Chapter 6

Interactive Pattern Recognition

Lately, the paradigm for Pattern Recognition (PR) systems design is shifting from the concept of full automation to schemes where the decision process is conditioned by human feedback. This is is motivated by the fact that many applications are expected to assist rather than replace human work; think for instance of systems for medical diagnosis or traffic control.

In this chapter, as an alternative to reviewing (or post-editing) the automatic output of PR systems, an interactive approach is proposed, where the human is placed "in the loop". This scenario leads the system to being able to leverage implicit information from user interactions, and use this information to improve its performance. Interactivity naturally entails multimodal operations, offering opportunities for even greater usability improvements. Multimodality arises when additional feedback signals are non-deterministic and, consequently, need to be decoded. Finally, interactivity offers an ideal framework for adaptive learning, which is expected to lead to further improvements in both performance and usability.

Chapter Outline

6.1 Introduction

Novel interfaces with high cognitive capabilities is a hot research topic that aims at solving challenging application problems in our society of information technology. The outstanding need for the development of such interactive systems is clearly reflected, for instance, in the MIPRCV[1] project, where these cognitive capabilities are included as one of the priority research challenges. Placing Pattern Recognition (PR) within an HCI framework requires changes to the way we look at problems in these areas [Vidal et al., 2007]. Classical PR minimum-error performance criteria should be complemented with better estimations of the amount of effort that the interactive process will demand from the user. As such, current existing PR techniques, which are intrinsically grounded on error-minimization algorithms, need to be revised and adapted to the new, minimum human-effort performance criterion.

Mining implicit data from user interactions provides research with a series of challenges and opportunities in order to rethink how Interactive PR approaches (IPR for short) may drive the dynamic environment of interactive systems. In this context, implicit interaction entails three types of opportunities in IPR:

- Feedback information derived from the interaction process can be used to significantly improve system performance.
- Interaction feedback signals are intrinsically multimodal, which means that we can study the synergy among different input modalities to enhance overall system behavior and usability.
- Each interaction generally yields ground-truth data, which can be advantageously used as valuable adaptive training data and tune system performance.

It should be noted that multimodal interaction may support two types of multimodality [Toselli et al., 2011]. One corresponds to the input signal itself, which can be a complex mixture of different data types, ranging, e.g., from conventional keystrokes to audio and video data streams. The other type, more subtle but also important, is derived from the often different nature of input and feedback signals. It is this second type the one that makes both multimodality and implicit interaction an inherent feature of human behavior.

Overall, the IPR framework proposes a radically different approach to correct the errors committed by a PR system. This approach is characterized by human and machine being tied up in a much closer loop than usually. That is, the user gets involved not only after the system has completed the production of its final recognition result, but also during the recognition process

[1] http://miprcv.iti.upv.es

itself. This way, errors can be avoided beforehand and correction costs can be dramatically reduced. Historically, this interactive-predictive approach was proposed by the so-called *conversation theory* from cybernetics, in which the system constructs its knowledge by means of a series of user interactions [Pask, 1975]. Currently, the Machine Learning community has renamed this approach to *corrective feedback* [Culotta et al., 2006], since every time the user amends an error, the system reacts by modifying the resulting hypothesis.

6.1.1 IPR Framework Overview

The IPR framework (Figure 6.1) is explained as follows [Vidal et al., 2007]:

- $\mathcal{X}$ is the system's input domain; i.e., the domain where input stimuli, observations, signals, or data come from.
- $\mathcal{H}$ is a theoretically infinite set of possible system outputs, results, or hypotheses. $h \in \mathcal{H}$ is a hypothesis which the system derives from a certain input $x \in \mathcal{X}$.
- $\mathcal{F}$ is the domain were feedback signals come from. $f(h, x)$, or just $f \in \mathcal{F}$ is a specific feedback signal which the user provides as a response to the system hypothesis h.
- $\mathcal{M}$ is any model which the system uses to derive its hypotheses.

