Serrano Martinez Santos, N.; Civera Saiz, J.; Sanchis Navarro, JA.; Juan, A. (2014). Effective balancing error and user effort in interactive handwriting recognition. Pattern Recognition Letters. 37(1):135-142. doi:10.1016/j.patrec.2013.03.010.

Effective Balancing Error and User Effort in Interactive Handwriting Recognition

N. Serrano, J. Civera, A. Sanchis and A. Juan

DSIC, Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain

Abstract

Transcription of handwritten text documents is an expensive and time-consuming task. Unfortunately, the accuracy of current state-of-the-art handwriting recognition systems cannot guarantee fully-automatic high quality transcriptions, so we need to revert to the computer assisted approach. Although this approach reduces the user effort needed to transcribe a given document, the transcription of handwriting text documents still requires complete manual supervision. An especially appealing scenario is the interactive transcription of handwriting documents, in which the user defines the amount of errors that can be tolerated in the final transcribed document. Under this scenario, the transcription of a handwriting text document could be obtained efficiently, supervising only a certain number of incorrectly recognised words. In this work, we develop a new method for predicting the error rate in a block of automatically recognised words, and estimate how much effort is required to correct a transcription to a certain user-defined error rate. The proposed method is included in an interactive approach to tran-

☆ *Tel:* (+34) 963877350 + 73533 *Fax:* (+34) 963877359

Email address: {nserrano,jcivera,josanna,ajuan}@dsic.upv.es (N. Serrano, J. Civera, A. Sanchis and A. Juan)

scribing handwritten text documents, which efficiently employs user interactions by means of active and semi-supervised learning techniques, along with a hypothesis recomputation algorithm based on constrained Viterbi search. Transcription results, in terms of trade-off between user effort and transcription accuracy, are reported for two real handwritten documents, and prove the effectiveness of the proposed approach.
Keywords: Handwriting Recognition, Computer-assisted Annotation, Accuracy prediction

1. Introduction

Information has been stored for posterity for centuries. The arrival of the digital era has led to efficient storage and access to this information, but in some cases its latter digestion and analysis present challenging problems. This is the case of handwritten text recognition (HTR). Nowadays, there is a great interest in the study of information stored in manuscripts in libraries all over the world. However, these manuscripts cannot be fully exploited by natural language processing (NLP) tools if transcriptions are not available in an electronic format. Furthermore, transcription of handwritten text documents is an expensive and time-consuming task, which in most cases has to be carried out by paleographic experts. Despite the fact that HTR has been studied since the beginning of Pattern Recognition (PR), current state-of-the-art systems (Graves et al., 2009) still cannot produce fully-automatic high quality transcriptions. This has led to the integration of automatic HTR systems as an assistive tool in the transcription process by experts. The idea behind this integration is to reduce the effort required to generate

transcriptions while guaranteeing high levels of accuracy. This approach is commonly referred as computer assisted transcription (CAT).

CAT systems deal with the interactive transcription of a handwritten text document, where the user is continuously aided by a system. The main problem with this approach is that user supervisions have to be efficiently employed, as their overuse may cause the user to ignore the system and transcribe the document manually. In previous works, we have focused on developing techniques to reduce user effort and maximise its utility. For instance, in (Serrano et al., 2009), active learning is used together with semi-supervised learning techniques to adapt (and improve) the system from partially-supervised transcription. Alternatively, in (Serrano et al., 2010a), we developed a technique to improve the current system hypothesis when a user interaction is performed, and thus improve the final transcription. These techniques were implemented on top of an open source interactive prototype called GIDOC (Serrano et al., 2010c).

Although the aim of CAT tools is to save on user effort when transcribing a document, its complete annotation still requires the manual revision of the whole document. It is therefore difficult to measure how much user effort is actually saved when transcribing a document with a CAT tool. In contrast, an alternative approach to CAT is to predefine the desired transcription accuracy after the transcription process. This means that we are accepting an amount of residual error in our transcriptions in order to save on user effort. For instance, an automatically transcribed document that has been partially supervised by a user may contain a small number of errors but still it can be sufficient to convey the meaning. Similarly, there

are many applications dealing with tasks that tolerate erroneous input. For example, the output of an Automatic Speech Recognition (ASR) system can be successfully used as input for well-known tasks such as dialogue act annotation (Stolcke et al., 2000), information retrieval (Grangier et al., 2003), or speech-to-speech translation (Matusov et al., 2006). All these applications may not require perfect annotation of the data, but only a sufficiently good annotation that guarantees the desired accuracy at lower user effort. In this scenario, the ideal CAT tool achieves the required transcription accuracy in exchange of the minimum user effort.

We have studied this latter scenario in the transcription of handwritten text documents (Serrano et al., 2010b) and, more recently, the transcription of speech (Sánchez-Cortina et al., 2012). In these works, we developed a simple yet effective algorithm for estimating the expected error of recognised words that have not been supervised yet. This algorithm was used to adjust the error of transcriptions produced by a CAT system to a given user-defined error threshold. However, even though the described approach guaranteed that the error on the final transcriptions was below the user-defined threshold, it was far too pessimistic and required from the user more effort than was actually needed. In this work, we proposed a new algorithm for predicting the error-rate of recognised words of a HTR system, which outperforms our previous algorithm. This improvement is mainly due to two factors. First, a more precise estimation of the error for each word. Second, the estimation of the error is now performed for a whole block of words, which is more accurate that the previous biased, line by line estimation. This new algorithm will be combined with the best-performing techniques presented in previous works.

Our CAT system was evaluated on two real handwritten text documents showing that user effort was closely estimated by the proposed algorithm.

The rest of this paper is organised as follows. First, a brief description of related work is provided in Section 2. In Section 3 we present our new error estimation algorithm. Section 4 shows the empirical results of the proposed approach. Finally, conclusions are drawn and future work is envisioned in Section 5.

2. Related Work

The present work deals with the interactive transcription of handwritten text documents, in which a defined quantity of errors in the transcriptions produced can be tolerated in exchange for a substantial savings of manual effort in the annotation process. This approach deals with multiple techniques to successfully complete the task, such as active learning, semi-supervised learning or error-rate prediction. In the following section, we describe the similarity between the diverse components of our approach and previous works, because to our knowledge there are not previous works integrating all the techniques in the same system.

User supervision is typically the most expensive and time-consuming resource in the transcription process. In our case, we deal with the correction of machine-generated output, in which user supervision is only employed to supervise recognised words. Consequently, two problems are tackled in our CAT system. First, the user effort available must be intelligently employed in supervising incorrectly-recognised words, and secondly, unsupervised correctly-recognised words should be identified to be incorporated as

training data. The first problem is solved by applying active learning algorithms (Settles, 2009), while the second is solved using semi-supervised learning techniques (Zhu, 2006).

