Inferring Relevance from Eye Movements: Feature Extraction

Jarkko Salojärvi[†], Kai Puolamäki[†], Jaana Simola[§], Lauri Kovanen[†]
Ilpo Kojo[§], Samuel Kaski[‡,†]

[†] Laboratory of Computer and Information Science
Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, FI-02015 HUT, Finland

[§] Center for Knowledge and Innovations Research
Helsinki School of Economics
Tammasaarenkatu 3, FI-00180 Helsinki, Finland

[‡] Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland

3 March 2005

Abstract

We organize a PASCAL EU Network of Excellence challenge for inferring relevance from eye movements, beginning 1 March 2005. The aim of this paper is to provide background material for the competitors: give references to related articles on eye movement modelling, describe the methods used for extracting the features used in the challenge, provide results of basic reference methods and to discuss open questions in the field.

1 Introduction

This technical report is written to complement the Inferring Relevance from Eye Movements challenge[1], one of the challenges partly funded by the EU network of excellence PASCAL. The challenge is organized in the form of a competition, where the contestants try to infer the relevance of a read document from the associated eye movement trajectory. We expect that the challenge will bring contributions to four different areas:

- Advances in machine learning methodology
- Establishing common practices for feature extraction in eye movements

[1] The Challenge has a web site at `http://www.cis.hut.fi/eyechallenge2005/`.

- Further the development of proactive user interfaces
- To learn of the psychology underlying eye movements in search tasks

The eye movement data is promising for advancing machine learning methods since it is very rich but noisy, and it is rather easy to collect in large quantities. The data is in the form of a time series which will pose challenges for optimal selection of features. For a simple case (Competition number 1), we will provide a comprehensive 22-dimensional set of eye movement features derived from the ones generally used in eye movement research (previously analysed in [31, 32]).

In psychological research of reading, it is common to segment the eye movement trajectory into fixations and saccades, and then compute summary measures of these modalities. The features used in Competition 1 are such summary measures. The controlled experimental setup used in the challenge makes it possible to test whether the established common practice is optimal for inferring relevance. In Competition 2 we give the full eye movement trajectory and the competitors can model it in any unorthodox way.

In information retrieval, relevance generally depends on the context, task, and individual competence and preferences of the user. Therefore relevance of articles suggested by a search engine could be improved by filtering them through an algorithm which models the interests of the user. This algorithm would be *proactive* [41]; it predicts the needs of the user and adapts its own behavior accordingly. Individual relevance can be learned from feedback given by the user. The usual way would be to ask after every document whether the user found it relevant, and to learn the user's preferences from the answers. However, giving this kind of explicit feedback is laborious, and people outside of research laboratories rarely bother. Alternatively, relevance can be inferred from implicit feedback derived traditionally from document reading time, or by monitoring other behavior of the user (such as saving, printing, or selecting of documents). The problem with the traditional sources is that the number of feedback events is relatively small. One of the motivations of the PASCAL challenge is to explore whether the traditional sources of implicit relevance information could be complemented with eye movements, and to find best methods for doing it.

In a typical information retrieval setup the user types in keywords to a search engine and is then given a list of titles of documents that possibly contain the information the user is looking for. Some of the documents suggested by the search engine will be totally irrelevant, some will handle the correct topic, and only few will be links to documents that the user actually will bother to read. Our experimental setting for collecting eye movement data was designed to simulate this natural situation, with the difference that in our case the relevance is known. By gathering data in a controlled setup we ensure that we know the ground truth, that is, the relevance associated with each eye movement trajectory. Machine learning methods can then be used for selecting a good set of features of eye movements, and for learning time series models to predict relevance of new measurements. If the eye movements contain any information about the relevance of a text, prediction should be possible. The modeling assumption behind our analysis is that attention patterns correlate with relevance; at the simplest, people tend to pay more attention to objects they find relevant or interesting.

2 Physiology of the Eye

Gaze direction is a good indicator of the focus of attention, since accurate viewing is possible only in the central *fovea* area (only 1–2 degrees of visual angle) where the density of photoreceptive cells is highly concentrated. For this reason, detailed inspection of a scene is carried out in a sequence of *saccades* (rapid eye movements) and *fixations* (the eye is fairly motionless). The trajectory is often referred to as a *scanpath*.

Information on the environment is mostly gathered during fixations, and the duration of a fixation is correlated with the complexity of the object under inspection. A simple physiological reason for this is that the amount of information the visual system is capable of processing is limited. During reading this complexity is associated with the frequency of occurrence of the words in general, and with how predictable the word is based on its context [29]. Naturally there are other factors affecting the reading pattern as well, such as different reading strategies and the mental state of the reader.

2.1 Eye movement details

Actually the eye does not lie completely still during fixations. In general we expect that the small movements during fixations will not play an important role in this challenge, since with the sampling rate of 50 Hz the average amount of samples from a fixation is around twelve. However, some basic knowledge on the fixations and saccades will be required if the competitors want to construct algorithms for fixation identification for Competition 2.

Clinical physiology text books [17] report that during fixation, the eye moves in an area which usually is less than 0.25 degrees of visual angle, meaning of the order of ten pixels in our experiment[2] (one should however also remember to take into account the measurement noise). During fixation, three different modes of movement can be separated: *tremor*, which is small amplitude (5–30 sec arc) and high frequency (30–100 Hz) oscillations, *drift*, which is slow velocity movement (1–8 min arc per second) and low frequency (<0.5 Hz), and *microsaccades*, low frequency (1–2 Hz) and small amplitude (1–8 min arc), saccade-like movements. Tremor and drift are commonly associated with the physiology of the eye, microsaccades on the other hand seem to have some cognitive basis [5, 19].

The saccades are ballistic, meaning that the target of the saccade will be decided before its initiation. The speed during a saccade depends on its length; for example during 5° saccade the peak velocity is around 260° per second, while during 20° saccade the peak velocity is around 660° per second. These characteristics are common to all people to the extent that one can use quantitative measurements of saccades to assess the function of the oculomotor system, to investigate the effects of drugs or lesions, and in some cases to aid diagnosis of disease or locating of lesions (see [10], for example).

The computation of a saccade requires some (latency) time in the fixation, meaning that fixations under 60 ms are not generally possible. However, it is possible to pre-program a sequence of saccades where the fixation duration will be shorter.

[2] with a subject distance of 60 cm from the 17" screen with a resolution of 1024x1280.

2.2 Pupillometry

In addition to eye movement features, the challenge also contains features computed from the pupil. There was some evidence in our experiments that the features correlated with relevance of the text [31]; the effect was very small at best, but it led us to discover the works reported in [16] or [2], where pupil diameter has been reported to increase as a sign of increased cognitive load.