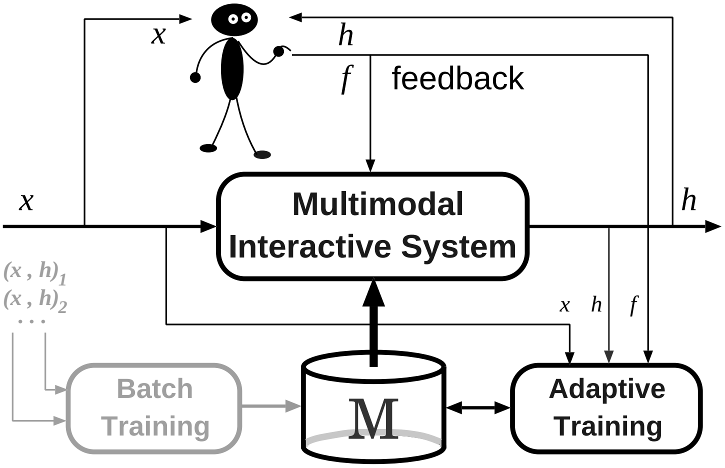


Figure 6.1: The IPR framework [Vidal et al., 2007]. Reproduced with permission.

Assume for simplicity that both the input x and the feedback f are unimodal. Interaction leads to the following modality fusion problem[2]:

$$\hat{h} = \arg\max_h \Pr(h|x, f) = \arg\max_h \Pr(x, f|h) \cdot \Pr(h) \tag{6.1}$$

In many applications x and f can be assumed to be independent given h. This allows for a naïve Bayes decomposition:

$$\hat{h} \approx \arg\max_h P_{\mathcal{M}_\mathcal{X}}(x|h) \cdot P_{\mathcal{M}_\mathcal{F}}(f|h) \cdot P_{\mathcal{M}_\mathcal{H}}(h) \tag{6.2}$$

Then, independent models, $\mathcal{M}_\mathcal{X}$, $\mathcal{M}_\mathcal{F}$ and $\mathcal{M}_\mathcal{H}$, can now be estimated separately for the input components and for the prior hypotheses distribution, respectively. This way, the resulting search problem accounts for the joint optimization of the conditional probability product.

6.1.2 Interaction Protocol

In the context of this thesis, the IPR framework has been successfully applied to four different IPR systems, where implicit interaction plays a crucial role: *1*) Handwritten Transcription, *2*) Machine Translation, *3*) Grammatical Parsing, and *4*) Image Retrieval. Indeed, the role of implicit interaction is crucial because the user can interact with an IPR system in an unimaginable number of ways. As such, the range of interaction possibilities has to be delimited or predicted in some way, so that the system can take maximum advantage of the expected user feedback. This leads to the creation of a user model, also known as an *interaction protocol*.

Depending on the application and the input modalities involved, very different types of protocols can be assumed for the user to interact with the system in a comfortable and productive way. But the chosen protocols must also allow an efficient implementation, because interactive processing is generally highly demanding in terms of response times [Toselli et al., 2011]. Eventually, the design of an efficient interaction protocol *and* an adequate UI are the most sensible design tasks for an IPR application. Concretely, once a specific interaction protocol is defined, it should be possible to apply decision theory in order to model the expected interaction effort of such protocol in terms of an adequate loss function. This would allow to search for a corresponding decision function that minimizes the loss; i.e., the expected interaction effort.

Within the two general types of interaction protocols identified in IPR [Toselli et al., 2011], we will focus in the *passive protocol*, that is, where the system requires human feedback to emit a hypothesis. This focus is motivated by the

[2] True probabilities are denoted as $\Pr(\cdot)$, while $P_\mathcal{M}(\cdot)$ or just $P(\cdot)$ denote probabilities computed with some model $\mathcal{M}$.

fact that it is a suitable scenario in which the system can take advantage of implicit interactions to a great extent. In contrast, under the *active protocol* it is the system, rather than the human, which is in charge of making the relevant decisions about the need of supervising errors. Clearly, this scenario is not as advantageous as the previous one to illustrate the role of implicit interactions in IPR.