It is worth noting that the combination of active and semi-supervised learning is really necessary for our CAT system to achieve a maximum improvement of transcription accuracy with minimum user effort. Active and semi-supervised learning are used to select the most suitable unannotated samples for user supervision and system adaptation respectively. They can be applied separately or, for better results, in combination, so as to boost their complementary beneficial effect. Indeed, their combination has recently been studied in areas other than HTR, such as ASR (Tur et al., 2005), image retrieval (Zhou et al., 2006) and other fields (Wang and Zhou, 2008). Usually, the key idea behind these learning techniques is the use of confidence measures (CMs) (Wessel et al., 2001; Sanchis et al., 2012) to measure the uncertainty of each hypothesis. In our HTR case, a recognised word with a low confidence value is likely to be an error, whereas a high confidence word is expected to be correctly recognised. Therefore, low confidence words are candidates for supervision, while high confidence words are likely to be useful for system adaptation (re-training).

CAT approaches exploit the impact of user supervision beyond the simplistic idea of correcting incorrectly-recognised words. An incorrectly-recognised word in a given text line, typically affects the surrounding words, generating more errors. When the user supervises a recognised word, the uncertainty of the system around that word is reduced. In this regard, one of the most successful approaches is the prefix-based approach. The main idea of this ap-

proach is to improve the system hypothesis on a sample by recomputing the best system hypothesis constrained to a correct prefix. Specifically, first, the user validates the prefix of a system hypothesis up to the first incorrect word, which is corrected. Next, the validated prefix and the user corrected word are employed to predict the remaining suffix by constraining the search process. This process is repeated until the whole transcription has been revised. This approach has been the base of many works dealing with very different applications, such as HTR (Toselli et al., 2007), ASR (Revuelta-Martínez et al., 2012) or syntactic tree annotation (Sánchez-Sáez et al., 2010). All these approaches successfully reduce the effort needed to obtain the required output. However, as mentioned above, the whole machine-generated transcription still has to be revised by a user. Although our approach also follows the idea of constrained search, it must not be confused with the described prefix-based approach. As explained above, in our case we consider a limited amount of user effort, which keep us from supervising the complete output, but only those words that are likely to be wrong. This leads to the supervision of individual words in the output transcription rather than complete prefixes or suffixes. Supervision of individual words saves a significant amount of user effort by focusing user attention on those parts most likely to need correcting. In order to perform a search process constrained to those isolated words supervised by the user, we extrapolated the constrained-Viterbi search proposed by (Kristjannson et al., 2004) for information retrieval to HTR.

So far, we have described some techniques to efficiently exploit a limited amount of user supervision. Nevertheless, in our approach, we must first estimate the error-rate of a set of recognised words, to then decide on the su-

pervision effort to achieve the error rate desired by the user. This problem is typically known in the literature as accuracy or error-rate prediction. In the following, we speak in terms of error-rate prediction (EP), as our results are reported in error rate. EP has been typically used on practical applications. In these applications, EP estimation typically employs CMs to validate system performance on a given task. For instance, Schlapbach et al. (2008b) used a EP system based on support vector regression in HTR, in which the estimation is employed to decide if a recognised text is readable enough. Similarly, Yoon et al. (2010) proposed a linear regression of multiple speech features to determine the quality of the English in real oral exams. Another application is to use the acoustic likelihood of an ASR system to better distribute effort in a speech transcription task (Roy et al., 2010). However, these applications were not related to computer-assisted scenarios.

In (Serrano et al., 2010b), we developed an EP estimation algorithm and employed it within a CAT approach for HTR. Although the error-rate threshold defined was not surpassed by the HTR system, the estimation was rather pessimistic, and user supervision was overused. An approach more closely related to our work was proposed by (Navarro-Cerdan et al., 2010) for optic character recognition. In their work, they develop a heuristic method to dynamically adjust the supervision given an error-rate threshold defined by the user, based on dynamic confidence intervals. In this work, we refine our previous algorithm using a probabilistic approximation based on CMs to estimate the expected error-rate in a set of recognised transcriptions.

3. Error Estimation in Automatically Recognised Words

In HTR, error is typically measured in terms of word error rate (WER). WER is calculated as

$$\mathrm{WER} = \frac{S + I + D}{N} = \frac{E}{N} \tag{1}$$

where S, I and D are the minimum number of elemental edit operations E (substitutions, insertions and deletions, respectively) needed to convert the recognised transcription into the reference transcription, and N is the number of words in the reference transcription.

Our objective is to estimate the WER of a set of unsupervised recognised words, whose reference transcription is unknown, in order to decide what level of supervision is required in order to reach the desired WER. Variables referring to the supervised and unsupervised parts are denoted with the plus and minus sign, respectively as superindices. Given a set of R^- unsupervised recognised words, its WER^- is calculated as

$$\mathrm{WER}^- = \frac{E^-}{N^-} \tag{2}$$

where E^- and N^- denote the number of editions and reference words in the unsupervised part, respectively.

In (Serrano et al., 2010b), we supposed that WER^- can be estimated as the basis of previously supervised recognised words. In that work, we assumed that errors in the supervised part occur with the same frequency as in the unsupervised part and that the ratio between recognised and reference words is also the same.

$$\frac{E^+}{R^+} \approx \frac{E^-}{R^-} \qquad \frac{R^+}{N^+} \approx \frac{R^-}{N^-} \tag{3}$$

This assumption is an upper bound of the ratio of number of errors in the recognised words since, as more blocks are added to the training set, this ratio should decrease. Therefore, by making this assumption, we guarantee that the error estimation on final transcriptions is below the user-defined error threshold.

So if we substitute our assumptions expressed in Eq. 3 into Eq. 2, we can estimate WER in the unsupervised part as

$$\mathrm{WER}^{-} \approx \frac{R^{-} \cdot \frac{E^{+}}{R^{+}}}{R^{-} \cdot \frac{R^{+}}{N^{+}}} \tag{4}$$

This estimation suffers from a major drawback in our approach. When the error is estimated, the system asks the user to correct some recognised words in order to bring the error down to the user-defined WER threshold. Even when the WER calculation is accurate, it considers that all words contribute to its calculation with the same number of editions. Specifically, the mean number of editions.

However, in practice, errors are not uniformly distributed among all recognised words. To illustrate this problem, we performed a recognition experiment on the RODRIGO database (Serrano et al., 2010c), represented in Fig. 1. In order to obtain this chart, first a block of lines are automatically recognised using our HTR system. Then, recognised words are ordered according to their CM from left (low) to right (high) in the x axis. We should note that CMs are basically defined as posterior probabilities and so their values range from zero to one. Confidence measures are expected to be correlated with the correctness of each word. In this way, low confidence words are likely to be incorrect, while high confidence words will be largely correct.