The main function of pupil is to control the amount of light falling onto the retina. However, in addition to reflexive control of pupillary size there also seem to be tiny, cognitively related fluctuations of pupillary diameter ([2] reports interesting results that are discussed below). The so called *task-evoked pupillary response* (TERP) amplitudes appear to provide a good measure of the cognitive demands [2] for a wide variety of tasks (see Appendix for a brief note on TERPs).

Besides being a measure of cognitive demands of the task, the pupil width is also reported to vary due to different emotions. In [25], affective stimuli has been reported to cause systematical effects in subjects' physiological reactions and subjective experiences. The pupil size variation could therefore be used as implicit feedback signal for example in an affective computing interface [25].

3 Some literature

In this Section we give a brief introduction to literature on eye movements. The emphasis is on the areas which are relevant to the challenge: eye movements during reading and eye movements used as an implicit feedback channel.

3.1 Eye movements and reading

In a typical reading situation, the reader fixates on each word sequentially. Some of the words are skipped, some fixated twice and some trigger a *regression* to preceding words (approx. 15 % of the saccades). The reader is often not conscious of these regressions. The typical duration of fixations varies between 60–500 ms, being 250 ms on the average [21].

Research on eye movements during reading is a well-established field (see [29] for a good overview). In psychological literature, several models for reading have been proposed (most recent [6, 20, 30]). Models of eye movement control during reading differ mainly by the extent to which eye movements are assumed to be governed by lexical (high-level) processes over a simple default (low-level) control system assuming certain mean saccade lengths and fixation durations [39].

Currently the most popular model, so called E-Z Reader [30], concentrates on modeling reading at the basic level, as a series of sequential fixations occurring from left to right without regressions which are assumed to be associated with higher order cognitive processes. The durations of the fixations are correlated with word occurrence frequency, that is, the access time for the concepts concerning more rarely occurring words is longer than the access time for more frequently occurring words (however, similar correlations with word predictability and word length have also been reported). In a more recent publication [6] this correlation is extended to explain also regressions as occurring to those words which did not receive enough processing time during the first pass reading.

3.2 Eye movements and implicit feedback

Eye movements have earlier been utilized as alternative input devices for either pointing at icons or typing text in human-computer interfaces (see [15, 44]).

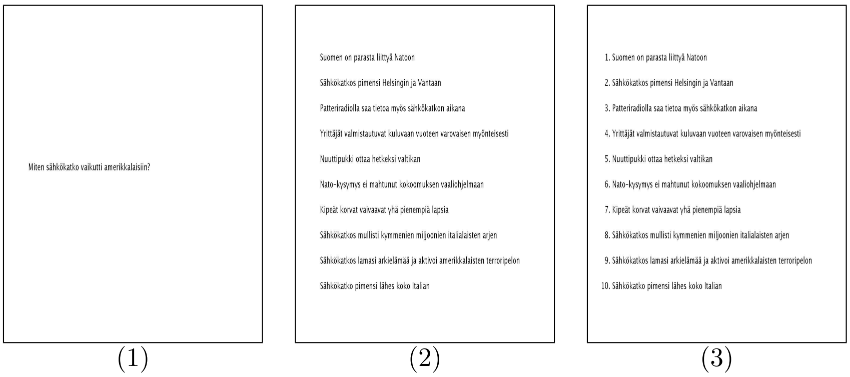
Use of eye movements as a source of implicit feedback is a relatively new concept. The first application where user interest was inferred from eye movements was an interactive story teller [38]. The story told by the application concentrated more on items that the user was gazing at on a display. Rudimentary relevance determination is needed also in [13], where a proactive translator is activated if the reader encounters a word which she has difficulties (these are inferred from eye movements) in understanding. A prototype attentive agent application (Simple User Interest Tracker, Suitor) is introduced in [22, 23]. The application monitors eye movements during browsing of web pages in order to determine whether the user is reading or just browsing. If reading is detected, the document is defined relevant, and more information on the topic is sought and displayed. Regretfully the performance of the application was not evaluated in the papers in any way. The (heuristic) rules for inferring whether the user is reading are presented in [4]. The eye movements have also been used as one feedback channel to identify critical driving events in intelligent driver assistance systems [24, 42].

The first analysis of eye movements in an information retrieval situation was published in [31, 32], where the experimental setup is quite similar to the Challenge. In [8] the goal was different: to investigate with quantitative measures how users behave in a real, less-controlled information retrieval task.

Implicit feedback information is also evaluated in usability studies[14, 7], where it is common to compute summary measures of eye movements on large areas of interest, such as images or captions of text (see [27] for an example study). The eye movements have also been used to give feedback of the subjective image quality [43].

4 Measurements

4.1 Experimental setup



Miten sähkökatko vaikutti amerikkalaisiin?

(1)

Suomen on parasta liittyä Natoon
Sähkökatkos pimensi Helsingin ja Vantaan
Patteriradiolla saa tietoa myös sähkökatkon aikana
Yrittäjät valmistautuvat kuluvaan vuoteen varovaisen myönteisesti
Nuuttipukki ottaa hetkeksi valtikan
Nato-kysymys ei mahtunut kokoomuksen vaaliohjelmaan
Kipeät korvat vaivaavat yhä pienempiä lapsia
Sähkökatkos mullisti kymmenien miljoonien italialaisten arjen
Sähkökatkos lamasi arkielämää ja aktivoi amerikkalaisten terroripelon
Sähkökatko pimensi lähes koko Italian

(2)

1. Suomen on parasta liittyä Natoon
2. Sähkökatkos pimensi Helsingin ja Vantaan
3. Patteriradiolla saa tietoa myös sähkökatkon aikana
4. Yrittäjät valmistautuvat kuluvaan vuoteen varovaisen myönteisesti
5. Nuuttipukki ottaa hetkeksi valtikan
6. Nato-kysymys ei mahtunut kokoomuksen vaaliohjelmaan
7. Kipeät korvat vaivaavat yhä pienempiä lapsia
8. Sähkökatkos mullisti kymmenien miljoonien italialaisten arjen
9. Sähkökatkos lamasi arkielämää ja aktivoi amerikkalaisten terroripelon
10. Sähkökatko pimensi lähes koko Italian

(3)

Figure 1: An example of stimuli used in the experiments.

The structure of an assignment is as follows: a subject was first shown a question (image 1 in Figure 1), and then a list of ten sentences (image 2 in Figure 1), one of which contained the correct answer (C). Five of the sentences were known to be irrelevant (I), and four relevant for the question (R). The subject was instructed to identify the correct answer and then press 'enter' (which ended the eye movement measurement) and then type in the associated number in the next screen (image 3 in Figure 1). The assignments were in Finnish, the mother tongue of the subjects.

The measurements were made for 11 subjects.