In general, the way of interacting with an IPR system following a passive protocol is described as follows:

1. The system automatically proposes a draft of the output of the task; e.g., a text transcription or a collection of images.
2. The user then validates the parts of the output which is error-free; e.g., indicating the correct prefix in a text-oriented task or selecting those images considered as relevant in image retrieval.
3. The system then suggests a suitable, new extended consolidated hypothesis based on the previously validated parts and implicit information derived from user feedback.
4. Steps 2 and 3 are iterated until a final, perfect output is produced.

In the following sections we delineate a series of real-world implementations of the MIPR framework.

6.2 IPR Systems Overview

The following prototypes are focused on an interactive-predictive strategy, fully integrating the user knowledge into the PR process. The prototypes have been classified into two categories, depending whether the user feedback comes in the form of *structured input* or not. The former category includes three examples of Natural Language Processing (NLP) systems, where the order in which errors are corrected is determinant for the system. The latter category includes as an example an image retrieval system, where the user feedback comes in the form of *desultory input*, i.e., the order is not determinant for the system.

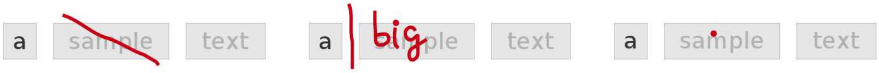
It is worth pointing out that these prototypes were not intended to be production-ready applications. Rather, they were developed to provide an intuitive interface which aims at showing the functionality of an IPR system in general, as well as illustrating the role of implicit interaction in particular.

6.2.1 Structured Input

In these systems, the user validates the longest prefix of the system hypothesis (e.g., a text transcription, a speech utterance, etc.) which is error-free. Such a validation can be performed by using, e.g., a keyboard, a computer mouse, a touchscreen, a microphone, or an e-pen. Once the first error is corrected, the system predicts the most probable continuation of the partial input. This new extended hypothesis is strongly based on the previously validated prefix and the decoding of the corrections submitted by the user—for instance, if an e-pen was used to write down a word, those pen strokes must be decoded. For the sake of simplicity, let us assume in this subsection that the system is producing text-based hypotheses; for instance, transcriptions, translations, or parse trees.

As observed, under this protocol, the user is asked to correct the first error found. Then, the system can make the reasonable assumption that the user is reading the text form left to right (or vice versa for right-to-left languages, such as Arabic). With this assumption, the search process of the next (best) hypothesis is constrained to a smaller subset of words regarding the initial hypothesis, which allows the system to make a better prediction. Moreover, this assumption allows to automate the evaluation of these IPR systems, by simulating a user that will perform a series of error amendments in an ordered sequence.

However, the role of implicit HCI has much to offer to this protocol, as the system can place a series of (safe) constraints to improve its hypotheses even further. For instance, some editing operations are expected to be performed by the user beyond simple word substitution, e.g., insertion, deletion, or rejection (Figure 6.2). More specifically, when the user is going to insert (or delete) a word, the system can assume that the word at the right of the insertion (or deletion) is correct. This constrains to an even smaller subset of words regarding the previous hypothesis, and therefore it is expected that the next prediction will be much better, since the system has more information that is implicitly validated. Going further, to replace an incorrect word the user needs to place the cursor over a text field and then start typing the corrected word. Nevertheless, this information about cursor placement can be leveraged to emit the next hypothesis *before* the user starts typing, offering thus a (hopefully) better proposal, if not the one the user had in mind.



(a) Delete **(b)** Insert **(c)** Reject

Figure 6.2: Examples of editing operations in IPR systems. By deleting or inserting a word, the system can assume that the neighboring words are implicitly correct, allowing thus for a better prediction in the next hypothesis. By making a rejection, though, the system can only assume that the word at the left is correct.

