggested method from the HCI perspective is the ability to efficiently obtain a focused model suitable for the required speech situation. The experiments in Finnish and English corpora [3, 4] show that using a properly focused language model leads to better speech recognition and higher language modeling accuracy.

References

- [1] Teuvo Kohonen. *Self-Organizing Maps*. Springer, Berlin, 2001. 3rd ed.
- [2] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Vesa Paatero, and Antti Saarela. Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585, May 2000.
- [3] Mikko Kurimo and Krista Lagus. An efficiently focusing large vocabulary language model. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN’02)*, pages 1068–1073, Madrid, Spain, 2002.
- [4] Mikko Kurimo, Bowen Zhou, Rongqing Huang, and John H.L. Hansen. Language modeling structures in audio transcription for retrieval of historical speeches. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004. (Submitted).
- [5] Steve Renals and Dan Ellis. Audio information access from meeting rooms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003. (In Special Session on Smart Meeting Rooms).