The *full training set* consists of 50 assignments, shown to all subjects. The lists were presented to the subjects in a randomized order. The measurements were carried out in sets of ten assignments, followed by a short break and recalibration. Some of the assignments were excluded for technical reasons (e.g. the subject gave a wrong answer), resulting in less than 50 assignments per subject. In the challenge, the full training set is divided into a training and validation data set. The distribution of the correct answers in the full training data set is balanced, so that the correct answer appeared five times in the place of the first sentence, and so on.

Of the 11 subjects, seven were randomly chosen to take part in test data measurements. The *test set* consists of 180 assignments. To make cheating harder, all assignments within the test set are unique, and each assignment was shown to only one of the subjects. The locations of the relevant lines and correct answers in the test stimuli was randomly chosen, without balancing. The test data is constructed to be more real life-like, with less controlled questions and candidate sentences. It can therefore be expected that the classification rate is lower with the test data than with the training data.

4.2 Equipment

The device used for measuring eye movements was Tobii 1750 eye tracker[3], shown in Figure 2. The eye tracker is integrated into a 17" TFT monitor. The tracker illuminates the user with two near infrared diodes (they can be seen in Figure 2) to generate reflection patterns on the corneas of the user. A video camera then gathers these reflection patters as well as the stance of the user. Digital image processing is then carried out for extracting the pupils from the video signal. The systems tracks pupil location and pupil width at the rate of 50 Hz. The pupil locations can be mapped to gaze locations on the screen by calibrating the system; during the process the user needs to gaze at sixteen pre-defined locations on the screen.

The manufacturer reports the spatial resolution (frame-to-frame variation of the measured gaze point) to be 0.25 degrees and the average accuracy (bias error, deviation between the measured and actual gaze point of the user) of approximately 0.5 degrees. Additionally, the calibration deteriorates over time due to changes in the pupil size or if the eyes become dry. The associated drift of calibration is less than 0.5 degrees. The system allows free head motion in a cube of 30x15x20 cm at 60 cm from tracker. The resolution of the tracker is 1280x1024, and the recommended distance of the user from the display is 60 cm.

[3] Web pages at http://www.tobii.com. On 24 February 2005 a product description of the Tobii 1750 was available at http://www.tobii.com/downloads/Tobii_50series_PD_Aug04.pdf

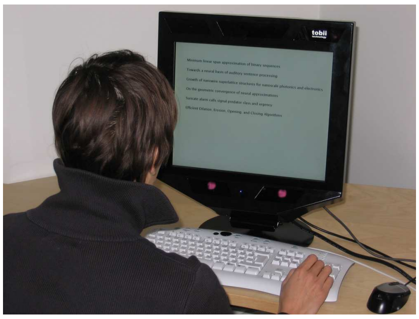


Figure 2: Eye movements of the subjects were measured with a Tobii 1750 eye tracker.

5 Feature Extraction

There are not many publications on the initial preprocessing of eye movement data (see [37] for an example). To our knowledge, the Tobii eyetracker does not preprocess the data[4].

5.1 Fixation Identification

Identifying fixations is still very much an open question within the eye tracking research, as there is no consensus of the method that best segments the eye movement trajectory (see [35] for discussion on the subject). Most of the eye movement measuring equipment manufacturers provide a *window-based* segmentation algorithm as a standard software. *Hidden Markov Model*-based algorithms have only recently gained some attention in the research area [33, 45].

5.1.1 Window-based Algorithms

In a window-based algorithm, a fixation is identified by drawing a square of x pixels around the currently measured gaze location. If the next measured gaze location falls within the block, it will be counted as a possible fixation. If in n consecutive gaze locations each falls within the block drawn around the gaze point preceding it, the n points will be counted as a fixation with a duration of n times the sampling interval (in our case 20 ms). In a Tobii eye tracker the standard setting is a 50-pixel window, with a time frame of 100 ms. For reading studies the manual recommends smaller window sizes. For the PASCAL challenge Competition 1, the fixations were computed using a 20 pixel window with a 80 ms time frame.

5.1.2 HMM-based Algorithms

The first application of Hidden Markov models (HMMs) to segment eye movement trajectories was [33], where a two-state HMM was applied. The model parameters were set manually, and the model was merely used for finding the most probable (Viterbi) path through the model for a given sequence in order to

[4] The Tobii however computes a *validity code* for each measurement, describing whether it tracks reliably both eyes or only one eye.

segment the trajectory. A more realistic application of the HMMs was presented in [45], where the parameters of a two-state HMM were learned from data.

Competitors taking part in the PASCAL Challenge Competition 2 may try to find the optimal segmentation method giving the best classification accuracy. Alternatively, they can of course decide to skip the segmentation part altogether.

5.2 Features for Competition 1

After segmenting the eye movement trajectory into fixations and saccades, they were assigned to the nearest word. After that, features for each word can be computed. All the features for the Competition 1 are listed in Table 1. We will next discuss the psychological justification behind the features.

The eye movement features used in psychological studies are often categorized into first-pass and second-pass measures, according to the order the region of text is encountered during reading. First-pass reading features are generally used as the primary measure of interest or as the measures of initial processing, whereas second-pass measures reflect the processes associated with re-analysis or “late processing” of the text region [29]. We expect the latter measures to play an important role in the challenge setup, for example in a case when the subject is choosing between two candidates of correct answers.

The eye movement features used in the challenge can additionally be divided into measures that are obtained from eye fixations, regressions, saccades, or pupil dilation data. In addition to the 22 features provided in the Competition 1, we will also briefly list some measures used in psychological studies for analysing the time series nature of the data, such as re-fixations and word skipping. These measures can be easily computed from the Competition 1 data.

Any single measure of processing would be an inadequate reflection of the reality of cognitive processing. To obtain a good description about the cognitive processes occurring during our task, a large number of different features need to be analysed. Features used in this paper and the challenge are listed in Table 1.

5.2.1 Fixation features

Typical measures of initial processing are first fixation duration (firstFixDur) and first-pass reading time or gaze duration (firstPassFixDur), which is the sum of all fixation durations on a region prior to moving to another region [3]. Additional measures for exploring early processes are the probability of fixating the target word (P1stFixation) when the region is initially encountered and the number of fixations received during first pass reading (FirstPassCnt). The duration of the fixation preceding the first fixation onto the current word (prevFixDur) and the duration of the next fixation after which the eyes moved to the next word (nextFixDur) were included in our analysis. In this paper, one measure of re-analysis or “late processing” was the probability that the word was fixated during second-pass reading (P2ndFixation). Measures covering all the fixations that landed on each word were also analysed. Mean fixation durations (meanFixDur), sums of all fixation durations on a word (totalFixDur) and the total number of fixations per word (fixCount) were computed, as well as the ratio between the total fixation duration and the total duration of fixations on the display (timePrctg).

