Multimodal interfaces

Chapter 7

Content-based information retrieval and analysis

Erkki Oja, Jorma Laaksonen, Markus Koskela, Ville Viitaniemi, Zhirong Yang, Mats Sjöberg, Hannes Muurinen

7.1 Introduction

Content-based image or information retrieval (CBIR) has been a subject of intensive research effort for more than a decade now. Content-based retrieval of images differs from many of its neighboring research disciplines in computer vision due to one notable fact: human subjectivity cannot totally be isolated from the use and evaluation of CBIR systems.

In our PicSOM[1] CBIR system, parallel Self-Organizing Maps (SOMs) have been trained with separate data sets obtained from the multimodal object data with different feature extraction techniques. The different SOMs and their underlying feature extraction schemes impose different similarity functions on the images, videos, texts and other media objects. In the PicSOM approach, the system is able to discover those of the parallel SOMs that provide the most valuable information for retrieving relevant objects in each particular query.

7.2 Benchmark tasks of natural image content analysis

In the course of previous years we have outlined and implemented our generic PicSOM system architecture for image and information retrieval tasks. The architecture is based on extraction of numerous different features from the feature descriptors from the information objects, performing inference separately based on each feature, and fusing the partial inferences. The architecture supports hierarchical organization of the information objects. In the case of image analysis, the hierarchy is used to describe the decomposition of images into segments.

We have investigated how our architecture can be applied to various benchmark tasks concerning generic domain photographic images. While individual components of the architecture have been improved during the studies, the general architecture has proven to be successful. The improvements include the incorporation of new feature extraction methods, most notably the Scale-Invariant Feature Transform (SIFT) features calculated from interest points, the use of Support Vector Machines (SVMs) as an alternative to SOMs as the classification method, and alternative early and late feature feature fusion methods.

Our group has participated in the annual PASCAL FP6 NoE Visual Object Classes (VOC) Challenges [1, 2]. The material of the Challenges consists of photographic images of natural scenes containing objects from predefined object classes. In 2006 there were approximately 5000 images and ten object classes, including objects such as "bicycle","bus","cat" and "cow". For the 2007 Challenge the number of images and object classes were both doubled. The Challenge included the classification task, ie. the determination whether an object of a particular class appears in the image, and the detection task for the object's bounding box. In addition, the 2007 Challenge also included a novel competition of pixel-wise object segmentation. Our performance in the Challenge has been satisfactory, the highlights being the best segmentation accuracy and the fourth best classification performance in the 2007 Challenge.

For the VOC benchmarks we have investigated and analyzed techniques of automatic image segmentation, especially in [3]. The devised techniques have been fundamental for performing the bounding box detection tasks. However, for the classification task the usefulness of segmented images does not currently seem to be competitive against state-of-the-art global image analysis techniques. Partly this is due to the strong correlation of

[1] `http://www.cis.hut.fi/picsom`

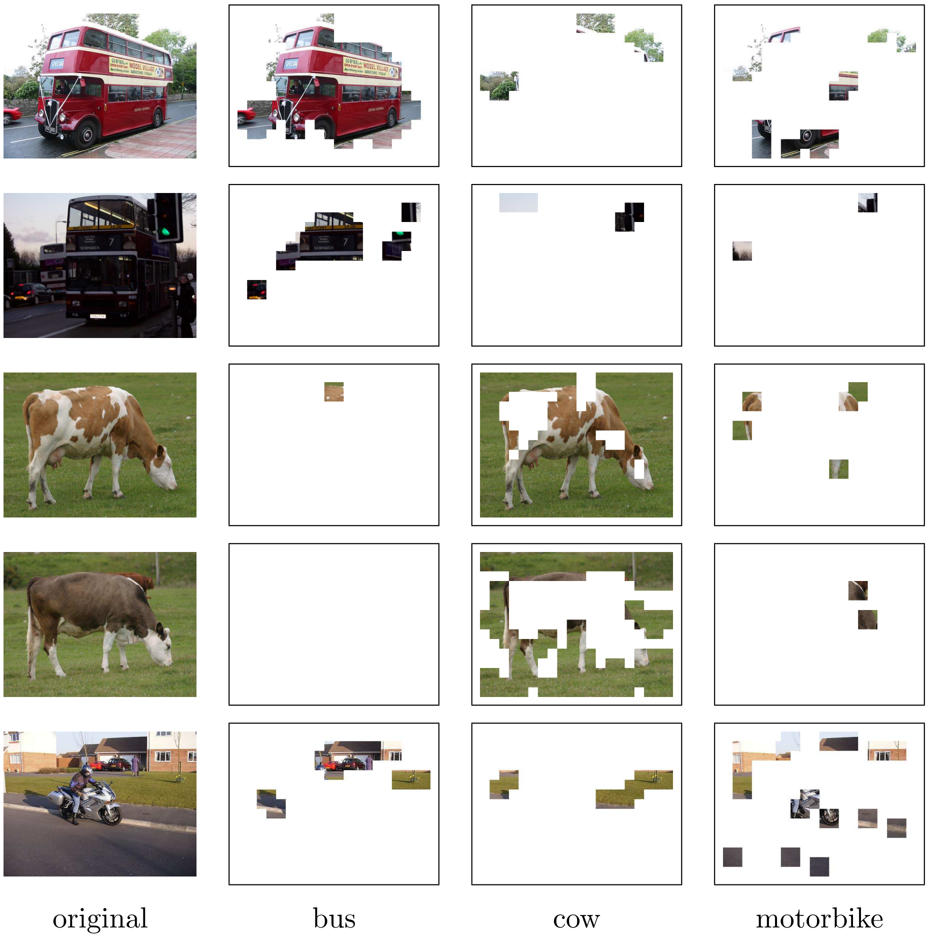


Figure 7.1: Images of the VOC2006 image collection shown together with those patches that on the collection level contribute most to the classification of the images as a “bus”, “cow” and “motorbike”.

actual target objects and the background both in the challenge databases and in natural images in general, which diminishes the advantage from focusing analysis exclusively to specific image locations. This effect is illustrated in Figure 7.1 where we have highlighted the image patches that contribute most to the decision of the image containing a particular object in the classification task.

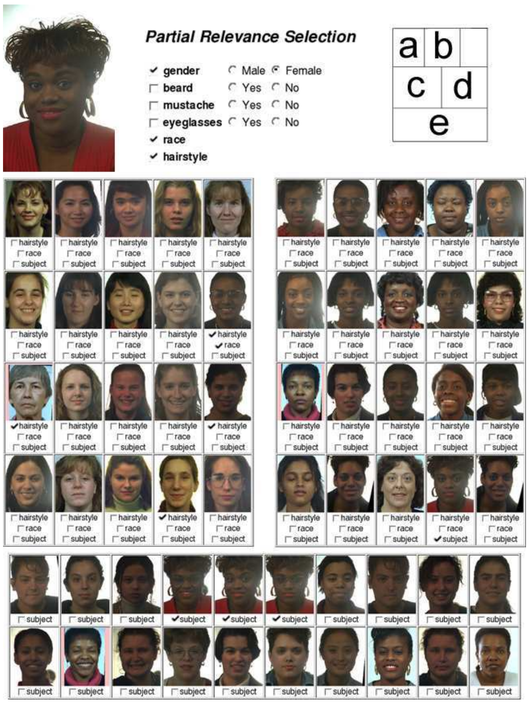
Other benchmark tasks we have studied include the ImageCLEF 2006 object annotation task, which we analyzed outside the competition, and the ImageCLEF 2007 object retrieval task, in which our results were clearly the best of the campaign submissions. We have also applied our CBIR system to benchmark tasks of automatic image annotation, performing clearly better than numerous state-of-the-art methods reported in literature [4, 5].

7.3 Interactive facial image retrieval

It is often desired to search for an image depicting a person only through an eyewitness' recalling about the appearance. Interactive computer-based systems for this purpose, however, confront the problem of evaluation fatigue due to time-consuming retrieval. We have addressed this problem by extending our PicSOM CBIR system to emphasize the early occurrence of the first subject image. Partial relevance criteria provide a common language understood by both the human user and the computer system. In addition to filtering by ground truth and hard classifier predictions, we have proposed Discriminative Self-Organizing Maps (DSOMs) [6] to adaptively learn the partial relevances.

A straightforward method to obtain DSOMs is to employ discriminant analysis as a preprocessing step before normal SOM training. We have applied the widely used method, PCA+LDA, in pattern recognition as our baseline. Furthermore, we have adapted the Informative Discriminant Analysis (IDA) to maximize the discrimination for more complicated distributions. Our Parzen Discriminant Analysis [7] regularizes the IDA objective by emphasizing the prior of piecewise smoothness in images. Both LDA and our PDA have been extended for handling fuzzy cases. The original IDA optimization algorithm is computationally expensive. We have presented three acceleration strategies [8]: First, the computation cost of batch gradients is reduced by using matrix multiplication. Second, the updates follow an geodesic flow in the Stiefel manifold without Givens reparameterization. Third, a more efficient leading direction is calculated by preserving only the principal whitened components of the batch gradient at each iteration.

Simulations have been performed on the FERET database. We have provided a query example (Figure 7.2) and also presented a quantitative study on the advantage in terms of the first subject hit and retrieval precisions at various recall levels [6].



Partial Relevance Selection

✓ gender ○ Male ● Female
□ beard ○ Yes ○ No
□ mustache ○ Yes ○ No
□ eyeglasses ○ Yes ○ No
✓ race
✓ hairstyle

Figure 7.2: A query example using the PicSOM system: (a) the target person; (b) specifying partial relevance; the displayed images in the first phase, with (c) the first round and (d) the second round; (e) the images displayed in the first round of the second phase.