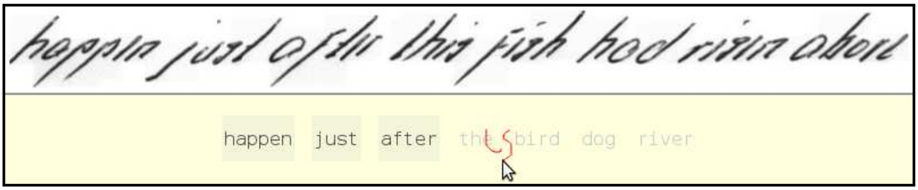
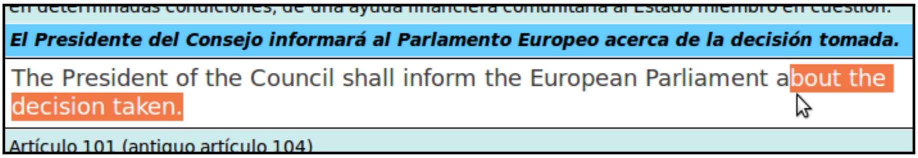


Figure 6.3: Interactive Handwritten Transcription prototype, an example of structured input. Some word-level editing operations that can be performed are substitution (shown in the image), insertion (6.2a), deletion (6.2b), or rejection (6.2c). In this example, the system assumes that the first 3 words plus the first 2 characters of the edited word are correct. This information is used to 1) decode the submitted pen strokes and 2) predict a suitable continuation of the implicitly validated segment: “happen just after this fish $\cdots$”. Prototype available at http://cat.iti.upv.es/iht/.



en determinadas condiciones, de una ayuda financiera comunitaria al Estado miembro en cuestión.

El Presidente del Consejo informará al Parlamento Europeo acerca de la decisión tomada.

The President of the Council shall inform the European Parliament about the decision taken.

Artículo 101 (antiguo artículo 104)

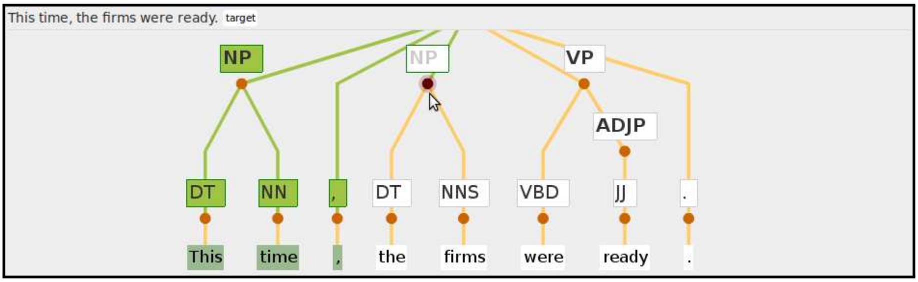
Figure 6.4: Interactive Machine Translation prototype, another example of structured input. Here, the system predicts a new hypothesis when the cursor is positioned over an erroneous character, before the user starts typing. As in Figure 6.3, the text at the left of the cursor is considered to be correct. Available at http://cat.iti.upv.es/imt/.

6.2.2 Desultory Input

As in the previous passive protocol, here the user is expected to supervise the system hypotheses in order to achieve a high-quality result. However, in this case the user can perform the amendments in a desultory order. This is especially useful when the elements of the output do not have a particular hierarchy. Many different scenarios can fall under this category. However, here we analyze the case of information retrieval, where the user initially submits a natural language description of an object she is looking for.

Under this protocol, the system outputs a set of objects matching the submitted query, so the user can select which ones fit her needs and which do not. The system then tries to fill the set with new objects taking into account the user preferences from the previous iterations. The procedure stops when the user chooses not to reject any further object from the set. The goal is to obtain such a set in the minimum number of interactions.

In image retrieval this protocol is known as *relevance feedback*, since the user typically categorizes the presented images into two (sometimes three) classes: relevant and non-relevant (and neutral in some cases). The role of implicit interaction in this scenario is particularly useful to unburden the user from having to think whether a particular image should be classified as non-relevant or neutral. As such, it is much easier for the user just to indicate which images



This time, the firms were ready. target

Figure 6.5: Interactive Grammatical Parsing prototype, an example of two-dimensional structured input. The same editing operations presented in Figure 6.2 can be performed. Tree visiting order is left-to-right depth-first, so resulting nodes at the left of (and above) the cursor are considered to be correct. Available at http://cat.iti.upv.es/ipp/.

are relevant. The system then classifies the rest of presented images into non-relevant, e.g., if they are very different to the ones the user has selected, or into neutral otherwise. Moreover, this strategy allows to automate the evaluation of these image retrieval systems, by simulating a user that will select only images considered as relevant in each iteration with the system.