5.2.2 Fixation position features

Landing position of first fixation on the word is used for exploring the early processing, whereas the launch site or the last location of the eyes before landing on the target word is used as a "control" for "parafoveal" preprocessing of the target word [3]. There is variability in where the eyes land on a word, but usually people tend to make their first fixation on a word about halfway between the beginning and the middle of a word [29]. This prototypical location is labelled as the *optimal viewing position*, where the word recognition time is minimized. Extensive research effort has been made to examine the consequences of making fixations at locations other than the optimal viewing position. It has been shown that the further the eyes land from the optimal position on a word the more likely there will be a refixation onto that word. We computed three measures that take the fixation position into account. The distance (in pixels) between the fixation preceding the first fixation on a word and the beginning of the word (prevFixPos), the distance of the first fixation on a word from the beginning of the word, and the launch site of the last fixation on the word from the beginning of the word (leavingPosition) were included.

5.2.3 Regressions

Approximately 10–15 % of fixations are regressions to previously read words. A common hypothesis is that eye movements during reading are mainly controlled by reasonably low-level processes in the brain, and higher level processes only interfere when something needs to be clarified. The second-pass measures such as regressions are therefore commonly accepted as indicators of higher-order cognitive processes. This may occur with a delay, since the transmission and processing of neural signals takes time.

In studies of reading it has been noted that the text difficulty has a strong influence on the number of regressions the readers make. Studies have also demonstrated that a regression was triggered when readers encountered a word indicating that their prior interpretation of a sentence was in error. Therefore it is likely that some of the regressions are due to comprehension failures [29].

Four regression measures were included in our set of features. We computed the number of regressions leaving from a word (nRegressionsFrom), the sum of durations of all regressions leaving from a word (regressDurFrom) and the sum of the fixation durations on a word during a regression (regressDurOn). It has been noted that sometimes the processing of a word "spills" on to reading the next word. Data analysis in [28] showed that most regressions originated from positions that were relatively close to a target word. In their dataset, of all the regressive saccades made within one line of text, 26 % came from within the same word (regressive refixations), 49.4 % came from the immediately following word, and 24.6 % came from more distant locations. We therefore included a binary feature (nextWordRegress) indicating whether the regression initiated from the following word.

5.2.4 Saccade features

Two saccade measures were included in the present paper. We computed the distance (in pixels) between the launch site of a saccade and its landing position, when the fixation following the saccade was the first fixation onto a word (firstSaccLen) and when the fixation was the last fixation on a word (lastSaccLen).

5.2.5 Pupil features

There is evidence that the processing of complex sentences not only takes longer but it also produces a larger change in pupil diameter [2, 16]. Therefore two measures of pupil diameter were included in our analysis.

The mean horizontal pupil diameter during fixations on the current word was computed (pupilDiam1), as well as the maximum of pupil dilation within $0.5 - 1.5$ seconds after encountering the word (pupilDiam2). The latter was the measure used in [16]. The measures were calibrated by subtracting the mean pupil diameter of the subject during the measurement.

5.2.6 Refixations

Refixation is a fixation to the currently processed word or text region. Some refixations occur because the gaze falls initially in a suboptimal place for processing the word, and a refixation takes the eyes to a more optimal viewing location [29]. The most frequent pattern is to first fixate near the beginning of the word followed by a fixation near the end of the word. Also contextual variables and incomplete lexical processing have been shown to have an effect on whether readers refixate on a current word. In [11] refixations were measured with sentences as the units of analysis. They computed the frequency and duration of reinspective fixations during the first reading of a sentence (reinspections). Hyönä [11] measured also the frequency and duration of looks back to a sentence that had already been read (look backs), and the frequency and duration of looks from a sentence back to an already read sentence (look froms). Reinspective and look-back fixations presented in [11] differ from regressions in that the saccadic direction is not decisive; rather, fixations that land on a previously fixated text region are defined either as reinspections (when reading parts of the currently fixated sentence) or look backs (when reading parts of a previously read sentence). All measures in [11] were computed as a ratio per character to provide adjustment for differences in length across sentences.

5.2.7 Skipping

There is experimental evidence that context has a strong effect on word skipping [29]. When the following words can be easily predicted from the context, they are more frequently skipped. Also high-frequency and short words are more easily skipped.

Note on the selected units of measures In psychology the most common unit of saccade lengths has been visual angle, which has the benefit of being independent of distance from stimuli. In studies of reading, saccade lengths have also been reported to scale with respect to font size. Both of these measures naturally demand that the subject's head is kept fixed throughout the measurements. Since the subject is allowed to move quite freely in our experiment (without losing too much accuracy), we will report saccade lengths in pixels, because converting them to angles or letter sizes would only add noise to the measures due to movement of the subjects. The pixel measures with respect to each subject are comparable, since the stimuli were the same for all subjects, as was the the average distance of the subject to the display. Finally, the fixation identification algorithms provided by manufacturers of measuring equipment use the same units.

5.3 Features for Competition 2

In the challenge Competition 2, the raw eye movement data will be provided. The competitors are free to compute their own features from the x- and y-coordinates of gaze location and the pupil diameter. The given values are averages of the left and right eye.

6 Baseline Methods

6.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is one of the simplest means of classification, and it is discussed in most textbooks on applied statistics or multivariate techniques. The presentation here follows the one in [36].

The idea in LDA is to find new variables which are linear combinations of the original ones, such that different classes are discriminated as well as possible. Discrimination is measured by $SS_{between}/SS_{within}$, where $SS_{between}$ is the sum of squares between classes and SS_{within} the sum of squares inside a single class, defined by

$$SS_{within} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} x_{gi}^2 \quad , \tag{1}$$

$$SS_{between} = \sum_{g=1}^{G} n_g (\bar{x}_g - \bar{x})^2 \quad , \tag{2}$$

where x_{gi} is the observation number i in class g, n_g is the number of observations in class $g = 1, \cdots, G$, $\bar{x}_g$ the mean of the observables in class g, and $\bar{x}$ the mean over all observations. In [36], the calculations needed to find optimal new axes are covered. We will next discuss how new observations are classified.

Let p_j be the prior probability and $f_j(x)$ the density function for class π_j. The observation x is allocated to the class π_j for which the probability of misclassification,

$$\sum_{i=1, i \neq j}^{G} p_i f_i(x) \quad , \tag{3}$$

is minimal. Clearly, this is the same as maximizing

$$\ln[p_j f_j(x)] \quad . \tag{4}$$

Assuming that x comes from a normal distribution, we get the classification rule (ignoring constants)

$$\text{argmax}_j \ [\ln p_j - 1/2 \ln |\Sigma_j| - 1/2(x - \mu_j)\Sigma_j^{-1}(x - \mu_j)], \tag{5}$$

where Σ_j is the covariance matrix and μ_j the mean vector for class π_j in the training set.