7.4 Content analysis and change detection in earth observation images

Earth observation (EO) data volumes are growing rapidly, with an increase in both the number of satellite sensors and in their resolutions. Yet, it is estimated that only 5% of all EO data collected up to now has been used. Therefore, traditional remote sensing archiving systems – with queries made typically on sensor type, geographical extents or acquisition date – could become obsolete as the amount of data to be stored, accessed and processed explodes. Using image content indexing would allow a more efficient use of these databases. This has led to the emergence of content-based image retrieval systems for archive management of remote sensing images and for annotation or interpretation of satellite images. In co-operation with the VTT Technical Research Centre of Finland, we have applied the PicSOM system for analysis of multispectral and polarimetric radar (PolSAR) satellite images divided in small patches or *imagelets*.

With the high-resolution optical images the aim has been to detect man-made structures and changes on the studied land cover. Fusion of panchromatic and multispectral information was done conveniently within the PicSOM framework, in which several SOMs are trained in parallel, one SOM per feature. Qualitative and quantitative evaluation of the methods were carried out for man-made structure detection and change detection, using partially labeled datasets. The results were encouraging, considering that a totally new approach was presented to the challenging problem of change detection in very high-resolution images [9]. Possible applications of this work are high-resolution satellite image annotation and monitoring of sensitive areas for undeclared human activity, both in an interactive way.

With the radar images, the availability of dual-polarization and fully-polarimetric data, instead of earlier single-polarization data, will in the near future enable a deeper analysis of backscattering processes. This development will in turn pave the way for many new applications for spaceborne SAR data. At the same time, these satellite missions generate a huge amount of data at a higher resolution than previous spaceborne SAR sensors. It is still quite unclear what low-level features will be the most efficient ones for the automatic content analysis of the satellite polarimetric SAR data. In our research [10] we have compared six different types of polarimetric features and their different postprocessings, including averages and histograms, to gain quantitative knowledge of their suitability for the land cover classification and change detection tasks. The results proved that different features are most discriminative for different land cover types, and the best overall performance can be obtained by using a proper combination of them.

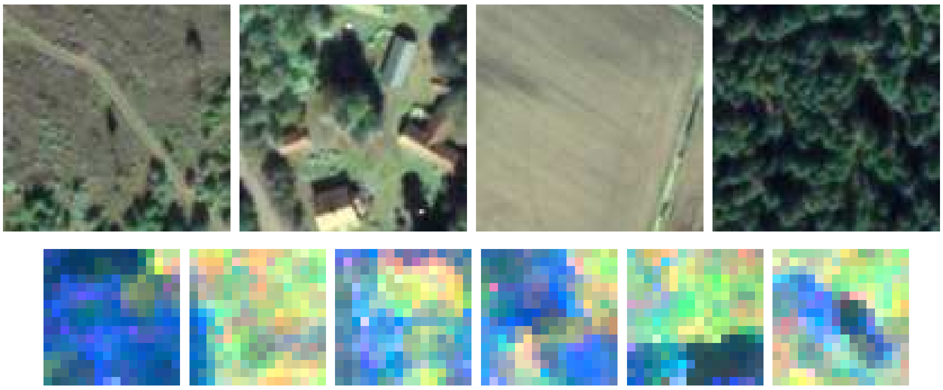


Figure 7.3: 100×100-pixel optical and 16×16-pixel SAR (Pauli decomposition) imagelets.

7.5 Multimodal hierarchical objects in video retrieval

The basic ideas of content-based retrieval of visual data can be expanded to multimodal data, where we consider multimodal objects, for example video or images with textual metadata. The PicSOM system has been extended to support general multimodal hierarchical objects and to provide a method for relevance sharing between these objects [11]. For example a web page with text, embedded images and links to other web pages can be modeled as a hierarchical object tree with the web page as the parent object and the text, links and images as children objects. The relevance assessments originally received from user feedback will then be transferred from the object to its parents, children and siblings. For example, if we want to search for an image of a cat from a multimedia message database, we can let the system compare not only the images, but also the related textual objects. If the reference message text contains the word "cat" we can find images which are not necessarily visually similar, but have related texts containing the same keyword.

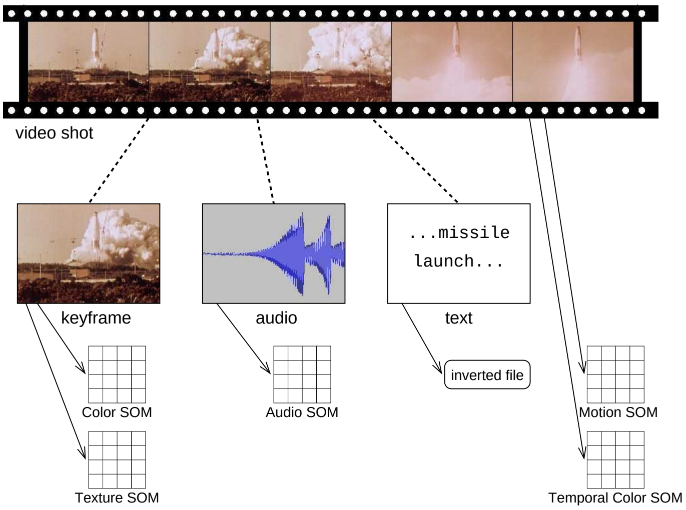


Figure 7.4: The hierarchy of video and multimodal SOMs.

The multimodal hierarchy used for indexing video shots and supporting multimodal fusion between the different modalities is illustrated in Fig. 7.4. The video shot itself is considered as the main or parent object in the tree structure. The keyframes (one or more) associated with the shot, the audio track, and text obtained with automatic speech recognition are linked as children of the parent object. All object modalities may have one or more SOMs or other feature indices, and thus all objects in the hierarchy can have links to a set of associated feature indices.

A common approach to semantic video retrieval is to combine separate retrieval results obtained with low-level visual features and text-based search. The relative weights of these sub-results are specified based on e.g. validation queries or query categorization.

An important catalyst for research in video retrieval is provided by the annual TREC Video Retrieval Evaluation (TRECVID) workshop. The goal of the workshop series is to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested to compare their results. The search task in TRECVID models the task of an intelligence analyst who is looking for specific segments of video containing persons, objects, events, locations, etc. of current interest. The task is defined as follows: given a search test collection and a multimedia statement of information need, return a ranked list of shots which best satisfy the need. We have successfully participated in TRECVID annually since 2005 [12, 13].

7.6 Semantic concept detection

Extracting semantic concepts from visual data has attracted a lot of research attention recently. The aim of the research has been to facilitate semantic indexing and concept-based retrieval of unannotated visual content. The leading principle has been to build semantic representations by obtaining intermediate semantic levels (objects, locations, events, activities, people, etc.) from automatically extracted low-level features. The modeling of mid-level semantic concepts can be useful in supporting high-level indexing and querying on multimedia data, as such concept models can be trained off-line with considerably more positive and negative examples than what are available at query time.

We treat semantic concept detection from shot-segmented videos as a general supervised classification task by utilizing the hierarchical approach shown in Fig. 7.4 and by extracting multiple low-level features from the different data modalities [14]. A set of SOMs is trained on these features to provide a common indexing structure across the different modalities. The particular features used for each concept detector are obtained using sequential forward feature selection. The method has proven to be readily scalable to a large number of concepts, which has enabled us to model e.g. a total of 294 concepts from a large-scale multimedia ontology [15] and utilize these concept models in TRECVID video search experiments [12]. Figure 7.5 lists and exemplifies the 36 semantic concepts detected for the TRECVID 2007 high-level feature extraction task.

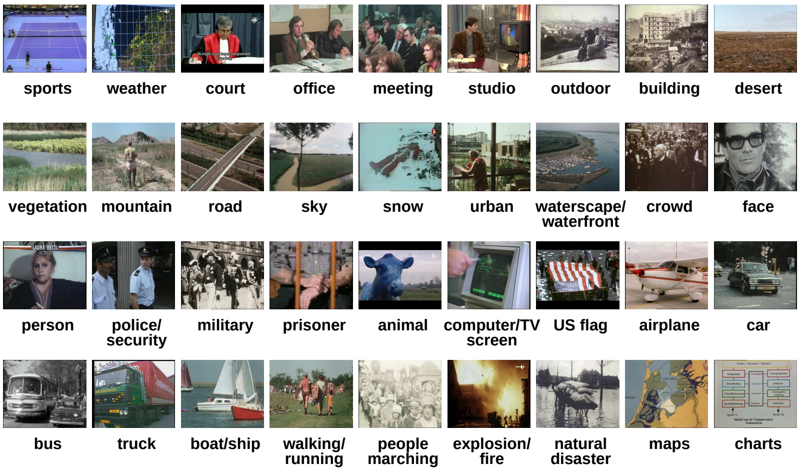


Figure 7.5: The set of 36 semantic concepts used in TRECVID 2007.

Semantic concepts do not exist in isolation, but have different relationships between each other, including similarities in their semantic and visual (low-level) characteristics, co-occurrence statistics, and different hierarchical relations if a taxonomy has been defined for the concepts. We have studied how multimedia concept models built over a general clustering method can be interpreted in terms of probability distributions and how the quality of such models can be assessed with entropy-based methods [16].

In addition we also explored the possibility of taking advantage of temporal and inter-concept co-occurrence patterns of the high-level features using n-gram models and clustering of temporal neighborhoods. The method was found to be very useful in our TRECVID 2007 experiments [13].

7.7 Shot boundary detection

We have applied our general multimedia analysis framework to shot boundary detection and summarization of video data. Our approach for shot boundary detection utilizes the topology preservation properties of SOMs in spotting the abrupt and gradual shot transitions. Multiple feature vectors calculated from consecutive frames are projected on two-dimensional feature-specific SOMs. The transitions are detected by observing the trajectories formed on the maps.

Due to the topology preservation, similar inputs are mapped close to one another on the SOMs. The trajectory of the best-matching map units of successive frames thus typically hovers around some region of a SOM during a shot, provided that the visual content of the video does not change too rapidly. Abrupt cuts are characterized by sudden trajectory leaps from one region on the map to another, and gradual transitions on the other hand are characterized by a somewhat rapid drift of the trajectory from one region to another. The detector tries to detect these kinds of characteristic phenomena.