When in possession of the reference transcription, we are able to identify which words were incorrectly recognised, and compute the percentage of accumulated errors (y axis) in a set of words of increasing confidence. This set of words is characterised by its size, in terms of percentage with respect to the total number of recognised words (bottom x axis), or by the highest value of CM in that set (top x axis). Four curves representing alternative error estimators appear in Fig. 1.

The curve labelled *Real* assumes that the reference transcription is known beforehand, and so it accounts for the accumulative percentage of errors in a set of words ordered by CM. As expected, errors are more likely to occur on low confidence words, which accumulates most errors. The curve labelled as *Mean* has no access to the reference transcription and assumes that errors are uniformly distributed among recognised words, so estimating accumulative error according to Eq. 4. As observed, this is not an accurate error estimation.

At this point, it is logical to consider CMs in error estimation. As we have said, CMs are calculated as posterior probabilities which measure the probability of a recognised word being correct given its corresponding word image. Similarly, one minus the posterior probability directly represents the expected value of the error of a recognised word and could be used as an error estimator. The curve labelled as *CM* in Fig.1 shows the error estimation based on the CM of each word. As shown, this error estimator performs poorly when applied directly, since a large percentage of incorrect words are assigned high confidence values. In fact, over 40% of recognised words are assigned a confidence value of one.

Alternatively, we could also consider error estimation as a classification

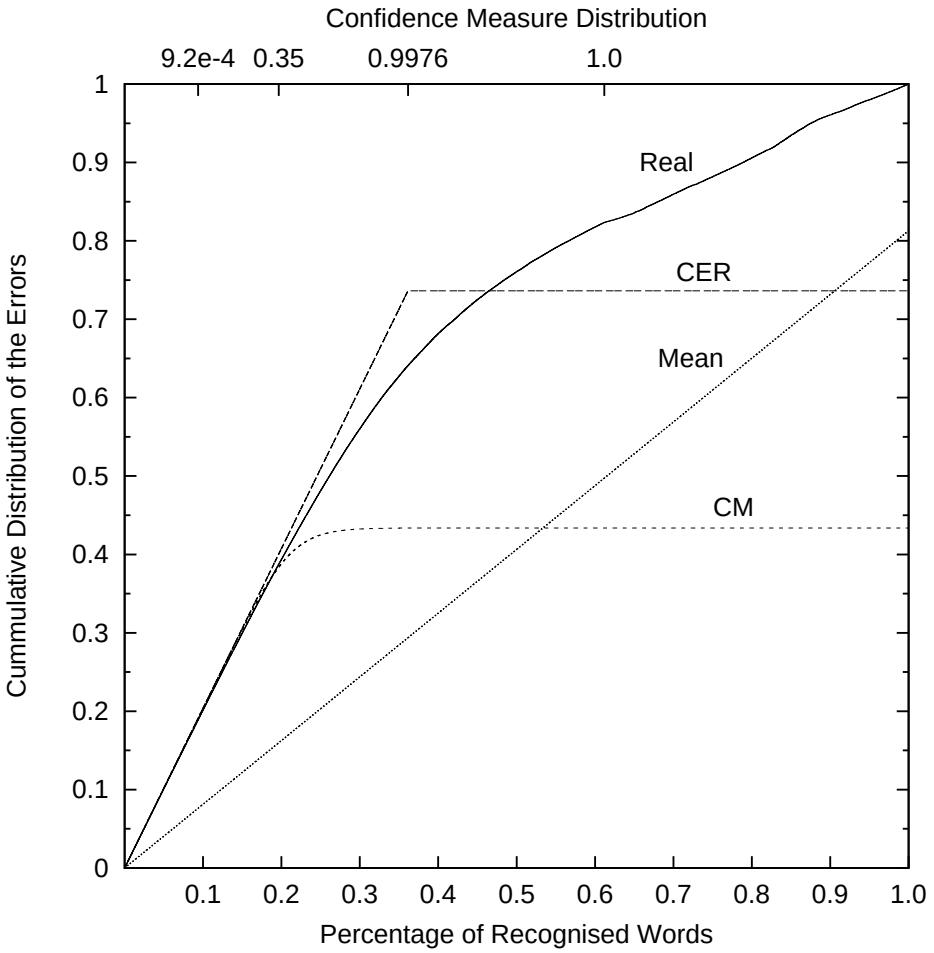

Figure 1: Cumulative distribution of errors on a set of recognised words ordered by Confidence Measure (CM). Actual error distribution represented by the curve labelled as *Real* is compared with other error estimators based on CMs.

problem, in which CMs are used to classify a recognised word as correct or incorrect (Schlapbach et al., 2008a). Classification is then performed by defining a threshold for CMs. All words below the threshold are considered incorrect, while those above are considered correct. The curve labelled as *CER* shows error estimation using a classifier based on CMs whose threshold was adjusted to optimise the Classification Error Rate (CER) on a validation set. As shown, it also results in a poor estimation because almost 25% of errors occur above the optimised threshold, the point above which errors are not considered. This empirical study reveals that confidence measures cannot be directly used to predict error on a set of recognised words.

The problem of error estimation based on CMs was slightly alleviated in our previous work by quantifying CMs in a step-wise fashion when applying Eq. 3. Due to the sequential processing (line by line) in our previous work, we supposed that errors were uniformly distributed over all lines, but not over words in the same line. Then, for each line, recognised words were ordered according to their CM and assigned to different error intervals. For instance, the first interval corresponds to the least confident word of each line; the second interval to the second least confident; the third interval to the third least confident; and the fourth interval includes the remaining words. However, this error estimation was rather pessimistic because of the limited number of confidence intervals and the naive assumption of uniform distribution of errors over lines. Hereafter, this error estimator is referred to as *line-based*.

To overcome the problems described above we proposed an innovative error estimation method. This method predicts the error rate in a block of

lines by estimating the number of edit operations for each recognised word. This method is referred to as *block-based*. Given a block of R^- recognised words, let I^- be the number of incorrect words in that block, and let E^- be the number of edit operations required to convert those recognised words into their reference. Then α can be calculated as

$$\alpha = \frac{E^-}{I^-} \tag{5}$$

which is the ratio between the number of edit operations and the number of incorrectly recognised words. The α variable is motivated by the fact that an erroneous word might cause more than one edit operation, as insertions of multiple words may occur.

Then, we can calculate the number of edit operations of E^- in Eq. 2 as

$$E^- = \alpha \mathbb{E}[I^-] \tag{6}$$

where $\mathbb{E}[I^-]$ is the expected value of incorrectly recognised words, since the reference transcription is not available.