6.2 Hidden Markov Models

In order to explain user behavior, the sequential nature of the reading process has to be modelled. Hidden Markov models are the most common methods for modeling sequential data. In eye movement research, hidden Markov models have earlier been used for segmenting the low-level eye movement signal to detect focus of attention (see [45]) and for implementing (fixed) models of cognitive processing [34], such as pilot attention patterns [9].

Hidden Markov models optimize the log-likelihood of the data Y given the model and its parameters Θ, that is, $\log p(Y|\Theta)$. The goal is to optimize the parameters of the model so that the distribution of the data is expressed as accurately as possible. HMMs are *generative models*; they attempt to describe the process of how the data is being generated. Therefore they can be said to *emit* (produce) observations.

Long-range time dependencies within the data are taken into account by adding hidden states to the model. The changes in the distributions of the emitted observations are associated with transitions between hidden states. The transitions (as well as the observation distributions) are modelled probabilistically. There exists a well-known algorithm for learning the HMMs, namely the Baum-Welch (BW) algorithm, if all the probabilities within the model are expressed using distributions which are within the exponential family [1]. Baum-Welch is a special case of Expectation-Maximization (EM) algorithm, and it can be proven to converge to a local optimum.

6.2.1 Simple Hidden Markov Model for Each Class

The simplest model that takes the sequential nature of data into account is a two-state HMM. We optimized one model individually for each class. In a prediction task the likelihood of each model is multiplied by the prior information on the proportions of the different classes in the data. As an output we get the maximum a posteriori prediction.

6.2.2 Discriminative Hidden Markov Models

In speech recognition, where HMMs have been extensively used for decades, the current state-of-the-art HMMs are discriminative. Discriminative models aim to predict the relevance $B = \{I, R, C\}$ of a sentence, given the observed eye movements Y. Formally, we optimize $\log p(B|Y, \Theta)$. In discriminative HMMs, a set of states or a certain sequence of states is associated with each class. This specific state sequence then gives the probability of the class, and the likelihood is maximized for the teaching data, versus all the other possible state sequences in the model [26]. The parameters of the discriminative HMM can be optimized with an extended Baum-Welch (EBW) algorithm, which is a modification of the original BW algorithm.

6.2.3 Discriminative Chain of Hidden Markov Models

A main difficulty in the information retrieval setup is that relevance is associated with titles, not with words in a title. For example, there are words in titles which are not needed in making the decision on whether the title is relevant or not. There could be many such non-relevant words in a sentence, and possibly only one word which is highly relevant. The situation thus resembles the setup in reinforcement learning: the reward (classification result) is only known in the end, and there are several ways to end in the correct classification.

In order to take into account the whole eye movement trajectory during a task, we model eye movements with a two-level discriminative HMM (see Figure 3). The first level models transitions between sentences, and the second level transitions between words within a sentence. Viterbi approximation is used to find the most likely path through the second level model (transitions between words in a sentence), and then the discriminative Extended Baum-Welch optimizes the full model (cf. [18, 40] for similar approaches).

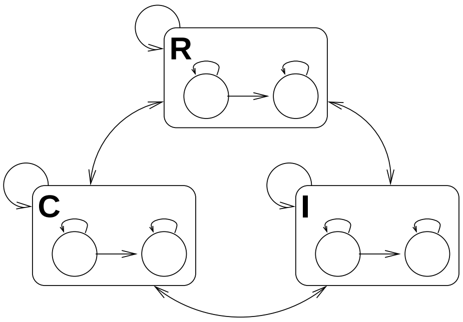


Figure 3: The topology of the discriminative chain of hidden Markov models.

In our implementation, the first level Markov model has three states, each state corresponding to one class of titles. Each of the three states in the first level have the following exponential family distributions:

1. A multinomial distribution emitting the relevance of the line, B. The parameters of this distribution were fixed, resulting in a discriminative Markov chain model in which each state corresponds to a known classification.

2. A *Viterbi distribution* emitting the probability of the sequence of words in a title.

The Viterbi distribution is defined by the probability of a Viterbi path trough a two-state Markov model forming the second level in our model. The two states of the second level model emit the observed word-specific distributions. The second level Viterbi distributions are further parameterized by the probabilities of beginning the sequence from that state (for example $\Pi^R = \pi_1^R, \pi_2^R$), and transition probabilities between states (e.g., a_{ij}^R, $i, j = 1, 2$). The second level Markov model is called a Viterbi distribution because when evaluating the emission probability only the most likely path over the two-state model is taken into account (the Viterbi path). After fixing the path the resulting Viterbi distribution is (a fairly complex) exponential family distribution that can be trained with the EBW algorithm.

6.2.4 Voting

The Markov models produce probabilities for the relevance classes (I, R, C) for each viewed sentence. However, the users may look at a sentence several times, and the resulting probabilities need be combined in a process we call *voting*.

We constructed a log-linear model for combining the predictions. Assume that the sentence-specific probability distribution, $p(B|Y_{1...K})$, can be constructed from the probability distributions of the kth viewings of the sentence, $P(B|Y_k)$, (obtained as an output from a Markov model) as a weighted geometric average, $p(B|Y_{1...K}, \alpha) = Z^{-1} \prod_k p(B|Y_k)^{\alpha_{Bk}}$, where Z is a sentence-specific normalization factor and the parameters α_{Bk} are non-negative real numbers, found by optimizing the prediction for the training data. The predicted relevance of a sentence is then the largest of $p(I)$, $p(R)$, and $p(C)$.

It is also possible to derive a simple heuristic rule for classification by assuming that the decision of relevance is made only once while reading the sentence. We will call this rule maxClass, since for each sequence we will select the maximum of the predicted relevance classes. A simple baseline for the voting schemes is provided by classifying all the sequences separately (i.e., no voting).

7 Data analysis

Below we will carry out an example analysis of the challenge data. We apply Linear Discriminant Analysis to the eye movement data to obtain a first classification result, to get first visualizations of the data, and to select features that will be used in time series modeling, with HMMs and discriminative HMMs.

7.1 Linear Discriminant Analysis

Linear Discriminant Analysis is a simple linear method for analyzing data. Besides classification, the method can be used for visualization and feature set selection. It has not been developed for time series, however, and we apply it on feature vectors averaged over each sentence.

Averaged Features

Simple averaging of features presented in Table 1 would be against their spirit. The probabilities {3,4,18} are obtained by diving the sum of the features by the number of words in the sentence. Features {1,2,14,16,17,19,22} are commonly used as summary measures for larger areas of interest, and hence were merely added up. Features {5, 6, 7, 8, 9, 10, 11, 12, 13, 15} were computed as means, and for the pupil measures {20, 21} a maximum was taken (since in [16] the best effect was reported in the maximum of pupil dilation). Before analysing the data with LDA, the data was standardized.