To increase detector robustness and prevent false positive cut detection decisions, e.g. due to flashlights, we do not only monitor the rate of change of the map position between two consecutive frames, but take small frame windows from both sides of the current point of interest, and compare the two frame windows. A circular area with a constant radius is placed over each map point in the given frame window as illustrated in Figure 7.6. We call the union of these circular areas the area spanned by the frame window. If the areas spanned by the preceding and following frame windows overlap, there are some similar frames in both of them, and we decide that the current point of interest is not a boundary point. If there is no overlapping, the frames in the frame windows are clearly dissimilar, and we decide that we have found a boundary. The flashlights are characterized by sudden trajectory leaps to some region on the map followed by a leap back to the original region. If the duration of the flashlight is smaller than the frame window size, the proposed method helps to avoid false positives.

The final boundary decision is done by a committee machine that consists of this kind of parallel classifiers. There is one classifier for each feature calculated from the frames, and each classifier has a weight value. The final decision is made by comparing the weighted vote result of the classifiers against a threshold value. Abrupt cuts and gradual transitions are detected using the same method. The detected boundary points that are close to one another are combined, and as the result we get the starting locations and lengths of the transitions. To facilitate detection of slow gradual transitions, our system also allows to use a frame gap of given length between the two frame windows. A more detailed description and quantitative results with the algorithm are given in [17].

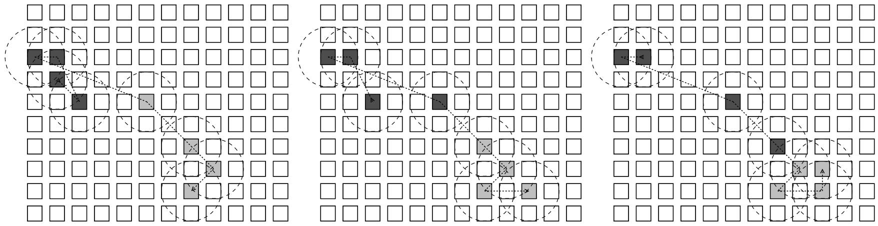


Figure 7.6: Segments of a trajectory at three consecutive time steps. The SOM cells marked with a dark gray color represent trajectory points belonging to the set of preceding frames, and light gray cells represent the following frames. The circles represent the area spanned by the preceding and following frame sets.

7.8 Video summarization

Video summarization is a process where an original video file is converted to a considerably shorter form. The video summary can then be used to facilitate efficient searching and browsing of video files in large video collections. The aim of successful automatic summarization is to preserve as much as possible from the essential content and overall structure. Straightforward methods such as frame subsampling and fast forwarding produce incoherent summaries that are strenuous to view and cannot usually be absorbed with a single viewing. The strategy of selecting parts of the video using a fixed interval can easily lose important information. More sophisticated summarization algorithms typically use shot-based segmentation and analysis. However, including each shot in the summary may not be optimal as certain shots may be almost duplicates of each other or there may be too many of them for a concise summary, depending on the original material.

There are two fundamental types of video summaries: *static abstracts or storyboards* and *video skims*. The former typically consist of collections of keyframes extracted from the video material and organized as a temporal timeline or as a two-dimensional display. Video skims consist of collections of selected video clips from the original material. Both these types of summaries can be useful, depending on the intended application. Storyboards provide static overviews that are easily presented and browsed in many environments, whereas skims preserve the original media type and can also contain dynamic content such as important events in the original video.

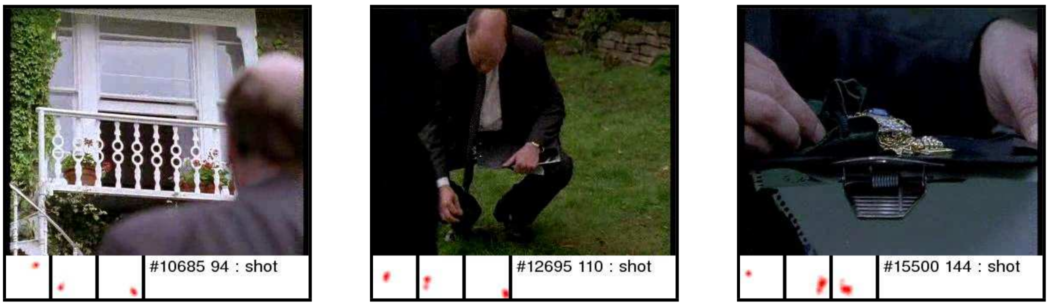


Figure 7.7: Representative frames and SOM signatures of three video shots.

We have developed a technique for video summarization as video skims [18] using SOMs trained with standard visual features that have been applied in various multimedia analysis tasks. The method is based on initial shot boundary detection providing us with lists of shots, which are used in the following stages as basic units of processing. We detect and remove unwanted “junk” shots (e.g. color bar test screens, empty frames) from the videos, and apply face detection and motion activity estimation. Next, we compute the visual similarities between all pairs of shots and remove overly similar shots. We trace the trajectory of the frames within the shot in question and record the corresponding BMUs. The set of BMUs constitutes a SOM-based signature for the shot, which can then be compared to other shots’ signatures to determine whether a shot is visually unique or similar to some other shots. Fig. 7.7 shows example frames from three shots and the convolved SOM-based trajectory signatures of those shots as red-colored responses on the SOM surfaces. Each remaining shot is then represented in the summary with a separately selected one-second clip. The selected clips are finally combined using temporal ordering and fade-outs and fade-ins from black.

We participated in the TRECVID 2007 rushes summarization task [18] and obtained very promising results. Our summarization algorithm obtained average ground-truth inclusion performance with the shortest overall summaries over all the submissions.

References

[1] Mark Everingham, Andrew Zisserman, Chris Williams, and Luc Van Gool. The Pascal Visual Object Classes Challenge 2006 (VOC2006) results. Technical report, 2006. Available on-line at `http://www.pascal-network.org/`.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[3] Ville Viitaniemi and Jorma Laaksonen. Techniques for still image scene classification and object detection. In *Proceedings of 16th International Conference on Artificial Neural Networks (ICANN 2006)*, volume 2, pages 35–44, Athens, Greece, September 2006. Springer.

[4] Ville Viitaniemi and Jorma Laaksonen. Evaluating the performance in automatic image annotation: example case by adaptive fusion of global image features. *Signal Processing: Image Communications*, 22(6):557–568, July 2007.

[5] Ville Viitaniemi and Jorma Laaksonen. Improving the accuracy of global feature fusion based image categorisation. In Bianca Falcidieno, Michela Spagnuolo, Yannis S. Avrithis, Ioannis Kompatsiaris, and Paul Buitelaar, editors, *Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies (SAMT 2007)*, volume 4669 of *Lecture Notes in Computer Science*, pages 1–14, Genova, Italy, December 2007. Springer.

[6] Zhirong Yang and Jorma Laaksonen. Interactive content-based facial image retrieval with partial relevance and parzen discriminant analysis. *Pattern Recognition Letters*, 2008. In submission.

[7] Zhirong Yang and Jorma Laaksonen. Face recognition using Parzenfaces. In *Proceedings of International Conference on Artificial Neural Networks (ICANN'07)*, volume 4669 of *Lecture Notes in Computer Science*, pages 200–209, Porto, Portugal, September 2007. Springer.

[8] Zhirong Yang and Jorma Laaksonen. Principal whitened gradient for information geometry. *Neural Networks*, 2008. In press.

[9] Matthieu Molinier, Jorma Laaksonen, and Tuomas Häme. Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):861–874, April 2007.

[10] Matthieu Molinier, Jorma Laaksonen, Yrjö Rauste, and Tuomas Häme. Detecting changes in polarimetric SAR data with content-based image retrieval. In *Proceedings of IEEE International Geoscience And Remote Sensing Symposium*, Barcelona, Spain, July 2007. IEEE.

[11] Erkki Oja, Mats Sjöberg, Ville Viitaniemi, and Jorma Laaksonen. Emergence of semantics from multimedia databases. In Gary Y. Yen and David B. Fogel, editors, *Computational Intelligence: Principles and Practice*, chapter 9. IEEE Computational Intelligence Society, 2006.

[12] Mats Sjöberg, Hannes Muurinen, Jorma Laaksonen, and Markus Koskela. PicSOM experiments in TRECVID 2006. In *Proceedings of the TRECVID 2006 Workshop*, Gaithersburg, MD, USA, November 2006.

[13] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen, and Philip Prentis. PicSOM experiments in TRECVID 2007. In *Proceedings of the TRECVID 2007 Workshop*, Gaithersburg, MD, USA, November 2007.

[14] Markus Koskela and Jorma Laaksonen. Semantic concept detection from news videos with self-organizing maps. In Ilias Maglogiannis, Kostas Karpouzis, and Max Bramer, editors, *Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations*, pages 591–599, Athens, Greece, June 2006. IFIP, Springer.

[15] Milind Naphade, John R. Smith, Jelena Tešić, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[16] Markus Koskela, Alan F. Smeaton, and Jorma Laaksonen. Measuring concept similarities in multimedia ontologies: Analysis and evaluations. *IEEE Transactions on Multimedia*, 9(5):912–922, August 2007.

[17] Hannes Muurinen and Jorma Laaksonen. Video segmentation and shot boundary detection using self-organizing maps. In *Proceedings of 15th Scandinavian Conference on Image Analysis (SCIA 2007)*, pages 770–779, Aalborg, Denmark, June 2007.

[18] Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Ville Viitaniemi, and Hannes Muurinen. Rushes summarization with self-organizing maps. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 45–49, Augsburg, Germany, September 2007. ACM Press.

Chapter 8

Automatic speech recognition

Mikko Kurimo, Kalle Palomäki, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Ville Turunen, Sami Virpioja, Matti Varjokallio, Ulpu Remes, Antti Puurula

8.1 Introduction

Automatic speech recognition (ASR) means an automated process that inputs human speech and tries to find out what was said. ASR is useful, for example, in speech-to-text applications (dictation, meeting transcription, etc.), speech-controlled interfaces, search engines for large speech or video archives, and speech-to-speech translation.