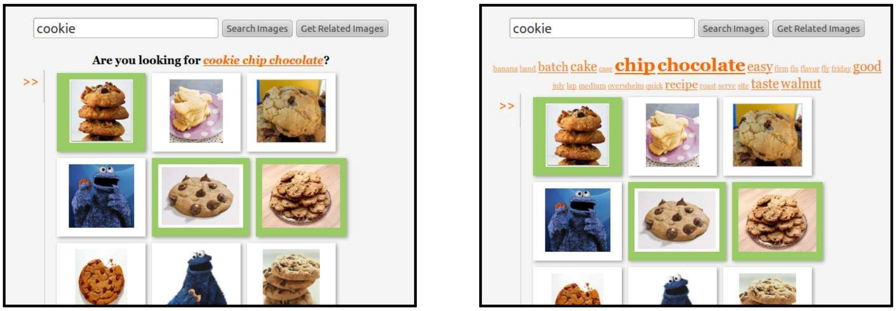
Again, the role of implicit HCI has much to offer to this protocol, as the system can take some initiative derived from user input to improve its hypotheses even further. For instance, using metadata from the presented images, it is possible to suggest a textual query that would allow the user to retrieve better images from scratch. In addition, the system can present a tag cloud to provide the user with a gist of the current set of images. Furthermore, when clicking on a tag, the system can refine the original query by adding the respective tag (or related information thereof) to the query.

6.3 Evaluation

Here we will focus on the evaluation of the IPR framework with real users. The IPR literature uses test-set-based estimates of user effort reduction, but only a few researchers have conducted controlled lab studies to verify whether the IPR framework proves to be superior to current baselines techniques [Alabau et al., 2012; Leiva et al., 2011a,b]. From the four applications previously examined, we will focus on three of them, which are the most mature technologies implemented so far.

6.3.1 Interactive Handwritten Transcription

The goal of this evaluation was aimed at improving Handwritten Text Recognition (HTR) technology. An Interactive Handwriting Transcription (IHT) system was used on a real-world task, and compared to a manual approach as


cookie
Search Images
Get Related Images
Are you looking for cookie chip chocolate?
>>
cookie
Search Images
Get Related Images
banana band batch cake case chip chocolate easy firm fix flavor fly friday good
july lap medium overwhelm quick recipe roast serve site taste walnut
>>

(a) Query refinement suggestion

(b) TagCloud based suggestion

Figure 6.6: Interactive Image Retrieval prototype, an example of desultory input. The user must select which images are relevant, in no particular order, and the system will mark the rest as non-relevant or neutral; depending on the considered image features. Moreover, this information can be used to make suggestions to the user, e.g., as a refined query (6.6a) or as a tag cloud (6.6b), which may help to disambiguate intent or to retrieve hopefully better images. Available at http://risenet.iti.upv.es/.

baseline. We compiled a test corpus from a 19th century handwritten document identified as "Cristo Salvador" (CS), which was kindly provided by the Biblioteca Valenciana Digital[3].

Participants Fourteen users from our Computer Science department volunteered to cooperate, aged 28 to 61 (M=37.3). Most of them were knowledgeable with handwriting transcription tasks, although none was a transcriber expert. One user could not finish the evaluation, so the end user sample was 13 subjects (3 females).

Assessment Measures We used two well-known objective test-set-based measures: word error rate (WER)[4] and word stroke ratio (WSR)[5], both normalized by the number of words in the reference transcription. We also measured the time needed to transcribe completely each page with each HTR system. Additionally, we measured the *probability of improvement* (POI), which estimates if a system is *a priori* better than another for a given user [Bisani and Ney, 2004].

Design We carried out a within-subjects repeated measures design. We tested two conditions: transcribing a page with the manual and the IHT system, taking into account that each one was tested twice—to compensate the above-mentioned learnability bias. We used the (non-parametric) two-sample Kolmogorov-Smirnov test, since normality assumptions did not hold.