Given a block of R^- recognised words, let $y_i \in \{0, 1\}$ be a random variable, which indicates if the word i is correct ($y_i = 0$) or incorrect ($y_i = 1$). Similarly, let $x_i \in \mathbb{R}$ be the CM of the i-th recognised word. We assume that y_i follows a Bernoulli distribution with probability $p(y_i \mid x_i)$, i.e $y_i \sim \text{Be}(p(y_i \mid x_i))$. The number of errors I^- in a block can be estimated as

$$I^- = y_1 + y_2 + \cdots + y_{R^-} \tag{7}$$

and its expected value is

$$\mathbb{E}[I^-] = \mathbb{E}[y_1] + \mathbb{E}[y_2] + \cdots + \mathbb{E}[y_{R^-}] \tag{8}$$

Then, the expected number of errors can be calculated as

$$\mathbb{E}[I^-] = \sum_{i=1}^{R^-} \mathbb{E}[y_i] = \sum_{i=1}^{R^-} p(y_i = 1 \mid x_i) \tag{9}$$

Under these assumptions, the estimated number of errors in a block of recognised words is calculated as the sum of the probabilities of each word being incorrect conditioned on its CM multiplied by α. Finally, putting Eqs. 2, 3, 6 and 9 together, the estimation of WER is

$$WER^- = \frac{\alpha \sum_{i=1}^{R^-} p(y_i = 1 \mid x_i)}{R^- \frac{R^+}{N^+}} \tag{10}$$

Obviously, the term $p(y_i = 1 \mid x_i)$ needs to be estimated in previous blocks that have been supervised. This term can be calculated simply, as

$$p(y = 1 \mid x) = \frac{N(y = 1, x)}{N(x)} \tag{11}$$

which is the frequency that words with CM x are incorrect.

However, the distribution of events $\{y, x\}$ is very sparse and we cannot estimate this posterior probability for all possible values of x. In this work, we have estimated $p(y_i = 1 \mid x_i)$ as a probability histogram, in which the domain of x is divided into a finite number of intervals.

In order to analyse the effect of the number of intervals on the accuracy of the error estimation, we performed the same experiment described in Fig. 1, exploring the number of intervals for 1,2,8 and 32 intervals of equal size. Fig. 2 presents a comparison of error estimation between block-based methods and real distribution. As observed, considering only one interval is equivalent to the mean error estimation in Eq. 4. Differently, each increment of the number or intervals results in a better estimation of the error. As observed,

considering 32 confidence intervals in the posterior calculation produces an accurate estimation of the error on the whole distribution. In practice, the number of intervals are optimised on a development set.

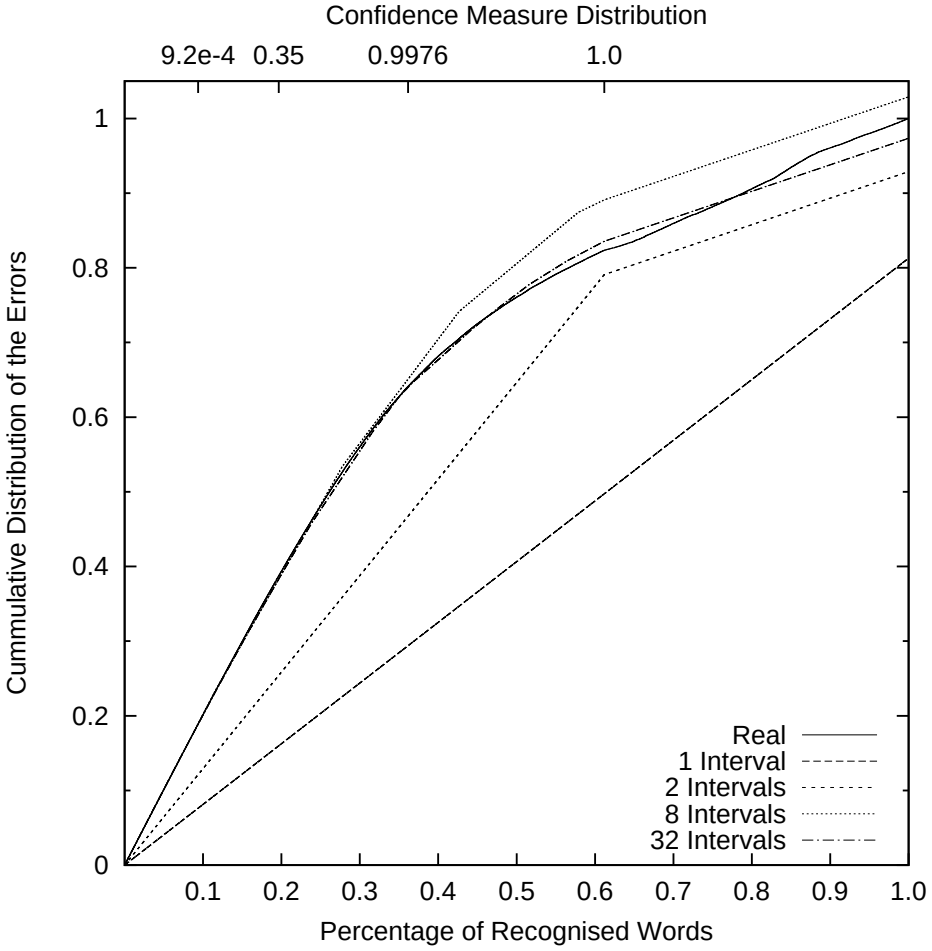

Figure 2: Cumulative distribution of errors on a set of recognised words ordered by Confidence Measure (CM). Actual error distribution is compared with the block-based estimation studying the effect of the number of intervals.

Finally, we should recall that the error estimation methods proposed so far aim at predicting how much supervision effort is required to achieve the WER

defined beforehand by the user. The more accurate the error prediction, the less user supervision effort is unnecessarily requested.

4. Experiments

Experimental results are reported on two old handwritten text documents called GERMANA (Pérez et al., 2009) and RODRIGO. Both documents were digitised and annotated by paleography experts and are freely available for research purposes. On one hand, GERMANA is a 764-page Spanish manuscript from 1981, which mostly contains written calligraphy text on well-separated lines in up to six different languages. On the other hand, RODRIGO is a 853-page manuscript completely written in Spanish. Despite the fact that its size and layout are similar to GERMANA, it comes from an older epoch (1545) and its writing style has clear Gothic influences. Table 1 shows some basic statistics of the two documents.

Table 1: Statistics of GERMANA and RODRIGO. Out-of-vocabulary words correspond to the percentage of running words, which do not appear in the training set. Perplexity is calculated using a ten-fold validation on the whole document.