Figure 8.1 illustrates the major modules of an ASR system and their relation to applications. In *feature extraction*, signal processing techniques are applied to the speech signal in order to dig out the features that distinguish different phonemes from each other. Given the features extracted from the speech, *acoustic modeling* provides probabilities for different phonemes at different time instants. *Language modeling*, on the other hand, defines what kind of phoneme and word sequences are possible in the target language or application at hand, and what are their probabilities. The acoustic models and language models are used in *decoding* for searching the recognition hypothesis that fits best to the models. Recognition output can then be used in various applications.

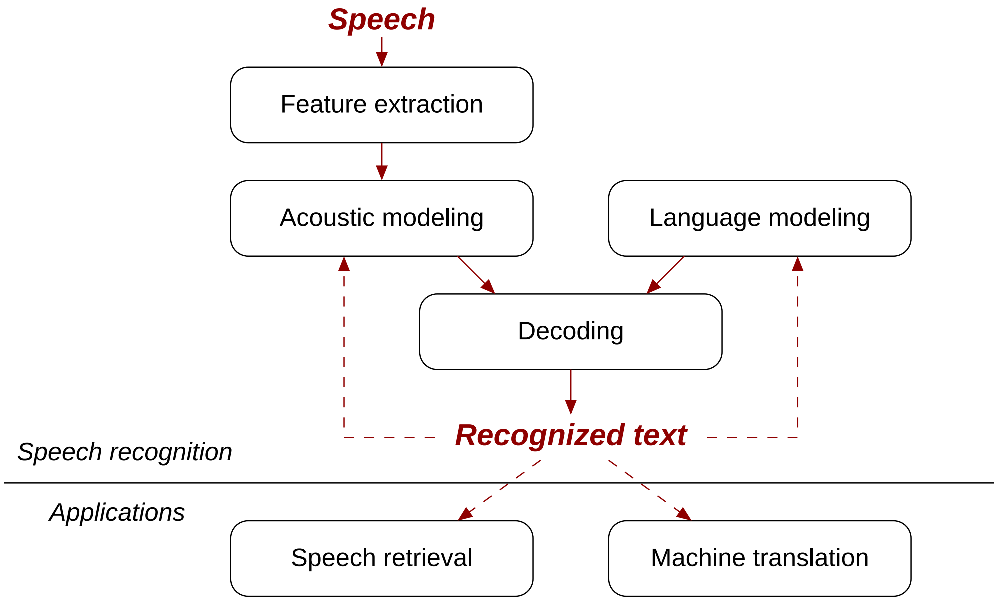


Figure 8.1: The main components of an automatic speech recognition system and their relation to speech retrieval and machine translation applications.

Our focus in ASR is large vocabulary continuous speech recognition (LVCSR). For several years, we have been developing new machine learning algorithms for each of the subfields and building a complete state-of-the-art recognizer to evaluate new methods and their impact. Originally, the recognizer was constructed for fluent and planned speech such as Finnish newsreading, where language models covering a very large vocabulary are required. Besides newsreading, other example tasks are political and academic speeches and other radio and television broadcasts where the language used is near the written style. Sofar, we have not seriously attempted to recognize Finnish spontaneous conversations, because enough Finnish training texts for learning the corresponding style do not exist. Our main training corpus for language modeling is the Finnish Language Bank at CSC. For acoustic modeling we use voice books, Finnish Broadcast Corpus at CSC and the SPEECON corpus.

In addition to the recognition of Finnish, we have performed experiments in English, Turkish and Estonian. To make this possible we have established research relations to different top speech groups in Europe and U.S., e.g. University of Colorado, International Computer Science Institute ICSI, Stanford Research Institute SRI, IDIAP, University of Edinburgh, University of Sheffield, Bogazici University, and Tallinn University of Technology. The forms of collaboration have included researcher exchanges, special courses, workshops and joint research projects. We have also participated in several top international and national research projects funded by EU, Academy of Finland, Tekes, and our industrial partners. In the close collaboration with our Natural Language Processing group 10 we are also organizing an international competition called Morphochallenge to evaluate the best unsupervised segmentation algorithms for words into morphemes for information retrieval, LVCSR and language modeling in different languages. This challenge project is funded by EU's PASCAL network and described in Chapter 10.

8.2 Acoustic modeling

Acoustic modeling in automatic speech recognition means building statistical models for some meaningful speech units based on the feature vectors computed from speech. In most systems the speech signal is first chunked into overlapping 20-30 ms time windows at every 10 ms and the spectral representation is computed from each frame. A commonly used feature vector consists of mel-frequency cepstral coefficients (MFCC) which are the result of the discrete cosine transform (DCT) applied to the logarithmic mel-scaled filter bank energies. Local temporal dynamics can be captured by concatenating the first and second order delta features (time differences) to the basic feature vector.

Although the use of DCT and delta features are well-established methods for processing speech features, they are by no means optimal. Better features can be constructed by learning from the data which features would best discriminate between speech sounds. A well known method for this is the linear discriminant analysis (LDA), which can be used to process the spectral input for creating new discriminative features. As a simple method LDA has its limitations, and therefore in [1] we studied different methods to enhance its operation. The result was the pairwise linear discriminant (PLD) features, which unlike most LDA extensions are simple to compute but still work in speech recognition better than the traditional methods.

Closely connected to the feature extraction is the speaker-wise normalization of the features. One commonly used method is the vocal tract length normalization (VTLN). It requires estimating only a single normalization parameter yet still provides significant improvements to the speech recognition. The estimation, however, can not be done in closed form, so an exhaustive search over a range of parameters is usually used. We have devised a method which greatly simplifies the estimation of the VTLN parameter but still gives competitive performance [2]. It is especially attractive when used with discriminative feature extraction, such as with PLD.

The acoustic feature sequence in ASR is typically modeled using hidden Markov models (HMM). In basic form each phoneme is modeled by a separate HMM, where the emission distributions of the HMM states are Gaussian mixtures. An example is shown in Figure 8.2. In practice, however, we need to take the phoneme context into account, so that for each phoneme there are separate HMMs for various phoneme contexts. This leads easily to very complex acoustic models where the number of parameters is in order of millions.

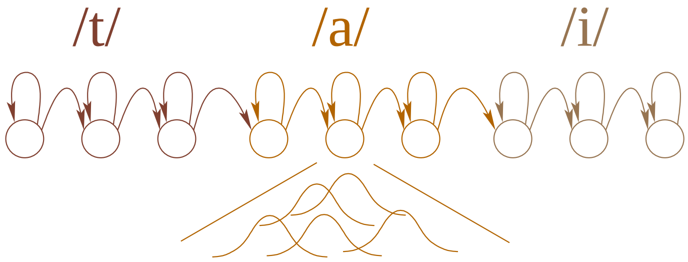


Figure 8.2: Each phoneme is modeled with a hidden Markov model, usually consisting of three states. The state distributions are modeled by Gaussian mixture models.

To limit the number of parameters and thereby allow robust estimation of the acoustic models, the covariance matrices of the Gaussian mixture components are usually assumed diagonal. This is a relatively reasonable assumption, because there is typically a whitening transform (DCT or similar) applied to the feature vector. The uncorrelatedness is, however, a global property and there are always correlations on the state level. The correlations can be modeled by adding more mixture components in the direction of most variance, which is sometimes called as *implicit covariance modeling.* Modeling covariances *explicitly* instead has some clear benefits as fewer modeling assumptions typically lead to more robust models. Constraining the exponential parameters of the Gaussians to a subspace is appealing for speech recognition, as the computational cost of the acoustic model is also decreased. A subspace constraint on the inverse covariance matrices was shown to give a good performance [3] for LVCSR tasks.

To ensure high quality research we constantly put considerable effort to keep our speech recognition system up-to-date. One major recent improvement to our system has been the introduction of discriminative acoustic training. The use of discriminative training has been a growing trend during the last decade and some form of it is now a necessity for a state-of-the-art system. Our implementation allows using several different training criteria such as maximum mutual information (MMI) and minimum phone error (MPE) [4] over the traditional maximum likelihood (ML) training. It also enables gradient based optimization in addition to the commonly used extended Baum-Welch method. Discriminative training techniques have already given very promising results and they will be an important research direction in the future.

Speaker segmentation

In addition to feature normalization methods such as the vocal tract length normalization (VTLN), acoustic model adaptation is often used for increased robustness against speaker variation. Speaker normalization and adaptation generally improve the speech recognition performance substantially, but they cannot be applied unless the speech recognition system knows who spoke and when. Often there is no such information about the speakers, but automatic speaker segmentation is needed. Speaker segmentation (i) divides the audio to speaker turns (speaker change detection) and (ii) labels the turns according to speaker (speaker tracking) as illustrated in Figure 8.3. While most speaker segmentation methods have been developed primarily for audio content or spoken dialogue analysis, we focused on speaker segmentation for speaker adaptation. We developed a speaker tracking method that seeks to directly maximize the feature likelihood when we assume the features are adapted to speaker using the segmentation results and acoustic model adaptation with constrained maximum likelihood linear regression (CMLLR). The proposed method performed well when tested on Finnish television news audio in [5].

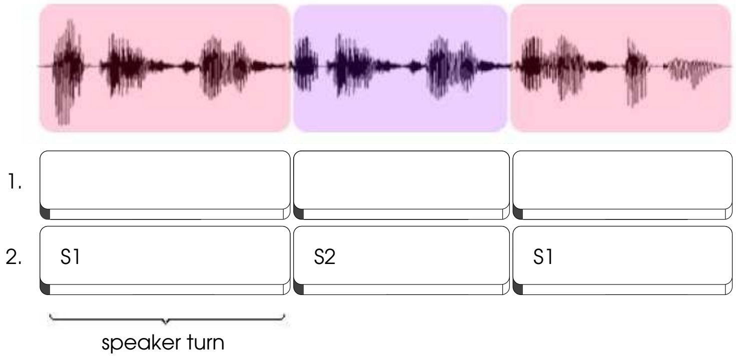


Figure 8.3: Speaker segmentation first divides the audio to speaker turns according to where speakers change and then labels the detected turns. Speaker labels are created on-line and no prior information about the speakers (e.g. training data or speaker models) is needed.