[3] http://bv2.gva.es/

[4] WER is the minimum number of editing operations to achieve the target transcription.

[5] WSR is the number of interactions needed to achieve the target transcription.

Apparatus We modified an IHT web-based prototype [Romero et al., 2009] to carry out the field study. We implemented two HTR engines to assist the document transcription on the same UI. In addition, a logging mechanism was embedded into the web application. It allowed us to register all user interactions at a fine-grained level of detail (e.g., keyboard and mouse events, client/server messages exchanging, etc.). Then, interaction log files were reported in XML format for later postprocessing.

Procedure Participants accessed the web-based application via a special URL that was sent to them by email. In order to familiarize with the UI, users informally tested each transcription engine with some test pages, different from the ones reserved for the real test. Then, people transcribed the two user-test pages with both transcription engines. These pages were selected according to their WER and WSR values, which were close to the median values of the test-set. To avoid possible biases due to human learnability, the first page (#45) was initially transcribed with the manual engine first; then the order was inverted for the second page (#46). Finally, participants filled out an online System Usability Scale (SUS) questionnaire [Brooke, 1996] for both systems. Such an online form included a text field to allow users submit free comments and ideas about their testing experience, as well as insights about possible enhancements and/or related applicability.

Results

In sum, we can assert that regarding *effectiveness* there are no significant differences, as expected, i.e., users can achieve their goals with any of the tested systems. However, in terms of *efficiency* the IHT system is the better choice. Regarding to *user satisfaction*, IHT again seems to be the most preferable option.

Quantitative Analysis Table 6.1 summarizes the main findings. We must emphasize that the daily use of any system designed to assist handwriting transcription would involve not having seen previously any of the pages (i.e., users would usually read a page once and at the same time they would transcribe it). Therefore, IHT seems to be slightly better than a manual approach in terms of WER, and clearly superior in terms of WSR.

Analysis of Task Completion Time We observed that, overall, there are no differences in transcription times [$D = 0.16$, $p = .75$, n.s.]. In general, the system used in second place always achieved the best time, because the user already knew the text. The remarkable result is that when the user reads a page in first place the chosen engine is not determinant, because one must spend time to accustom to the writing style, interpreting the calligraphy, etc. In this case the POI of IHT with respect to the manual engine is 53%.

System		Time	WER	WSR
Overall	Manual	11.1 (3.5)	8.6 (8.2)	97.8 (6.0)
	IHT	10.3 (3.7)	6.5 (3.7)	30.4 (6.1)
Page 45	Manual	12.8 (3.5)	12.8 (9.5)	97.3 (7.0)
	IHT	8.6 (3.2)	7.0 (4.1)	28.6 (4.1)
Page 46	Manual	9.4 (2.9)	4.1 (2.0)	98.4 (4.6)
	IHT	12.0 (3.4)	6.0 (3.3)	32.1 (7.1)

Table 6.1: Mean (and SD) per page for the measured variables: time (in minutes), WER (in %), and WSR (%).

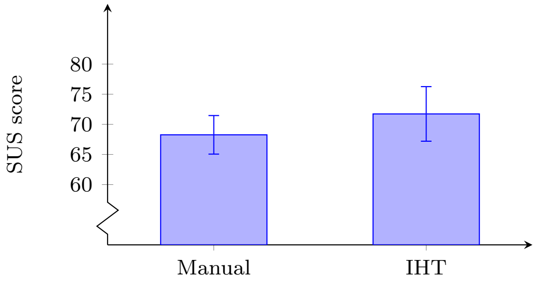
Analysis of WER Overall, IHT performs better regarding WER [$D = 0.11$, $p = .99$, n.s.]. Although the differences are not statistically significant, the interesting observation is that IHT is the most stable of the systems—even better than when using the manual engine on an already read page. We must recall that the more stable in WER a system is, the fewer residual errors are expected and therefore a high quality transcription is guaranteed. In this case, considering the first time that the user reads a page, the POI of the IHT engine over the manual engine is 69%.