	GERMANA	RODRIGO
Pages	764	853
Lines (K)	20.5	20.4
Running words (K)	217	232
Vocabulary size (K)	27.1	17.3
Out-Of-Vocabulary(%)	25.7	11.9
Character set size	115	115
Perplexity	274.1	177.1

In this paper, we performed the interactive transcription of these two documents and we compare it to a baseline, non-interactive approach. The baseline non-interactive approach (S) corresponds to an application in which a fixed quantity of user effort is used to fully transcribe the first part of a document. Then, an HTR system is trained on this first supervised part. Finally, the rest of the document is automatically transcribed with the trained HTR system. This approach is considered to be the baseline, because it is typically the first approach applied to these tasks and no form of interactive transcription is used. On the other hand, in the interactive experiments we compared two types of error estimation approaches. First, our previous line-based method for error estimation (Eq. 4). Second, the new block-based method for error estimation that we have described in Sec. 3. Furthermore, since as in our previous work dealing with error estimation hypothesis recomputation was not used, we performed an experiment to study its influence in the results.

Hypothesis recomputation was presented in our previous works, in which different strategies were tested. In this work, we employed the best performing strategy, which is called *Delayed*. In this strategy, hypothesis recomputation is performed after all user interactions with the same line have been performed. The combination of error prediction methods and hypothesis recomputation results in four different approaches: line-based (L), line-based with hypothesis recomputation (L+D), block-based (B) and block-based with hypothesis recomputation (B+D).

These four approaches were employed to interactively transcribe the document given several user-defined WER thresholds for which the system bal-

anced the supervision effort required. WER thresholds were selected taking into account the average number of words per line in both documents. GERMANA and RODRIGO lines have eleven words on average due to the fact that they have been written by a single author in well-defined templates. Then, we consider the interactive transcription of both documents when the user selects four different WER thresholds: 9% (one incorrect word per line on average), 18% (two incorrect words per line on average), 27% and 36%. It must be noted that, given that interaction with real users is expensive and our purpose is to study system behaviour for many different parameters, user supervision is simulated by means of an automatic process. Concretely, when supervising a recognised word, the simulated user performs the minimum number of edit operations according to the minimum distance path between the recognised and reference transcription. The user interaction model is explained in detail in (Serrano et al., 2009).

Due to the sequential structure of the documents, the transcription task is carried out from the beginning to the end of the document. On the one hand, in the baseline approach, we split the documents into blocks of 1000 lines. The first block is used to train an initial system from scratch and to tune the preprocessing, training and recognition parameters. All these optimised parameters remain unchanged for the rest of experiments. Then, starting from block two to the last. First, we trained a system from the first to the current block and used it to recognise the rest. Finally, we measured the WER of the resulting document, i.e. on both, the supervised and recognised parts. It must be noted that this error is a measure of the error produced by an autonomous system whose output was not been supervised.

Meanwhile, for the interactive experiments, each database was divided into 7 consecutive blocks of 3200 lines, except for the first block, which only contains 1000 lines, and the last block, which also includes the last remnant of the lines. It should be noted that the numbers of blocks is limited in our interactive experiments due to the higher computational cost compared to the baseline. The experimental setting for each database is performed as follows: the first block is devoted to train an initial system from scratch, and tune the preprocessing, training, error predicting and recognition parameters. All these optimised parameters, except for the ones related to error prediction, remain unchanged for the rest of the experiments. Starting from block two to the last block, each new block is processed as follows.

- First, the block is automatically recognised and CMs are estimated.
- Second, its recognised words are supervised according to the error estimation approach.

  Line-based approaches. For each recognised line, words are ordered by confidence. Then, from the least confident word to the highest, the system estimates the error of all unsupervised words so far considering that the current word is not going to be supervised, which will increment the previously estimated error. If the error threshold is surpassed, the word is supervised. Finally, each time a word is processed, the error prediction model parameters are updated.

  Block-based approaches. The system estimates expected error on the whole block using the method presented in Sec.3. Then, the

user supervises recognised words in order of CM, independently from the line order, until the error in the remaining words is below the defined threshold. It must be noted that, due to block segmentation of the document, the block-based approach adjust the error on the whole document by adjusting the error independently for each block. For instance, the 9% WER threshold is achieved by adjusting the WER of all blocks to 9%.

- Third, in the approaches using hypothesis recomputation. Once the user supervision is performed, the system recomputes its best hypothesis constrained to the newly supervised words and CMs are calculated again.

- Finally, once the whole block has been processed, it is added to the training set and the system is fully re-trained from the supervised and high-confidence words. At this stage, the error prediction model of the block-based approach is also trained.

Figs. 3 and 4 show the results of experiments for both corpora. On one hand, the x axis measures the quantity of supervision effort employed, which is calculated as the percentage of reference words supervised. A word is considered to be supervised once the user has been required to check it. Note that this includes the case of the supervision of correctly recognised words. On the other hand, the y axis measures the quality of the produced transcriptions in terms of WER. The imaginary diagonal of these plots would represent the manual transcription of the documents. For instance, the point at coordinates $(50, 50)$ would be the result of transcribing only 50% of the

words of the document, which would leave the rest untranscribed and it would result in 50% of WER. Similarly, the best results will correspond to a curve close to both axes, where with the minimum effort we obtain the best transcriptions.

Each curve represents the results for each of the described interactive approaches and each point of each curve represents the result of a whole experiment. For instance, the second point of the line-based approach with no hypothesis recomputation in RODRIGO corresponds to the experiment using a user-defined WER threshold of 36%. However, due the pessimistic WER prediction, the resulting WER is 27%, far below the user-defined WER threshold, and the supervision effort is 21%.

As observed, all interactive approaches obtained better results than the supervised approach. It must be noted that differences between the supervised and interactive approaches are statistically significant, as shown by a bootstrap evaluation (Efron and Tibshirani, 1994). This difference is mainly caused by the combination of active and semi-supervised learning, which intelligently selects the words that have to be supervised and then included as training data. In fact, all interactive experiments select words according to their CM, which is directly related to system uncertainty. We can also observe that, as typically happens in active learning applications (Hakkani-Tür et al., 2006), the improvement caused by active learning techniques decreases as the amount of available user supervision increases.