Visualizing the Data with LDA

The data can be visualized by projecting them to the eigenvectors of the LDA (see Figure 4). The two eigenvectors define a hyperplane in the original feature space that best discriminates the classes. The visualization makes it possible to evaluate which classes will be harder to separate. Judging from the plot in Figure 4, it seems that relevant and irrelevant sentences will be hard to separate.

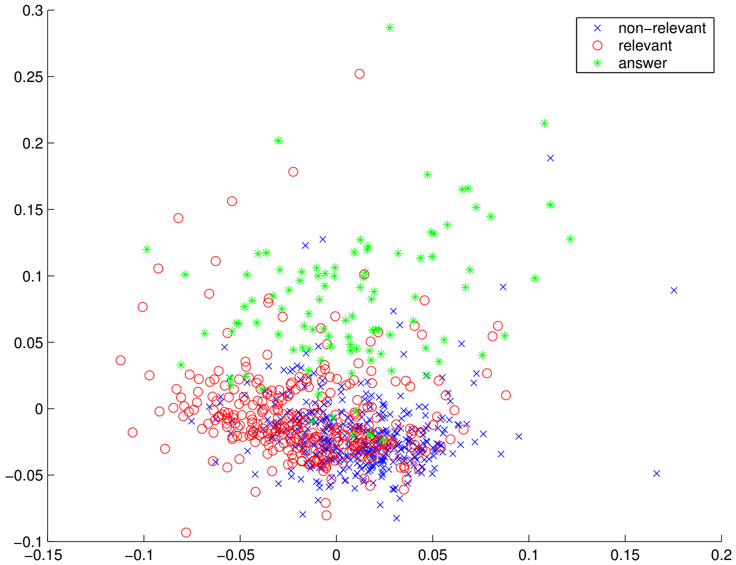


Figure 4: A visualization of the data using LDA.

Feature Set Selection with LDA

We may also plot the eigenvectors of the LDA in order to evaluate which components contribute most to the discriminating plane. Notice that if classification is not possible with LDA[5], the eigenvectors will be arbitrary. In our case, however, classification is possible as reported in Table 2. Judging from the eigenvectors plotted in Figure 5, it seems that less than ten features are sufficient.

7.2 Features for Time Series Analysis

Feature selection for the HMMs was carried out with the methods that use averaged data (LDA). In other words, we chose to model a representative set of features which can be used to construct the best discriminating averaged measures.

The resulting set of features were modeled with the following exponential family distributions: (1) One or many fixations within the word (binomial). (2) Logarithm of total fixation duration on the word (assumed Gaussian). (3)

[5] that is, the classification rate does not differ from a dumb classifier classifying all to the largest class

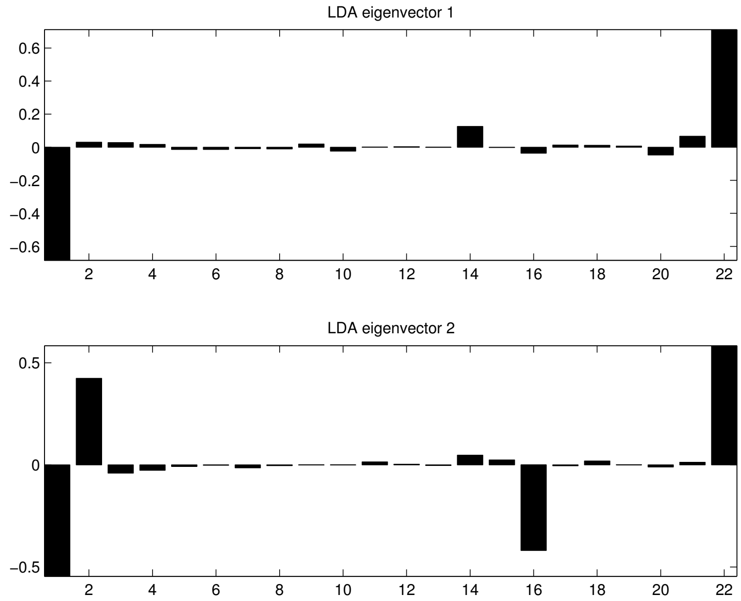


Figure 5: Eigenvectors of LDA. The histogram bars are loadings of the features, ordered according to Table 1.

Reading behavior (multinomial): skip next word, go back to already read words, read next word, jump to an unread line, or last fixation in an assignment.

7.3 Classification results

The prediction accuracy was assessed with 50-fold cross validation, in which each of the assignments was in turn used as a test data set. In order to test how the method would generalize to new subjects, we also ran an 11-fold cross validation where each of the subjects was in turn left out. The ultimate baseline is given by the "dumb model," which classifies all sentences to the largest class I. Table 2 lists the classification accuracies, that is, the fraction of the viewed sentences in the test data sets for which the prediction was correct. The methods generalize roughly equally well both to new assignments and to new subjects. The performance of the two different voting methods (log-linear and maxClass) seems to be nearly equal, with log-linear voting having a slight advantage.

Table 3 shows the confusion matrix of the discriminative HMMs. Correct answers (C) are separated rather efficiently. Most errors result from misclassifying relevant sentences (R) as irrelevant (I). It is also possible to compute precision and recall measures common in IR, if the correct answers are treated as the relevant documents. The resulting precision rate is 90.1 % and recall rate 92.2 %.

8 Discussion

The physiology and psychology of eye movements has been studied quite extensively. However, the improved multimodal interfaces, combined with proactive information retrieval, provide us with a whole new setup. The eye movements are a rich, complex, and potentially very useful time series signal. Efficient extraction of relevance information from it is not trivial, however, and requires development and application of advanced machine learning methods.

The features used in eye movement research have been based mostly on the segmentation of the eye movement trajectory to fixations and saccades. This segmentation, though useful, is neither unique nor always optimal. The optimal set of features is likely to depend on the task at hand. One of the goals of Competition 2 of this Challenge is to find and propose a new set eye movement features, not necessarily based on the division to fixations and saccades, for use in eye movement studies and proactive applications.

In the study of eye movements in psychology the basic goal is to understand the underlying psychological processes. Our objective is different and more application-oriented: we want to extract maximal amount of useful information from the real-world eye movement signal, to be used in proactive information retrieval. Our approach also differs from usability studies, another common application of eye movement analysis, where the objective has been to analyze qualitatively and quantitatively the behavior of a user when she for instance visits a web site. The quantitative measures have been mostly based on fixation durations and eye scan patterns. This Challenge differs from much of the prior work in its application and experimental setup (information retrieval task where the ground truth is known) and in the use of advanced probabilistic methods optimized for the task at hand (relevance extraction).