Recognition of reverberant speech

Research in the acoustic modeling for large vocabulary continuous speech recognition was concentrated mostly on fairly noise free conditions (see Sect. 8.2). In the field noise robust speech recognition we have been developing techniques suitable for recognition in highly reverberant spaces. This research has been collaborative with the University of Sheffield. Our approach is based on missing data approach [9], in which noisy, reverberated regions are treated as unreliable and noise free regions as reliable evidence of speech. Different treatments of reliable and unreliable parts of speech is achieved by a modification of Gaussian mixture model proposed by Cooke et al. [9]. Our approach to reverberant speech recognition is based on detecting reliable regions of speech from strong onsets at modulation rates characteristic to speech [8]. In recent developments of the model we have sought modeling solutions that more closely match on perceptual data considering the recognition of reverberant speech by human listeners [6, 7].

References

[1] J. Pylkkönen, LDA Based Feature Estimation Methods for LVCSR. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh PA, USA, pages 389–392, 2006.

[2] J. Pylkkönen, Estimating VTLN Warping Factors by Distribution Matching. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, pages 270–273, 2007.

[3] M. Varjokallio, M. Kurimo, Comparison of Subspace Methods for Gaussian Mixture Models in Automatic Speech Recognition. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, pages 2121-2124, 2007.

[4] D. Povey and P. C. Woodland, Minimum Phone Error and I-smoothing for Improved Discriminative Training. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, pages I-105–108, 2002.

[5] U. Remes, J. Pylkkönen, and M. Kurimo, Segregation of Speakers for Speaker Adaptation in TV News Audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, USA, pages IV-481–484, 2007.

[6] G. J. Brown and K. J. Palomäki Reverberation, in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, eds. by DeLiang Wang and Guy J. Brown, Wiley/IEEE Press, 2006.

[7] G. J. Brown and K. J. Palomäki A reverberation-robust automatic speech recognition system based on temporal masking, Research abstract accepted to Acoustics 2008, Paris, France.

[8] K. J. Palomäki, G. J. Brown and J. Barker, Recognition of reverberant speech using full cepstral features and spectral missing data,*Proceedings the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tolouse, France, vol. 1, 289-292, 2006.*

[9] M.P. Cooke, P. Green, L. Josifovski, and A. Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data,*Speech Comm.*, vol. 34, pp. 267 285, 2001.

8.3 Language modeling

For Finnish, estimating the language model probabilities for words is difficult since there is a vast number of different word forms. For example, a single verb has theoretically thousands of inflected word forms. The natural way to attack the problem is to split words into smaller fragments and build the language models on the fragments instead of whole words. Since it is not obvious how the words should be split, we have studied what kind of word fragments are optimal for speech recognition systems. Experiments in Finnish, Turkish and Estonian recognition tasks indicate that an unsupervised data-driven splitting algorithm called Morfessor (see Section 10.1) improves recognition of rare words. [1]

N-gram models are the most widely used language models in large vocabulary continuous speech recognition. Since the size of the model grows rapidly with respect to the model order and available training data, many methods have been proposed for pruning the least relevant n-grams from the model. However, correct smoothing of the n-gram probability distributions is important and performance may degrade significantly if pruning conflicts with smoothing. In the journal paper [2] we show that some of the commonly used pruning methods do not take into account how removing an n-gram should modify the backoff distributions in the state-of-the-art Kneser-Ney smoothing. We also present two new algorithms: one for pruning Kneser-Ney smoothed models, and one for growing them incrementally. Experiments on Finnish and English text corpora show that the proposed pruning algorithm provides considerable improvements over previous pruning algorithms on Kneser-Ney smoothed models and is also better than the baseline entropy pruned Good-Turing smoothed models.

Representing the language model compactly is important in recognition systems targeted for small devices with limited memory resources. In [3], we have extended the compressed language model structure proposed earlier in the literature. By separating n-grams that are prefixes to longer n-grams, redundant information can be omitted. Experiments on English 4-gram models and Finnish 6-gram models show that extended structure can achieve up to 30 % lossless memory reductions when compared to the baseline structure.

Another common method for decreasing the size of the n-gram models is clustering of the model units. However, if size of the lexicon is very small, as in models based on statistical morpheme-like units (see, e.g., [1]), clustering of individual units is not so useful. Instead, we have studied how sequences of the morpheme-like units can be clustered to achieve improvements in speech recognition. When the clustered sequences are histories (context parts) of the n-grams, it is easy to combine the clustering to the incremental growing of the model applied in, e.g., [2]. Maximum a posteriori estimation can be used to make a compromise between the model size and accuracy. The experiments show that the clustering is useful especially if very compact models are required. [4]

References

[1] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, and A. Stolcke. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing*, 5(1), 2007.

[2] V. Siivola, T. Hirsimäki, and S. Virpioja. On Growing and Pruning Kneser-Ney Smoothed N-Gram Models. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5), pages 1617–1624, 2007.

[3] T. Hirsimäki. On Compressing N-gram Language Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, USA, pages IV-949–952, 2007.

[4] S. Virpioja and M. Kurimo. Compact N-gram Models by Incremental Growing and Clustering of Histories. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh PA, USA, pages 1037–1040, 2006.

8.4 Applications and tasks

Speech retrieval and indexing

Large amounts of information is produced in spoken form. In addition to TV and radio broadcasts, more and more material is distributed on the Internet in the form of podcasts and video sharing web sites. There is an increasing need for content based retrieval of this material. Speech retrieval systems consist of two parts as illustrated in Figure 8.4. First, an automatic speech recognition system is used to transcribe the speech into textual form. Second, an index is built based on this information.

The vocabulary of the speech recognizer limits the possible words that can be retrieved. Any word that is not in the vocabulary will not be recognized correctly and thus can not be used in retrieval. This is especially problematic since the rare words, such as proper names, that may not be in the vocabulary are often the most interesting from retrieval point of view. Our speech retrieval system addresses this problem by using morpheme-like units produced by the Morfessor algorithm. Any word in speech can now potentially be recognized by recognizing its component morphemes. The recognizer transcribes the text as a string of morpheme-like units and these units can also be used as index terms.

One problem of using morpheme-like units as index terms is that different inflected forms of the same word can produce different stems when they are split to morphemes. However, we would like to retrieve the speech document no matter what inflected form of the word is used. This resembles the problem of synonyms. We have countered this problem by applying Latent Semantic Indexing to the morpheme-based retrieval approach [1]. The method projects different stems of the same word to the same dimension that represents the true, latent, meaning of the term.

Speech recognizers typically produce only the most likely string of words, the 1-best hypothesis. Retrieval performance is decreased if a relevant term is misrecognized and is thus missing from the transcript. However, it is possible that the correct term was considered by the recognizer but was not the top choice. Thus, retrieval performance can be improved by extracting these alternative results from the recognizer and adding them to the index. A *confusion network* [2] provides a convenient representation of the competing terms along with a probability value for each term. However, as most terms in the network were in fact not spoken, the indexing method must be designed so that it is not degraded by these spurious terms. In [3], we compare methods that use the probability and rank of the terms to weigh the index terms properly and show improved performance of the retrieval system.

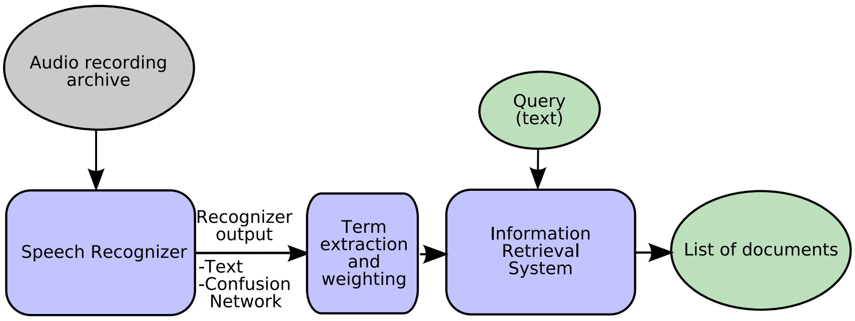


Figure 8.4: Overview of a spoken document retrieval system.

Estonian speech recognition

For agglutinative languages, like Finnish, Estonian, Hungarian and Turkish, it is practically impossible to build a word-based lexicon for speech recognition that would cover all the relevant words. The problem is that words are generally formed by concatenating several prefixes and suffixes to the word roots. Together with compounding and inflections this leads to millions of different, but still frequent word forms that can not be trivially split into meaningful parts. For some languages there exists rule-based morphological analyzers that can perform this splitting, but they are laborious to create and due to the handcrafted rules, they also suffer from an out-of-vocabulary problem.

In a pilot study of language and task portability of our speech recognition and language modeling tools, we created an Estonian speech recognizer. The text corpus used to learn the morph units and train the statistical language model consisted of newspapers and books, altogether about 55 million words [4]. The speech corpus consisted of over 200 hours and 1300 speakers, recorded from telephone [5], i.e. 8 kHz sampling rate and narrow band data instead of 16 kHz and normal (full) bandwidth that we have used for Finnish data. The speaker independence, together with the telephone quality and occasional background noises, made this task more difficult than our Finnish ones, but with the help of our learning and adaptive models we were still able to reach good recognition results and demonstrate a performance that was superior to the word-based reference systems [6, 7].