Although all interactive approaches efficiently employ the user effort available, there are significant differences between them. The main reason for this difference can be explained by the error prediction method. As observed in

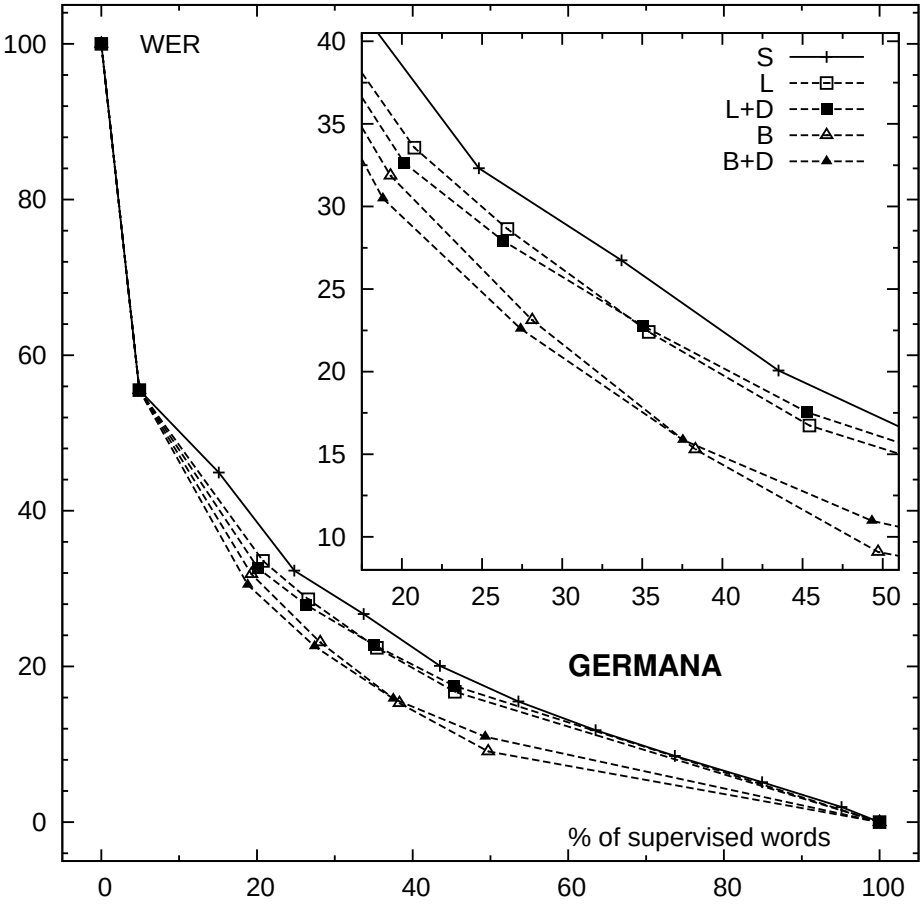

Figure 3: WER results from the interactive transcription experiments performed on the GERMANA database. Word Error Rate (WER) of the final transcriptions is shown for each approach using a limited user effort. A close-up is shown in the upper right corner depicting interactive approaches.

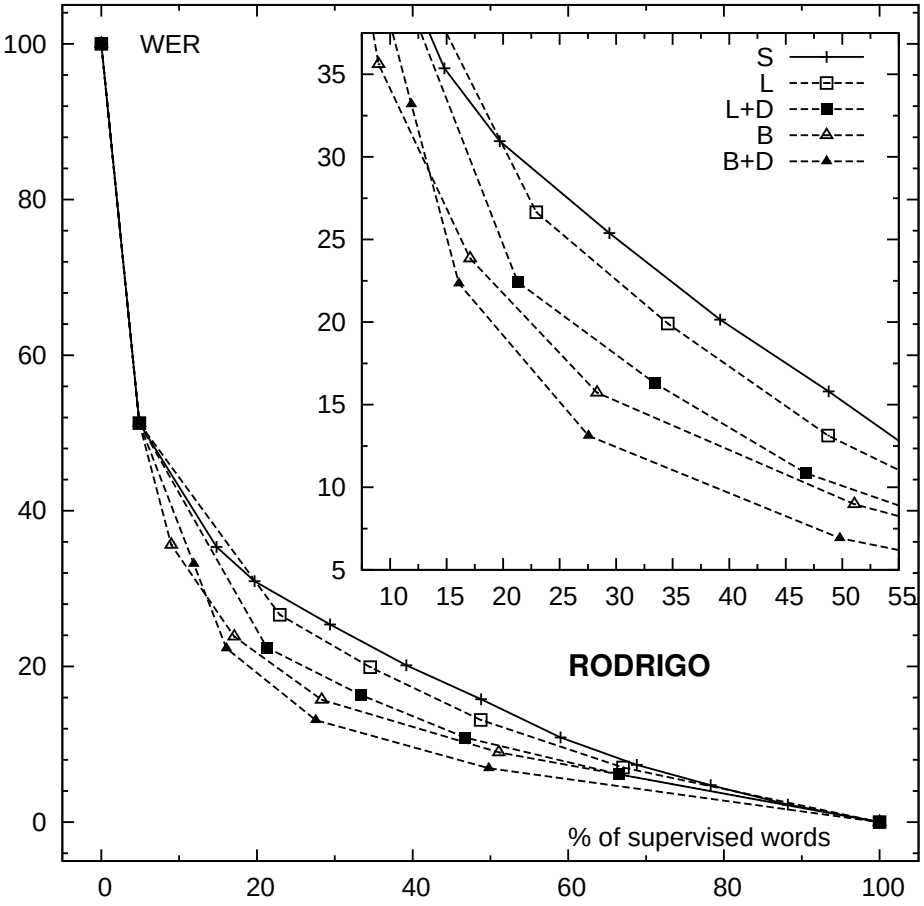

Figure 4: WER results from the interactive transcription experiments performed on the RODRIGO database. Word Error Rate (WER) of the final transcriptions is shown for each approach using a limited user effort. A close-up is shown in the upper right corner depicting interactive approaches.

both corpora, there is little difference between the supervised and the line-based approach. This is due to two problems, the ill-defined confidence intervals mentioned in Sec. 3 and the constraint of supervising words within a line.

The problems of line-based approaches were overcome by two features of the newly proposed block-based approach. First, the error estimation was significantly improved by the new estimation method. Second, word supervisions are decided at block level and not constrained to line level, so better decisions can be taken to select low confidence words inside a block.

In our experiments, as observed in Figs. 3 and 4 , the block-based approach improves upon the line-based approach in terms of both system performance and efficient use of supervision effort. For instance, when comparing the supervision effort of both approaches in RODRIGO for the same transcription error. Here, we observed that the block-based experiment for a WER threshold of 9% resulted in a transcription with about 9% WER and requires a supervision effort of 51.1%. In contrast, using the same threshold in the line-based experiment results in 7% WER and it requires a much greater quantity of supervision effort, 67%. On the other hand, when comparing the error resulting from both approaches for the same supervision effort, we observed that for a supervision effort of 22.5%, the line-based approach would obtain a transcription with 27% WER, while the block-based approach transcriptions would contain only 20%. Similar improvements can also be observed in the experiments performed in GERMANA. Again, a boostrap evaluation has shown that differences between the line-based and block-based results are statistically significant.