The Challenge will hopefully result in a toolbox of new machine learning methods and a set of features, optimal for extracting relevance information from the real world eye movement signals.

We are looking forward to an interesting competition and wish all participants the best of success!

Acknowledgments

This work was supported by the Academy of Finland, decision no. 79017, and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. The authors would like to thank people in the PRIMA project for useful discussions, and acknowledge that access rights to the data sets and other materials produced in the PRIMA project are restricted due to other commitments.

The authors are greatly indebted to Maria Sääksjärvi and Johanna Gummerus, Center for Relationship Marketing and Service Management at Swedish School of Economics and Business Administration at Helsinki, for providing us the measurement time. We would also like to thank Kari-Jouko Räihä, Aulikki Hyrskykari and Päivi Majaranta from Tampere Unit for Computer-Human Interaction for useful discussions and measurement time.

References

[1] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic func-

tions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, February 1970.

[2] Jackson Beatty and Brennis Lucero-Wagoner. The pupillary system. In John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson, editors, *Handbook of Psychophysiology*, chapter 6. Cambridge University Press, Cambridge, UK, 2000.

[3] Manuel G. Calvo and Enrique Meseguer. Eye movements and processing stages in reading: Relative contribution of visual, lexical and contextual factors. *The Spanish Journal of Psychology*, 5(1):66–77, 2002.

[4] Christopher Campbell and Paul Maglio. A robust algorithm for reading detection. In *Workshop on Perceptive User Interfaces (PUI '01)*. ACM Digital Library, November 2001. ISBN 1-58113-448-7.

[5] Ralf Engbert and Reinhold Kliegl. Microsaccades uncover the orientation of covert attention. *Vision Research*, 43:1035–1045, 2003.

[6] Ralf Engbert, André Longtin, and Reinhold Kliegl. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42:621–636, 2002.

[7] Joseph H. Goldberg, Mark J. Stimson, Marion Lewenstein, Neil Scott, and Anna M. Wichansky. Eye tracking in web search tasks: design implications. In *ETRA '02: Proceedings of the symposium on Eye tracking research & applications*, pages 51–58. ACM Press, 2002.

[8] Laura A. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of SIGIR'04*, pages 478–479. ACM Press, 2004.

[9] Miwa Hayashi. Hidden markov models to identify pilot instrument scanning and attention patterns. In *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, pages 2889–2896, 2003.

[10] S.B. Hutton, I. Cuthbert, T.J. Crawford, C. Kennard, T.R.E. Barnes, and E.M. Joyce. Saccadic hypometria in drug-naïve and drug-treated schizophrenic patients: A working memory deficit? *Psychophysiology*, 38:125–132, 2001.

[11] J. Hyönä, R.F. Lorch Jr, , and J.K. Kaakinen. Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology*, 94:44–55, 2002.

[12] J. Hyönä, J. Tommola, and A.M. Alaja. Pupil dilation as a measure of processing load in simultaneous interpreting and other language tasks. *Quarterly Journal of Experimental Psychology*, 48A:598–612, 1995.

[13] Aulikki Hyrskykari, Päivi Majaranta, and Kari-Jouko Räihä. Proactive response to eye movements. In G. W. M. Rauterberg, M. Menozzi, and J. Wesson, editors, *INTERACT'03*. IOS press, 2003.

[14] R.J.K. Jacob and K.S. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises (section commentary). In J. Hyona, R. Radach, and H. Deubel, editors, *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pages 573–605. Elsevier Science, Amsterdam, 2003.

[15] Robert J. K. Jacob. *Eye tracking in advanced interface design*, pages 258–288. Oxford University Press, 1995.

[16] Marcel Adam Just and Patricia A. Carpenter. The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, 47(2):310–339, 1993.

[17] K.J. Ciuffreda KJ and B. Tannen. *Eye Movement Basics for the Clinician.* Mosby Yearbook, St. Louis, 1995.

[18] Mikko Kurimo. *Using Self-Organizing Maps and Learning Vector Quantization for Mixture Density Hidden Markov Models.* PhD thesis, Helsinki University of Technology, Espoo, Finland, 1997.

[19] Jochen Laubrock, Ralf Engbert, and Reinhold Kliegl. Microsaccade dynamics during covert attention. *Vision Research*, 45:721–730, 2003.

[20] Gordon E. Legge, Timothy S. Klitz, and Bosco S. Tjan. Mr. chips: An ideal-observer model of reading. *Psychological Review*, 104(3):524–553, 1997.

[21] Simon P. Liversedge and John M. Findlay. Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4(1):6–14, 2000.

[22] Paul P. Maglio, Rob Barrett, Christopher S. Campbell, and Ted Selker. Suitor: an attentive information system. In *Proceedings of the 5th international conference on Intelligent user interfaces*, pages 169–176. ACM Press, 2000.

[23] Paul P. Maglio and Christopher S. Campbell. Attentive agents. *Commun. ACM*, 46(3):47–51, 2003.

[24] Bradford Miller, Chung Hee Hwang, Kari Torkkola, and Noel Masseya. An architecture for an intelligent driver assistance system. In *Proceedings of IEEE Intelligent Vehicles Symposium*, pages 639–644, June 2003.

[25] Timo Partala, Maria Jokiniemi, and Veikko Surakka. Pupillary responses to emotionally provocative stimuli. In *Proceedings of Eye Tracking Research and Applications (ETRA2000)*, pages 123–129. ACM press, 2000.

[26] D. Povey, P.C. Woodland, and M.J.F. Gales. Discriminative map for acoustic model adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)*, volume 1, pages 312–315, 2003.

[27] `http://www.poynter.org/eyetrack2000/`.

[28] R. Radach and G.W. McConkie. Determinants of fixation positions in words during reading. In G. Underwood, editor, *Eye Guidance in Reading and Scene Perception*, pages 77–100. Elsevier, Oxford, 1998.

[29] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

[30] Erik D. Reichle, Alexander Pollatsek, Donald L. Fisher, and Keith Rayner. Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125–157, 1998.

[31] Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel Kaski. Can relevance be inferred from eye movements in information retrieval? In *Proceedings of WSOM'03, Workshop on Self-Organizing Maps*, pages 261–266. Kyushu Institute of Technology, Kitakyushu, Japan, 2003.

[32] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Relevance feedback from eye movements for proactive information retrieval. In Janne Heikkilä, Matti Pietikäinen, and Olli Silvén, editors, *workshop on Processing Sensory Information for Proactive Systems (PSIPS 2004)*, Oulu, Finland, 14-15 June 2004.

[33] Dario D. Salvucci and John R. Anderson. Tracing eye movement protocols with cognitive process models. In *Proceedings of the Twentieth Annual Conference of the Cognitive Society*, pages 923–928, Hillsdale, NJ, 1998. Lawrence Erlbaum Associates.