Speech-to-speech translation

Speech-to-speech machine translation is in some ways the peak of natural language processing, in that it deals directly with our (humans') original, oral mode of communication (as opposed to derived written language). As such, it presents several important challenges:

1. Automatic speech recognition of the input using state-of-the-art acoustic and language modeling, adaptation and decoding

2. Statistical machine translation of either the recognized most likely speech transcript or the confusion network or the whole lattice including all the best hypothesis

3. Speech synthesis to turn the translation output into intelligible speech using the state-of-the-art synthesis models and adaptation

4. Intergration of all these components to aim at the best possible output and tolerate errors that may happen in each phase

A pilot study of Finnish-English speech-to-speech translation was carried out in the lab as a joint effort of the speech recognition, Natural Language Processing 10 and Computational Cognitive Systems 10.3 groups. The domain selected for our experiments was heavily influenced by the available bilingual (Finnish and English) and bimodal (text and speech) data. Because none is readily yet available, we put one together using the Bible. As the first approach we utilized the existing components, and tried to weave them together in an optimal way. To recognize speech into word sequences we applied our morpheme-based unlimited vocabulary continuous speech recognizer [8]. As a Finnish acoustic model the system utilized multi-speaker hidden Markov models with Gaussian mixtures of mel-cepstral input features for state-tied cross-word triphones. The statistical language model was trained using our growing varigram model [9] with unsupervised morpheme-like units derived from Morfessor Baseline [10]. In addition to the Bible the training data included texts from various sources including newspapers, books and newswire stories totally about

150 million words. For translation, we trained the Moses system [11] on the same word and morpheme units as utilized in the language modeling units of our speech recognizer. For speech synthesis, we used Festival [12], including the built-in English voice and a Finnish voice developed at University of Helsinki.

References

[1] V. Turunen and M. Kurimo Using Latent Semantic Indexing for Morph-based Spoken Document Retrieval, *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh PA, USA, pages 389–392, 2006.

[2] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech And Language*, 14:373–400, 2000.

[3] V. Turunen and M. Kurimo Indexing Confusion Networks for Morph-based Spoken Document Retrieval, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in information retrieval*, Amsterdam, The Netherlands, pages 631–638, 2007.

[4] Segakorpus. 2005. Segakorpus - Mixed Corpus of Estonian. Tartu University. *http://test.cl.ut.ee/korpused/*.

[5] Einar Meister, Jürgen Lasn and Lya Meister 2002. Estonian SpeechDat: a project in progress. In *Proceedings of the Fonetiikan Päivät - Phonetics Symposium 2002 in Finland*, 21–26.

[6] Mikko Kurimo, Antti Puurula, Ebru Arisoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumae and Murat Saraclar 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics.* HLT-NAACL 2006. New York, USA

[7] Antti Puurula and Mikko Kurimo 2007. Vocabulary Decomposition for Estonian Open Vocabulary Speech Recognition. In *Proceedings of ACL 2007.*

[8] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja and Janne Pylkkönen 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language* 20(4):515–541.

[9] Vesa Siivola Language models for automatic speech recognition: construction and complexity control. Doctoral thesis, Dissertations in Computer and Information Science, Report D21, Helsinki University of Technology, Espoo, Finland, 2006.

[10] Mathias Creutz. Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition. Doctoral thesis, Dissertations in Computer and Information Science, Report D13, Helsinki University of Technology, Espoo, Finland, 2006.

[11] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar, Alexandra Constantin, and Evan Herb. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, 2007.

[12] The Festival Speech Synthesis System. University of Edinburgh. *http://festvox.org*

Chapter 9

Proactive information retrieval

Samuel Kaski, Kai Puolamäki, Antti Ajanki, Jarkko Salojärvi

9.1 Introduction

Successful proactivity, that is anticipation, in varying contexts requires generalization from past experience. Generalization, on its part, requires suitable powerful (stochastic) models and a collection of data about relevant past history to learn the models.

The goal of the PROACT project is to build statistical machine learning models that learn from the actions of people to model their intentions and actions. The models are used for disambiguating the users' vague commands and anticipating their actions.

Our application area is information retrieval, where we investigate to what extent the laborious explicit relevance feedback can be complemented or even replaced by implicit feedback derived from patterns of eye fixations and movements that exhibit both voluntary and involuntary signs of users intentions. Inference is supported by models of document collections and interest patterns of users.

The PROACT project has been done in close collaboration with researchers in the European Union's Pascal Network of Excellence within a Pump Priming Programme (2005–2007); the collaborators are from University of Helsinki, University of Southampton and University College London. The project continues in a STREP project PinView from 2008 onwards.

9.2 Implicit queries from eye movements

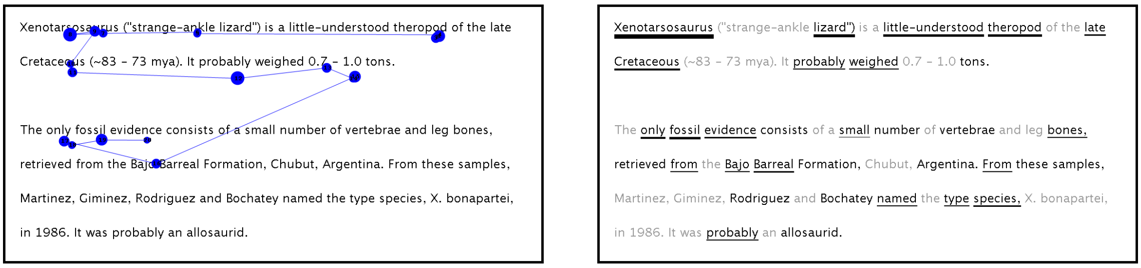
Eye movements measured during reading are a promising new source of implicit feedback. During complex tasks such as reading, attention approximately lies on the location of the reader's gaze. Therefore the eye movements should contain information on the reader's interests. Inferring interest of the user from a reading pattern is difficult however, since the signal is complex and very noisy, and since interestingness or relevance is higly subjective and thus hard to define. We have earlier developed machine learning and signal processing methods for this task, and hosted a research challenge where the task was to predict relevance from eye movement patterns [1].

The motivation for the next stage of the research was that formulating a good query in a web search engine, for example, is known to be difficult. Implicit feedback collected by observing user's behavior might reveal the true interest of the user without the need to explicitely label the documents as relevant or not relevant. We performed a feasibility study in which we used eye movements to formulate a query that reflects user's interest while he was reading [2].

We constructed a controlled experimental setting in which it is known which documents are relevant. The users read short documents searching for the ones related to a topic that was given to them beforehand. The eye movements were recorded during reading. We trained a regressor that predicts how relevant a term is for user's current query given the eye movement measurements on that term. This regressor can then be applied to new topics, with no training data available, to estimate relevance of words.

The learned model was then used to infer relevant query terms based on eye movement recorded while the user was performing a new search task, where the true query was unknown. The inferred query terms can be used to retrieve and suggest new documents that might be important to user's information need.

A SVM model that uses eye movements and textual features outperformed a similar model without the eye movement features. This indicates that eye movements contain exploitable information about relevance in information retrieval tasks.



Xenotarsosaurus ("strange-ankle lizard") is a little-understood theropod of the late Cretaceous (~83 - 73 mya). It probably weighed 0.7 - 1.0 tons.

The only fossil evidence consists of a small number of vertebrae and leg bones, retrieved from the Bajo Barreal Formation, Chubut, Argentina. From these samples, Martinez, Giminez, Rodriguez and Bochatey named the type species, X. bonapartei, in 1986. It was probably an allosaurid.

Xenotarsosaurus ("strange-ankle lizard") is a little-understood theropod of the late Cretaceous (~83 - 73 mya). It probably weighed 0.7 - 1.0 tons.

The only fossil evidence consists of a small number of vertebrae and leg bones, retrieved from the Bajo Barreal Formation, Chubut, Argentina. From these samples, Martinez, Giminez, Rodriguez and Bochatey named the type species, X. bonapartei, in 1986. It was probably an allosaurid.

Figure 9.1: Left: A sample eye movement pattern of a test subject during reading a document. Right: The term weights depicted on the same document inferred from the eye movements of all test subjects who were searching for information about dinosaurs. The magnitudes of the weights are depicted as the thickness of the underlining.

References

[1] Kai Puolamäki and Samuel Kaski, editors. *Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling.* Helsinki University of Technology, Espoo, Finland, 2006.

[2] David R. Hardoon, John Shawe-Taylor, Antti Ajanki, Kai Puolamäki, and Samuel Kaski. Information Retrieval by Inferring Implicit Queries from Eye Movements. In *Proceedings of the 11th International Conference on International Conference on Artificial Intelligence and Statistics*. San Juan, Puerto Rico, 2007.

Chapter 10

Natural language processing

Krista Lagus, Mikko Kurimo, Timo Honkela, Mathias Creutz, Jaakko J. Väyrynen, Sami Virpioja, Ville Turunen, Matti Varjokallio

10.1 Unsupervised segmentation of words into morphs

In the theory of linguistic morphology, morphemes are considered to be the smallest meaning-bearing elements of language, and they can be defined in a language-independent manner. It seems that even approximative automated morphological analysis is beneficial for many natural language applications dealing with large vocabularies, such as speech recognition and machine translation. These applications usually make use of *words* as vocabulary units. However, for highly-inflecting and agglutinative languages, this leads to very sparse data, as the number of possible word forms is very high. Figure 10.2 shows the very different rates at which the vocabulary grows in various text corpora of the same size. The number of different unique word forms in the Finnish corpus is considerably higher than in the English ones, for example.