Figs. 3 and 4 also include the results of both approaches when hypothesis recomputation is applied. In RODRIGO, we observe that recomputation improves the results for both approaches in all the experiments performed. However, the improvement from this technique is much higher in the line-based approach, as the error on this approach is higher than the error of the block-based approach. In contrast, in GERMANA, it can be observed that hypothesis recomputation only improved the results slightly when supervision effort was lower, while it performed worse when supervision effort was higher. The main cause of this problem is the explicit blank modelling used in GERMANA to tackle the problem of out-of-vocabulary words (OOVs). In GERMANA, words are only considered when delimited by the blank character (or space). This method is able to generate some OOVs by concatenating words in the lexicon. For example, the word "natural" can be generated by the characters "n-a-t-u-r-a-l" or "n-a-t-u-r-a-l-blank". The recognition of the word "naturalmente" could be performed recognising two words: "natural", not followed by blank, and "mente" followed by blank. An additional problem of the hypothesis recomputation technique is that it is not considered in the estimation of the error in either of the methods proposed. As a result, the error on final transcriptions was below the user-defined WER threshold and thus a minor supervision effort could have been employed. A more accurate method, in which the improvement due to hypothesis recomputation is included in error estimation, remains as future work.

An additional experiment was carried out to evaluate the effectiveness of user supervision in the best performing approach, i.e. the (B+D) approach (see Figure 5). In this experiment, we performed the interactive transcription

of both documents, but considering the case in which the user adjusted the amount of user effort available rather that the WER threshold.

In this scenario, the objective of the system is to generate the best possible transcriptions given the amount of user effort available. Here we followed the same interactive approach except for the error estimation method. Instead, the decision of which words were supervised was taken by uniformly distributing the user effort available across the blocks. Then, for each block, the system asked the user to supervise the corresponding least confident words. Hence, the results obtained with this approach can be directly compared with those obtained in the previous experiments, as the only difference is the user effort applied on each block.

It should be noted that the approach presented so far in this paper applies a variable number of supervisions per block depending on the estimated error within the block. However, the latter approach uniformly distributes the user effort available among all blocks. As a result, a comparison between a fixed and a variable number of supervisions can be performed. The results of transcribing both corpora, GERMANA and RODRIGO, using the best approach (B+D) with the same error threshold, and using the previously presented fixed user effort approach (U) when supervising the first block, and $\{10\%, 20\%, 30\%, 40\%\}$ of the remainder blocks, are depicted in Fig. 5.

As observed, the curves of both approaches overlap, from which we can draw two conclusions. First, the interactive transcription approach is effective for cases in which either the error or the user effort is fixed. Secondly, even though a fixed and a variable number of supervisions per block achieved similar results in terms of WER and percentage of supervised words, there

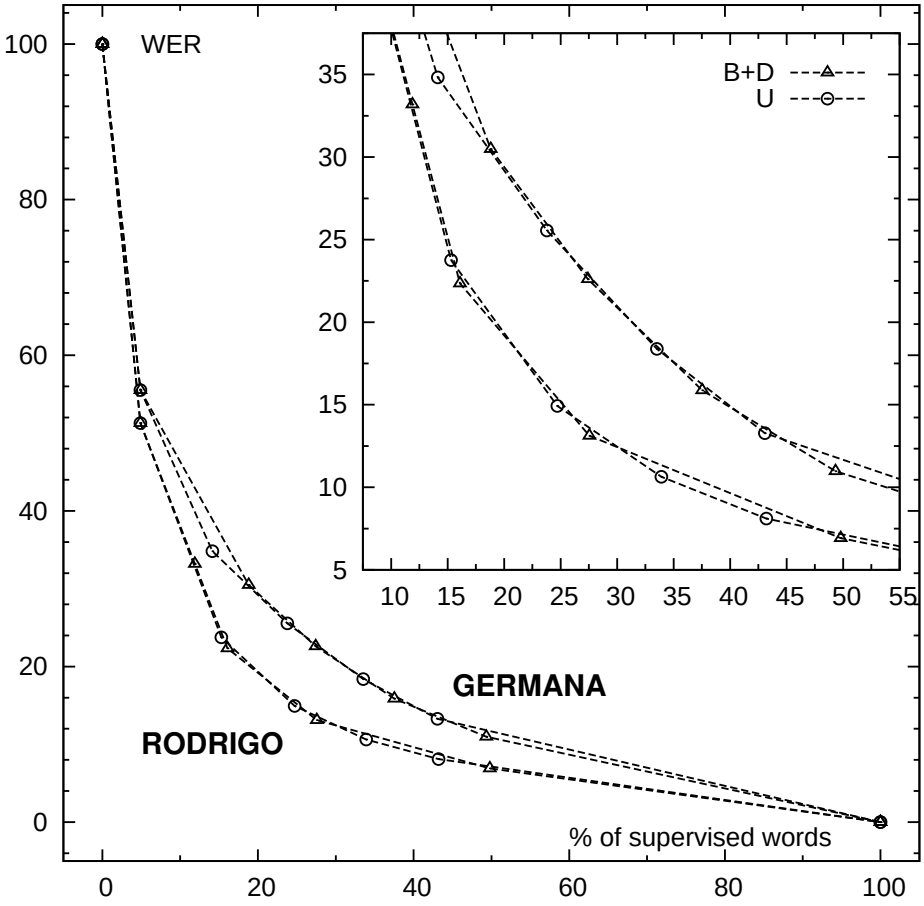

Figure 5: WER results from the interactive transcription experiments performed on the GERMANA and RODRIGO databases. Word Error Rate (WER) of the final transcriptions is shown for each approach using a limited user effort. A close-up is shown in the upper right corner depicting the results.

are notable differences in the number of incorrectly supervised words. A further analysis revealed that the method presented in this paper, i.e. variable number of supervisions, supervises more incorrect words than the uniform approach, as the supervision degree is higher for the first blocks when the system is still learning. In contrast, in the case of a fixed number of supervisions per block, when the last blocks are processed and the system is better trained, the system is more likely to ask the user to supervise correct words, which wastes the available user effort.

Finally, we observed that the accuracy of the block-based error estimation method degrades as more blocks are taken into account to compute its parameters. This is mainly caused by the data used to train these parameters, since we only consider recognised words that have been supervised by the user. However, the HTR system is continuously re-trained and thus its performance is improved. This improvement goes unnoticed by the error estimation method and it results in a pessimistic estimation in the last blocks. A solution to this problem could be to train the error estimation method using only the last n blocks.

5. Conclusions & Future Work

In this work, we have presented an interactive approach to HTR when a user-defined amount of error is tolerated. We proposed a method to estimate the WER of a set of recognised words. This method estimates the expected number of edit operations of a recognised word by calculating the expected error of a word subjected to its CM. The error estimation method is included in a CAT approach that efficiently employs user supervisions by means of active

and semi-supervised learning techniques, along with hypothesis recomputation to include user supervision as new search constraints. Experiments were performed on the transcription of two real handwritten text documents. The results obtained significantly outperformed our previous results in terms of both system performance and user effort reduction. We also measured the improvement due to hypothesis recomputation when user supervisions are performed. Hypothesis recomputation improved WER results but employed more user effort that would be required, as words corrected due to hypothesis recomputation are not considered in our error estimation method.