Analysis of WSR Interestingly, the WSR when using the manual engine was below 100%, since there are inherent errors (some users were unable to read all lines correctly). This means that some users wrote less words in their final transcriptions than they really should have to when using the manual engine. In both conditions IHT was the best performer, and differences were found to be statistically significant [$D = 1, p < .001$]. The POI of the IHT engine regarding the manual engine is 100%. This means that the number of words a user must write and/or correct under the IHT paradigm is always much lower than with a manual system. Additionally, this fact increases the probability of achieving a high-quality final transcription, since users perform fewer interactions and are prone thus to less errors.

Qualitative Analysis Regarding user subjectivity, the SUS scores could be considered normally distributed. Thus, a Welch two-sample t-test was employed to measure the differences between both groups. We observed a tendency in favor to IHT [$t(22) = 0.25$, $p = .80$, n.s.], since users generally appreciate the guidance of the IHT system to suggest partial predictions, considering the difficulty of the task proposed in the field study.

Limitations of the Study First, taking into account that our participants were not experts in transcribing ancient documents, a dispersed behavior was expected (i.e., some users were considerably faster/slower than others). Second, the pages were really deteriorated, making more difficult the reading for the users. For that reason, there is a strong difference between the first time that

Figure 6.7: User satisfaction, according to the SUS questionnaire. Error bars denote 95% confidence intervals.



a user had to transcribe a page and subsequent attempts. Third, most of the participants had never faced neither any of the implemented engines nor the web UI before the study, so it is expected a learning curve prior to using such systems in a daily basis. Finally, a simplified starting level would minimize this effect for the task; however we tried to select a scenario as close as possible to a realistic setting.

Evaluation Discussion

Despite of the above mentioned limitations, there is a comprehensible tendency to choose the IHT paradigm over the manual system. Additionally, as observed, the probability of improvement of an IHT engine over manual transcription revealed that the interactive-predictive paradigm worked better for all users.

The advantage of IHT over traditional HTR post-editing approaches goes beyond the good estimates of human effort reductions achieved. When difficult transcription tasks with high post-editing effort are considered, expert users generally refuse to post-edit conventional HTR output [Toselli et al., 2009]. In contrast, the proposed interactive approach constitutes a much more natural way of producing correct texts. With an adequate user interface, IHT lets the user be dynamically in command: if predictions are not good enough, then the user simply keeps typing at her own pace; otherwise, she can accept (partial) predictions, thereby saving both thinking and typing effort.

6.3.2 Interactive Machine Translation

The goal of this evaluation was aimed to assess a Machine Translation (MT) system that is based on the IPR framework (IMT for short), and compare it to a state-of-the-art post-editing (PE) MT system. Translating manually from scratch was not considered, since this practice is being increasingly displaced by assistive technologies at present. Indeed, PE of MT systems is found frequently in a professional translation workflow [TT2, 2001].

Initially, we modified an IMT web-based prototype [Ortiz-Martínez et al., 2010] to carry out the evaluation. We targeted specific IMT features, e.g., confidence measures in translated words or click-based operations. We will refer to this system as the advanced version (IMT-AV).

Evaluation of an Advanced Version

In addition to IMT-AV, a post-editing version of the prototype (PE-AV) was developed to make a fair comparison with state-of-the-art PE systems. PE-AV used the same interface as IMT-AV, but the IMT engine was replaced by autocompletion-only capabilities, as it is found in popular text editors.

Participants A group of 10 users (3 females) aged 26–43 from our research group volunteered to perform the evaluation as non-professional translators. They were proficient in Spanish and had an advanced knowledge of English. While none of them had worked with IMT systems before, all knew the theoretical foundations of the technology.

Assessment Measures Both systems were evaluated on the basis of the ISO 9241-11 standard (ergonomics of human-computer interaction). Three aspects were considered: efficiency, effectiveness, and user satisfaction. For the former, we computed the average time in seconds that took to complete each translation. For the second, we evaluated the BLEU[6] against the reference and a crossed multi-BLEU among users' translations. For the latter, we formulated 10 questions inspired by the system usability scale (SUS) questionnaire. Users would answer the questions in a 5-point Likert scale (1: strongly disagree, 5: strongly agree), plus a text area to submit free-form comments.