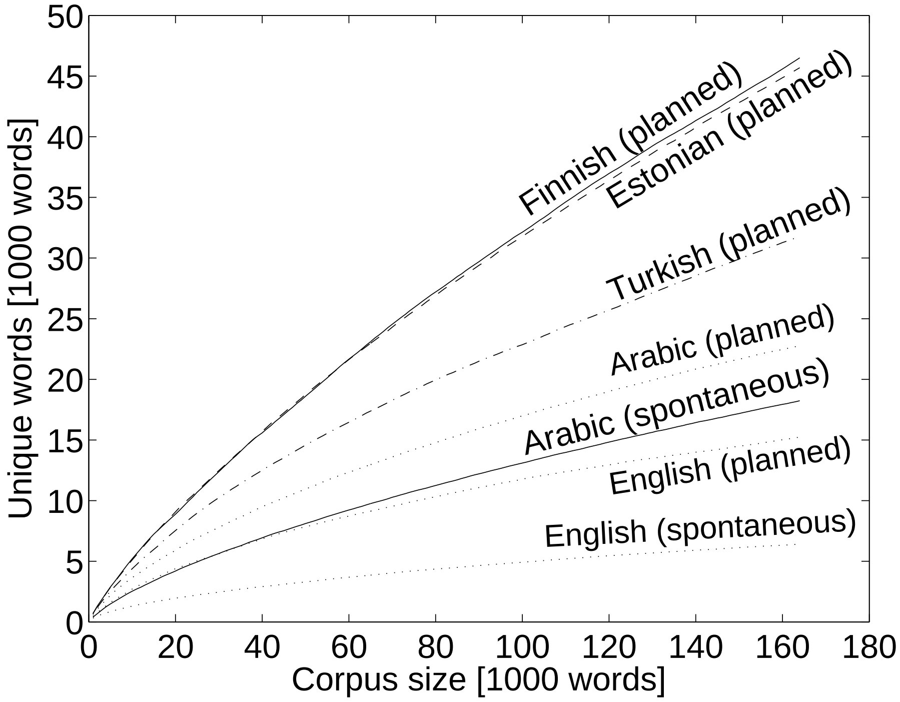


Figure 10.2: The number of different word forms (types) encountered in growing portions of running text (tokens) of various languages.

We have developed *Morfessor*, a language-independent, data-driven method for the unsupervised segmentation of words into morpheme-like units. There are different versions of Morfessor, which correspond to consecutive steps in the development of the model [1, 2, 3, 4]. All versions can be seen as instances of a general model, as described in [5].

The general idea behind the Morfessor model is to discover as compact a description of the data as possible. Substrings occurring frequently enough in several different word forms are proposed as *morphs* and the words are then represented as a concatenation of morphs, e.g., "hand, hand+s, left+hand+ed, hand+ful".

An optimal balance is sought between compactness of the *morph lexicon* versus the compactness of the representation of the *corpus*. The morph lexicon is a list of all distinct morphs (e.g., "hand, s, left, ed, ful") together with some stored properties of these morphs. The representation of the corpus can be seen as a sequence of pointers to entries in the morph lexicon; e.g. the word "lefthanded" is represented as three pointers to morphs in the lexicon.

Among others, de Marcken [6], Brent [7], and Goldsmith [8] have shown that the above type of model produces segmentations that resemble linguistic morpheme segmentations, when formulated mathematically in a probabilistic framework or equivalently using the Minimum Description Length (MDL) principle [9].

A shortcoming of previous splitting methods is that they either do not model *context-dependency* or they *limit the number of splits* per word to two or three. Failure to incorporate context-dependency in the model may produce splits like "s+wing, ed+ward, s+urge+on" on English data, since the morphs "-s" and "-ed" are frequently occurring suffixes in the English language, but the algorithm does not make this distinction and thus suggests them in word-initial position as prefixes. By limiting the number of allowed segments per word the search task is alleviated and context-dependency can be modeled. However, this makes it impossible to correctly segment compound words with several affixes (pre- or suffixes), such as the Finnish word "aka+n+kanto+kiso+i+ssa" (transl. "in the wife-carrying contests").

We have focused our efforts on developing a segmentation model that incorporates context-dependency without restricting the number of allowed segments per word. This has resulted in two model variants, Categories-ML [3] and Categories-MAP [4]. The former is based on Maximum Likelihood (ML) optimization, in combination with some heuristics, whereas the latter applies a more elegant model formulation within the Maximum a Posteriori (MAP) framework. The MAP formulation, along with a thorough comparison to the other Morfessor variants, is provided also in [5] and [10].

Some sample segmentations of Finnish, English, as well as Swedish words, are shown in Figure 10.3. These include correctly segmented words, where each boundary coincides with a linguistic morpheme boundary (e.g., "aarre+kammio+i+ssa, edes+autta+isi+vat, abandon+ed, long+fellow+'s, in+lopp+et+s"). In addition, some words are over-segmented, with boundaries inserted at incorrect locations (e.g., "in+lägg+n+ing+ar" instead of "in+lägg+ning+ar"), as well as under-segmented words, where some boundary is missing (e.g., "bahama+saari+lla" instead of "bahama+saar+i+lla").

In addition to segmenting words, Morfessor suggests likely grammatical categories for the segments. Each morph is tagged as a prefix, stem, or suffix. Sometimes the morph categories can resolve the semantic ambiguity of a morph, e.g., Finnish "pää". In Figure 10.3, "pää" has been tagged as a stem in the word "pää+hän" ("in [the] *head*"), whereas it functions as a prefix in "pää+aihe+e+sta" ("about [the] *main* topic").

Evaluation

In the publications related to the development of Morfessor, the algorithm has been evaluated by comparing the results to linguistic morpheme segmentations of Finnish and English words [1, 2, 3, 4, 5]. In order to carry out the evaluation, linguistic reference segmentations needed to be produced as part of the project, since no available resources were applicable as such. This work resulted in a morphological "gold standard", called *Hutmegs* (Helsinki University of Technology Morphological Evaluation Gold Standard) [11, 12]. When the latest context-sensitive Morfessor versions [3, 4] are evaluated against the Hutmegs gold standard, they clearly outperform a frequently used benchmark algorithm [8] on Finnish data, and perform as well or better than the benchmark on English data.

Morfessor algorithms have also been evaluated in the Morpho Challenge competitions described in Section 10.2. Morpho Challenge 2007 included evaluation in four languages (English, Finnish, German and Turkish) and two competitions: comparison against linguistic standards and evaluation in information retrieval tasks. Morfessor managed fairly

aarre + **kammio** + *i* + *ssa*, **aarre** + **kammio** + *nsa*, **bahama** + **saar** + *et*,
bahama + **saari** + *lla*, **bahama** + **saar** + *ten*, **edes** + **autta** + *isi* + *vat*,
edes + **autta** + *ma* + *ssa*, nais + **auto** + *ili* + *ja* + *a*, pää + **aihe** + *e* + *sta*,
pää + **aihe** + *i* + *sta*, **pää** + *hän*, taka + **penkki** + *lä* + *in*+ *en*, **voi** + *mme* + *ko*

abandon + *ed*, **abandon** + *ing*, **abandon** + *ment*, **beauti** + *ful*,
beauty + *'s*, **calculat** + *ed*, **calculat** + *ion* + *s*, **express** + *ion* + *ist*,
micro + **organ** + *ism* + *s*, **long** + **fellow** + *'s*, **master** + **piece** + *s*,
near + *ly*, **photograph** + *er* + *s*, **phrase** + *d*, un + **expect** + *ed* + *ly*

ansvar + *ade*, **ansvar** + *ig*, **ansvar** + *iga*, **ansvar** + *s* + för + **säkring** + *ar*,
blixt + ned + **slag**, **dröm** + *de*, **dröm** + *des*, **drömma** + *nde*, in + **lopp** + *et* + *s*,
in + **lägg** + *n* + *ing* + *ar*, **målar** + *e*, **målar** + **yrke** + *t* + *s*, o + ut + **nyttja** + *t*,
poli + *s* + **förening** + *ar* + *na* + *s*, **trafik** + **säker** + *het*, över + **fyll** + *d* + *a*

Figure 10.3: Examples of segmentations learned from data sets of Finnish, English, and Swedish text. Suggested prefixes are underlined, stems are rendered in **boldface**, and suffixes are *slanted*.

well in all the evaluations, especially with Finnish and Turkish languages.

Applications

Morfessor has been extensively tested as a component of a large vocabulary speech recognition system. By allowing a compact but flexible vocabulary for the system, Morfessor improves especially recognition of rare words. For several languages such as Finnish, Estonian and Turkish, this approach outperforms the state-of-the-art solutions. The speech recognition experiments are described in Section 8.3.

In addition to speech recognition, Morfessor has been used in speech retrieval and statistical machine translation systems. These experiments are described in Section 8.4 and 13, respectively.

Demonstration and software

There is an online demonstration of Morfessor on the Internet: `http://www.cis.hut.fi/projects/morpho/`. Currently, the demo supports three languages (Finnish, English, and Swedish) and two versions of the Morfessor (Baseline and Categories-ML). Those interested in larger-scale experiments can download the Morfessor program and train models using their own data sets. Two versions are available: Morfessor 1.0 software implements the Morfessor Baseline algorithm described in [13] and Morfessor Categories-MAP 0.9.2 software implements the Morfessor Categories-MAP algorithm described in [4]. During 2007, a monthly average of 10 downloads has been registered for both versions.

References

[1] Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. In *Proc. Workshop on Morphological and Phonological Learning of ACL'02*, pages 21–30, Philadelphia, Pennsylvania, USA, 2002.

[2] Mathias Creutz. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proc. ACL'03*, pages 280–287, Sapporo, Japan, 2003.

[3] Mathias Creutz and Krista Lagus. Induction of a simple morphology for highly-inflecting languages. In *Proc. 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona, July 2004.

[4] Mathias Creutz and Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, 2005.

[5] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, January 2007.

[6] C. G. de Marcken. *Unsupervised Language Acquisition.* PhD thesis, MIT, 1996.

[7] M. R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105, 1999.

[8] John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

[9] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific Series in Computer Science, Singapore, 1989.

[10] Mathias Creutz. Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition. Doctoral thesis, Dissertations in Computer and Information Science, Report D13, Helsinki University of Technology, Espoo, Finland, 2006.

[11] Mathias Creutz and Krister Lindén. Morpheme segmentation gold standards for Finnish and English. Technical Report A77, Publications in Computer and Information Science, Helsinki University of Technology, 2004.

[12] Mathias Creutz, Krista Lagus, Krister Lindén, and Sami Virpioja. Morfessor and Hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, pages 107–112, Tallinn, Estonia, 4 – 5 April 2005.

[13] Mathias Creutz and Krista Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology, 2005.