An additional improvement in the error estimation could be obtained by taking into he contribution of hypothesis recomputation using information theory metrics as was shown by Culotta et al. (2006). On the other hand, even though a more accurate error estimation was performed, further analysis revealed that the proposed method may be pessimistic because of the training data used. A better idea would be to make a better selection of the training data to estimate an error distribution similar to that of the next block. On the other hand, an online adaptation of the error estimation parameters each time a word is supervised could be useful in some applications and remains as future work.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287755 (transLectures). Also supported by the EC (FEDER, FSE), the Spanish Government (MICINN, MITyC, "Plan E", under grants

MIPRCV "Consolider Ingenio 2010", MITTRAL (TIN2009-14633-C03-01), iTrans2 (TIN2009-14511), and FPU (AP2007-02867), and the Generalitat Valenciana (grants Prometeo/2009/014 and GV/2010/067). Special thanks to Jesus Andrés for his fruitful discussions.

References

Culotta, A., Kristjansson, T., McCallum, A., Viola, P., 2006. Corrective feedback and persistent learning for information extraction. Artificial Intelligence 170, 1101–1122.

Efron, B., Tibshirani, R.J., 1994. An Introduction to Bootstrap . Chapman & Hall/CRC.

Grangier, D., Vinciarelli, A., Bourlard, H., 2003. Information Retrieval on Noisy Text. Technical Report. IDIAP.

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J., 2009. A novel connectionist system for unconstrained handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 855 –868.

Hakkani-Tür, D., Riccardi, G., Tur, G., 2006. An active approach to spoken language processing. ACM Transactions on Speech and Language Processing 3, 1–31.

Kristjannson, T., Culotta, A., Viola, P., McCallum, A., 2004. Interactive information extraction with constrained conditional random fields, in: Proc.

of the 19th Nat. Conf. on Artificial Intelligence (AAAI 2004), San Jose, CA (USA). pp. 412–418.

Matusov, E., Kanthak, S., Ney, H., 2006. Integrating speech recognition and machine translation: Where do we stand?, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France.

Navarro-Cerdan, J.R., Arlandis, J., Perez-Cortes, J.C., Llobet, R., 2010. User-defined expected error rate in OCR postprocessing by means of automatic threshold estimation, in: Proc. of the 2010 12th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR 2010), Washington, DC, USA. pp. 405–409.

Pérez, D., Tarazón, L., Serrano, N., Ramos-Terrades, O., Juan, A., 2009. The GERMANA database, in: Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009), Barcelona (Spain). pp. 301–305.

Revuelta-Martínez, A., Rodríguez, L., García-Varea, I., 2012. A computer assisted speech transcription system, in: Proc. of the 13th Conf. of the European Chapter of the Association for computational Linguistics (EACL 2012), pp. 41–45.

Roy, B., Vosoughi, S., Roy, D., 2010. Automatic estimation of transcription accuracy and difficulty, in: Proc. of INTERSPEECH 2010, pp. 1902–1905.

Sánchez-Cortina, I., Serrano, N., Sanchis, A., Juan, A., 2012. A prototype for interactive speech transcription balancing error and supervision effort, in: Proc. of the 2012 ACM Int. Conf. on Intelligent User Interfaces (IUI 2012), pp. 325–326.

Sánchez-Sáez, R., Leiva, L.A., Sánchez, J.A., Benedí, J.M., 2010. Interactive predictive parsing using a web-based architecture, in: Proc. of The 11th Annual Conf. of the North American Chapter of the Association for Computational Linguistics Demonstration Session (HLT-DEMO 2010), Stroudsburg, PA, USA. pp. 37–40.

Sanchis, A., Juan, A., Vidal, E., 2012. A word-based naïve bayes classifier for confidence estimation in speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 20, 565–574.

Schlapbach, A., Wettstein, F., Bunke, H., 2008a. Automatic estimation of the readability of handwritten text, in: Proc. of the 16th European Conf. on Signal Processing (EUSIPCO 2008), pp. 2–6.

Schlapbach, A., Wettstein, F., Bunke, H., 2008b. Estimating the readability of handwritten text - a support vector regression based approach., in: Proc. of the 20th Int. Conf. on Pattern Recognition (ICPR 2008), pp. 1–4.

Serrano, N., Giménez, A., Sanchis, A., Juan, A., 2010a. Active Learning Strategies for Handwritten Text Transcription, in: Proc. of the 12th Int. Conf. on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010), Beijing (China).

Serrano, N., Pérez, D., Sanchis, A., Juan, A., 2009. Adaptation from Partially Supervised Handwritten Text Transcriptions, in: Proc. of the 11th Int. Conf. on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009), Cambridge, MA (USA). pp. 289–292.

Serrano, N., Sanchis, A., Juan, A., 2010b. Balancing error and supervision effort in interactive-predictive handwriting recognition, in: Proc. of the 15th Int. Conf. on Intelligent User Interfaces (IUI 2010), Hong Kong (China). pp. 373–376.

Serrano, N., Tarazón, L., Pérez, D., Ramos-Terrades, O., Juan, A., 2010c. The GIDOC prototype, in: Proc. of the 10th Int. Workshop on Pattern Recognition in Information Systems (PRIS 2010), Funchal (Portugal). pp. 82–89.

Settles, B., 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van, C., Meteer, M., 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics 26, 339–373.

Toselli, A.H., V.Romero, Rodríguez, L., Vidal, E., 2007. Computer Assisted Transcription of Handwritten Text, in: Proc. of the 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007), Curitiba (Brazil). pp. 944–948.

Tur, G., Hakkani-Tür, D., Schapire, R.E., 2005. Combining active and semi-supervised learning for spoken language understanding. Speech Communication 45, 171 – 186.

Wang, W., Zhou, Z.H., 2008. On multi-view active learning and the combi-

nation with semi-supervised learning, in: Proc. of the 25th Int. Conf. on Machine learning (ICML 08), pp. 1152–1159.

Wessel, F., Schlüter, R., Macherey, K., Ney, H., 2001. Confidence measures for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing 9, 288–298.

Yoon, S.Y., Chen, L., Zechner, K., 2010. Predicting word accuracy for the automatic speech recognition of non-native speech, in: Proc. of INTERSPEECH 2010, pp. 773–776.

Zhou, Z.H., Chen, K.J., Dai, H.B., 2006. Enhancing relevance feedback in image retrieval using unlabeled data. ACM Transactions on Information Systems 24, 219–244.

Zhu, X., 2006. Semi-Supervised Learning Literature Survey. Technical Report. Computer Sciences, University of Wisconsin-Madison.