[34] Dario D. Salvucci and John R. Anderson. Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16:39–86, 2001.

[35] D.D. Salvucci and J.H. Goldberg. Identifying ixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research and Applications Symposium 2000 (ETRA2000)*, 2000.

[36] Subhash Sharma. *Applied Multivariate Techniques*. John Wiley & Sons, Inc., 1996.

[37] Dave M. Stampe. Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavior Research Methods, Instruments & Computers*, 25(2):137–142, 1993.

[38] India Starker and Richard A. Bolt. A gaze-responsive self-disclosing display. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 3–10. ACM Press, 1990.

[39] M.S. Starr and K. Rayner. Eye movements during reading: some current controversies. *Trends in Cognitive Sciences*, 5(4):156–163, 2001.

[40] A. Stolcke and S. Omohundro. Hidden markov model induction by bayesian model merging. In S.J. Hanson, J.D. Cowan, and C.L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 11–18, San Francisco, CA, 1993. Morgan Kaufmann.

[41] David Tennenhouse. Proactive computing. *Commun. ACM*, 43(5):43–50, 2000.

[42] Kari Torkkola, Noel Masseya, Bob Leivian, Chip Wood, John Summers, and Snehal Kundalkar. An architecture for an intelligent driver assistance system. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pages 81–85, June 2003.

[43] Tero Vuori, Maria Olkkonen, Monika Pölönen, Ari Siren, and Jukka Häkkinen. Can eye movements be quantitatively applied to image quality studies? In *Proceedings of the third Nordic conference on Human-computer interaction*, pages 335–338, 2004.

[44] David J. Ward and David J.C. MacKay. Fast hands-free writing by gaze direction. *Nature*, 418:838, 2002.

[45] Chen Yu and Dana H. Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. In *Proc. ICMI'03*. ACM, 2003. To appear.

A Notes on TERP

Because TERP amplitudes appear to be independent of baseline pupillary diameter, it is possible to compare the amplitude of TERPs obtained in different laboratories. Analysis of pupillometric data in memory storage and recall tasks have shown that there is variation in peak pupillary dilation as a function of the length of the target string to be stored or recalled. The item difficulty in memory tasks has also been associated with greater pupillary dilations.

There is evidence of response and movement-related pupillary responses. Results from experiments where immediate or delayed response selection and preparation were studied indicated that the rate of pupil dilation was inversely proportional to the length of the foreperiod preceding the imperative stimulus. It was shown that the pupil dilations were greater in Go-trials than dilations to No-Go stimuli in both immediate- and delayed-response conditions. Additionally, both peak pupil diameter and peak latency have been found to vary with the complexity of movements in motor tasks.

It has been reported that pupil dilations are elicited not only by external stimuli but also by a stimulus mismatch or by an orientation to a task important stimuli. An inverse relationship has been found between pupil amplitude and probability. Pupil dilations were found to be larger in amplitude and longer in latency for stimuli with low probability of occurrence.

TERP amplitude is also a sensitive and reliable reporter of differences in the structure of cortical language processing and decision. In a letter matching task, physically identical letter pairs evoked smaller TERPs than did pairs identical only at the level of naming. [16] found that more complex sentence types produced larger changes in pupil diameter. [12] reported, that increases in semantic demands of sentence processing resulted in increases of the TERP.

Table 1: Features

	Feature	Description
1	**fixCount**	Total number of fixations to the word
2	**FirstPassCnt**	Number of fixations to the word when the word is first encountered
3	**P1stFixation**	Did a fixation to a word occur when the sentence that the word was in was read for the first time ('1' or '0')
4	**P2ndFixation**	Did a fixation to a word occur when the sentence that the word was in was read for the second time ('1' or '0')
5	**prevFixDur**	Duration of the previous fixation when the word is first encountered
6	**firstFixDur**	Duration of the first fixation when the word is first encountered
7	**firstPassFixDur**	Sum of durations of fixations to a word when it is first encountered
8	**nextFixDur**	Duration of the next fixation when the gaze initially moves on from the word
9	**firstSaccLen**	Distance (in pixels) between the launching position of the previous fixation and the landing position of the first fixation
10	**lastSaccLen**	Distance (in pixels) between the launching position of the last fixation on the word and the landing point of the next fixation
11	**prevFixPos**	Distance (in pixels) between the fixation preceding the first fixation on a word and the beginning of the word
12	**landingPosition**	Distance (in pixels) of the first fixation on the word from the beginning of the word
13	**leavingPosition**	Distance (in pixels) between the last fixation before leaving the word and the beginning of the word
14	**totalFixDur**	Sum of all durations of fixations to the word
15	**meanFixDur**	Mean duration of the fixations to the word
16	**nRegressionsFrom**	Number of regressions leaving from the word
17	**regressDurFrom**	Sum of durations of fixations during regressions initiating from the word
18	**nextWordRegress**	Did a regression initiate from the following word ('1' or '0')
19	**regressDurOn**	Sum of the durations of the fixations on the word during a regression
20	**pupilDiam1**	Mean of pupil diameter during fixations on the word (minus mean pupil diameter of the subject during the measurement)
21	**pupilDiam2**	Maximum of pupil dilation within 0.5 – 1.5 seconds after encountering the word (minus mean pupil diameter of the subject during the measurement)
22	**timePrctg**	Total fixation duration on a word divided by the total duration of fixations on the display

Table 2: Performance of the different models in predicting relevance of the sentences. Differences between LDA and dumb classifier, and HMM and LDA tested significant (McNemar's test), as well as difference between discriminative HMM and simple HMMs (with leave-one-assignment-out cross validation) Left column: obtained by 50-fold cross-validation where each of the assignments was left out in turn as test data. Right column: Obtained by 11-fold cross-validation where each of the subjects was left out in turn to be used as test data.

Method	**Accuracy (%)** (leave-one-assignment-out)	**Accuracy (%)** (leave-one-subject-out)
Dumb	47.8	47.8
LDA	59.8	57.9
simple HMMs(no vote)	55.6	55.7
simple HMMs(maxClass)	**63.5**	**63.3**
simple HMMs(loglin)	**64.0**	**63.4**
discriminative HMM(loglin)	**65.8**	**64.1**

Table 3: Confusion matrix showing the number of sentences classified by the discriminative HMM, using loglinear voting, into the three classes (columns) versus their true relevance (rows). Cross-validation was carried out over assignments. The percentages (in parentheses) denote row- and column-wise classification accuracies.

		Prediction	
	I (62.4 %)	*R* (61.8 %)	*C* (90.1 %)
I (77.3 %)	1432	395	25
R (43.6 %)	845	672	24
C (92.2 %)	17	21	447