10.2 Morpho Challenge

Morpho Challenge is a series of scientific competition organized by Adaptive Informatics Research Centre for an evaluation of unsupervised morpheme analysis algorithms. The challenge is part of the EU Network of Excellence PASCAL Challenge Program and in 2007 organized in collaboration with Cross-Language Evaluation Forum CLEF. The objective of the challenge is to design statistical machine learning algorithms that discover which morphemes (smallest individually meaningful units of language) words consist of. Ideally, these are basic vocabulary units suitable for different tasks, such as text understanding, machine translation, information retrieval, and statistical language modeling. The challenge has sofar been organized two times: the results of the 2005 challenge were published in a workshop in April 2006 in Venice, Italy [1]. The 2007 challenge workshop was held in September 2007 in Budapest, Hungary [2, 3].

In the original challenge, the words were segmented in unsupervised morphemes and the results were evaluated by a comparison to linguistic gold standard morphemes. The organizers also used the results to for training statistical language models and evaluated the models in large vocabulary speech recognition experiments [1]. The 2007 challenge was a more difficult one requiring morpheme analysis of words instead of just segmentations into smaller units. The evaluation of the submissions was performed by two complementary ways: *Competition 1*: The proposed morpheme analyses were compared to a linguistic morpheme analysis gold standard by matching the morpheme sharing word pairs [2]. *Competition 2*: Information retrieval (IR) experiments were performed, where the words in the documents and queries were replaced by their proposed morpheme representations and the search was based on morphemes instead of words [3]. The IR evaluations were provided for Finnish, German, and English and participants were encouraged to apply their algorithm to all of them. The organizers performed the IR experiments using the queries, texts, and relevance judgments available in CLEF forum and morpheme analysis methods submitted by the challenge participants. The results show that the morpheme analysis has a significant effect in IR performance in all languages, and that the performance of the best unsupervised methods can be superior to the supervised reference methods.

References

[1] Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy and Murat Saraclar. Unsupervised segmentation of words into morphemes - Challenge 2005, An Introduction and Evaluation Report. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes.* Venice, Italy, April 12, 2006.

[2] Mikko Kurimo, Mathias Creutz, Matti Varjokallio. Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard – Morpho Challenge 2007. In *Working Notes of the CLEF 2007 Workshop.* Edited by Alessandro Nardi and Carol Peters. 19-21 September, Budapest, Hungary.

[3] Mikko Kurimo, Mathias Creutz, Ville Turunen. Unsupervised Morpheme Analysis Evaluation by IR experiments – Morpho Challenge 2007. In *Working Notes of the CLEF 2007 Workshop.* Edited by Alessandro Nardi and Carol Peters. 19-21 September, Budapest, Hungary.

10.3 Emergence of linguistic features using independent component analysis

We have been able to show that Independent Component Analysis (ICA) [1] applied on word context data provides distinct features that reflect syntactic and semantic categories [2]. The difference to latent semantic analysis (LSA) is that the analysis finds features or categories that are not only explicit but can also easily be interpreted by humans. This result can be obtained without any human supervision or tagged corpora that would have some predetermined morphological, syntactic or semantic information.

It is important to compare the capability of single features or feature pairs to separate categories because this measures how well the obtained features correspond with the categories. In fact, when all features are used, the separation capabilities of ICA and LSA are comparable because the total information present is the same. We have also shown that the emergent features match well with categories determined by linguists by comparing the ICA results to linguistic word category information [3].

We have shown how the features found by the ICA method can be further processed by simple nonlinear methods, such as thresholding, that gives rise to a sparse feature representation of words [4, 5]. We performed thresholding for each found word feature vector separately. The values closest to zero were set to zero and only a selected number of features were left to their original values. An analogical approach can be found from the analysis of natural images, where a soft thresholding of sparse coding is a denoising operator.

We compared the original representation and the thresholded representations in multiple choice vocabulary tasks, which measure the semantic information captured by the representation. An illustrative result is shown in Figure 10.4, which compares the feature thresholding with the two methods, latent semantic analysis and independent component analysis. The graph shows that the thresholded ICA representation is able to capture the most important semantics with fewer components, as the quality of the thresholded ICA representation degrades more slowly than both LSA representations. Several tests were run with three languages, including two different corpora, with quite similar results.

We have also shown how independent component analysis gives rise to a multilingual word feature space when trained with a parallel corpus [6]. The feature space created by the found features is also multilingual. Words that are related in different languages appear close to each other in the feature space, which makes it possible to find translations for words between languages. Table 10.1 shows the closest words for the English word 'finland' in the feature space, which include different forms of the Finnish equivalent, but also the name of a neighboring country ('sweden') as well as Austria ('itävalta'). The latter might be caused by shared work during the Finnish EU presidency. The single features also carry multilingual semantic information, as can be seen from Table 10.2, that lists the most prominent words in three features.

The attained results include both an emergence of clear distinctive categories or features and a distributed representation. In the emergent representation, a word may thus belong to several categories simultaneously in a graded manner. We see that further processing of the features is possible and thresholding produces a more sparse representation that can have greater interpretability without too much information loss. The method is also applicable to multilingual textual data, and is able to find representations where the multilingual semantic space can be used to mine translations and related words.

We wish that our model provides additional understanding on potential cognitive mechanisms in natural language learning and understanding Our approach attempts to show that it is possible that much of the linguistic knowledge is emergent in nature and based

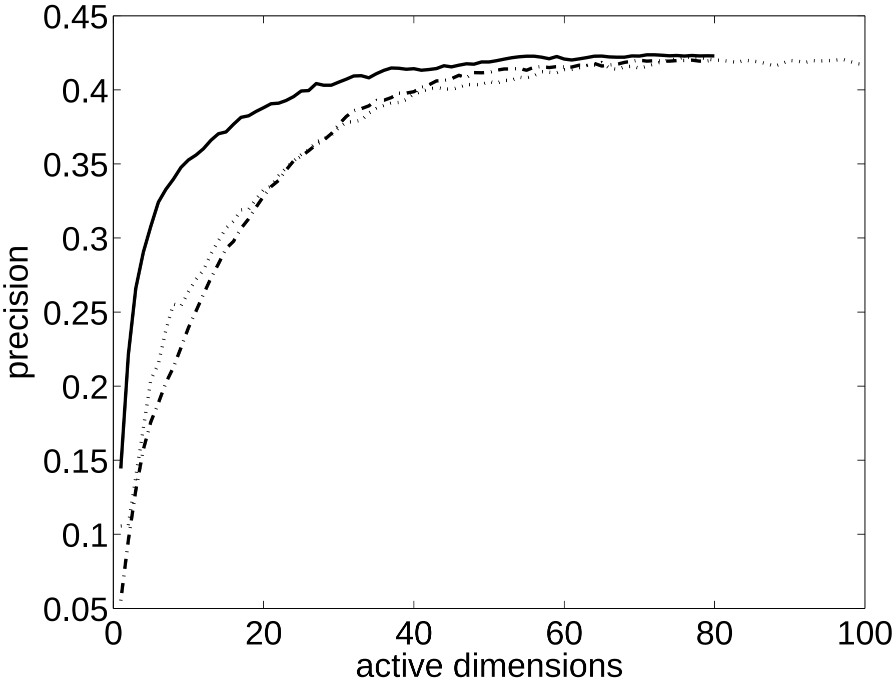


Figure 10.4: Rates of correctly answered questions with unthresholded LSA (dotted), LSA with thresholding with 80 components (dashed) and ICA with thresholding with 80 components (solid) set w.r.t. the number of non-zero features (after thresholding). The features where calculated from free electronic English books extracted from the Gutenberg project. The test questions were based on synonyms and related words extracted from the Moby thesaurus.

Table 10.1: The closest words in the multilingual feature space to the word 'finland'.

word	match
finland	1.00
suomen	0.83
suomi	0.82
sweden	0.79
suomessa	0.77
austria	0.73
...	...

on specific learning mechanisms.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis.* John Wiley & Sons, 2001.

[2] T. Honkela, and A. Hyvärinen. Linguistic feature extraction using independent component analysis. In *Proceedings of IJCNN 2004, International Joint Conference on*

Table 10.2: Most prominent words for three example features (columns) that list clearly related words in both languages.

saksan	values	eroja
ranskan	rauhan	different
germany	demokratian	difference
france	vapauden	välillä
french	democracy	erilaista
german	ihmisoikeuksien	differences
sweden	arvoja	erot
netherlands	solidarity	toisiaan
ranska	peace	disparities
belgian	arvojen	eri
ruotsin	kunnioittaminen	erilaiset
saksa	oikeusvaltion	differ
italian	principles	differing
kingdom	continent	eroavat
...	...	...

Neural Networks, Budapest, Hungary, 25–29 Jul 2004, pp. 279–284.

[3] J.J. Väyrynen, T. Honkela, and A. Hyvärinen. Independent component analysis of word contexts and comparison with traditional categories. In: Jarmo M. A. Tanskanen (ed.), *Proceedings of NORSIG 2004, Sixth Nordic Signal Processing Symposium*, Espoo, Finland, 9–11 Jun 2004, pp. 300–303.

[4] J. J. Väyrynen, L. Lindqvist and T. Honkela. Sparse distributed representations for words with thresholded independent component analysis. In *Proceedings of IJCNN 2007, International Joint Conference on Neural Networks*, Orlando, Florida, 12–17 Aug 2007, pp. 1031–1036.

[5] J. J. Väyrynen, T. Honkela and L. Lindqvist. Towards explicit semantic features using independent component analysis. In: M. Sahlgren and O. Knuttson (eds.), *Proceedings of SCAR 2007 Workshop, Semantic Content Acquisition and Representation*, SICS Technical Report T2007-06, Swedish Institute of Computer Science, Stockholm, Sweden, ISSN 1100-3154, Tartu, Estonia, 24 May 2007, pp. 20–27.

[6] J. J. Väyrynen and T. Lindh-Knuutila. Emergence of multilingual representations by independent component analysis using parallel corpora. In *Proceedings of SCAI 2006, Ninth Scandinavian Conference on Artificial Intelligence*, Espoo, Finland, 25–27 Oct 2006, pp. 101–105.