Apparatus Since participants were Spanish natives, we decided to perform translations from English to Spanish. We chose a medium-sized corpus, the EU corpus, typically used in IMT [Barrachina et al., 2009], which consists of around 200K sentences from legal documents. We built a glossary for each source word by using the 5-best target words from a word-based translation model. We expected this would cover the lack of knowledge for our non-expert translators toward this particular task. In addition, a set of 9 keyboard shortcuts was designed, aiming to simulate a real translation scenario.

Furthermore, autocompletion was added to PE-AV, i.e., words with more than 3 characters were autocompleted using a task-dependent word list. In addition, IMT-AV was set up to predict at character level interactions. We disabled complementary features for the evaluation to focus on basic IMT.

Procedure Three disjoint sentence sets (C1, C2, C3) were randomly selected from the test dataset. Each set consisted of 20 sentence pairs and kept the sequentiality of the original text. Sentences longer that 40 words were discarded. C3 was used in a warm up session, where users gained experience with the IMT system (5–10 min per user on average) before carrying out the actual

[6] BLEU is a standard measure of the quality of machine-translated text.

	PE-AV	IMT-AV
Avg. time (s)	62 (SD=51)	67 (SD=65)
BLEU	40.7 (13.4)	41.5 (13.5)
Crossed BLEU	77.4 (4.5)	78.9 (4.8)
User Satisfaction	2.5 (1.2)	2.1 (1.2)

Table 6.2: Summary of the results for the first test.

evaluation. Then, C1 and C2 were evaluated by two user groups (G1, G2) in a counterbalanced fashion: G1 evaluated C1 on PE-AV and C2 on IMT-AV, while G2 did C1 on IMT-AV and C2 in PE-AV.

Results Although results were not strongly conclusive (there were no statistical differences between groups), some trends were observed. First, time spent per sentence (efficiency) on average in IMT was higher than in PE (67 vs. 62 s). However, effectiveness was slightly higher for IMT in BLEU with respect to the reference sentence (41.5 vs. 40.7) and with respect to a cross-validation with other user translations (78.9 vs. 77.4). This suggested that the IMT system helped to achieve more consistent and standardized translations.

Finally, users perceived the PE system to be more adequate than the IMT system, although global scores were 2.5 for PE and 2.1 for IMT, which suggested that users were not very comfortable with none of the systems (Likert scores were comprised between 1 and 5). IMT failed to succeed in questions regarding the system being easy to use, consistent, and reliable. This was corroborated by the submitted comments.

Users complained about having too many shortcuts and available edit operations, some operations not working as expected, and some annoying common mistakes regarding predictions of the IMT engine (e.g., inserting a whitespace instead of completing a word, which would be interpreted as two different words by the UI). One user stated that the PE system "was much better than the [IMT] predictive tool". Regarding PE, users mainly questioned the usefulness of the autocompletion feature.

Evaluation of a Simplified Version

Results from the first evaluation were quite disappointing. Not only participants took more time to complete the evaluation with IMT-AV, but they also perceived that IMT-AV was more cumbersome and unreliable than PE-AV. However, we still observed that IMT-AV had been occasionally beneficial, and probably the bloated UI was the cause for IMT to fail. Thus, we developed a simplified version of the original prototype (IMT-SV).

Participants Fifteen participants aged 23–34 from university English courses (levels B2 and C1 from the Common European Framework of Reference for Languages) were paid to perform the evaluation (5 euro each). A special price of 20 euro was given to the participant who would contribute with the most useful comments about both prototypes. It was found that, by following this method, participants were more verbose when it came to reporting feedback.

Apparatus In this case, the editing interface was presented as a simple text area. In addition, the editing operations were simplified to allow only word substitutions and single-click rejections. Besides, we expected that the simplification of the interface logic would reduce some of the programming bugs that bothered users in the first evaluation. The PE interface was simplified the same way (PE-SV). Furthermore, the autocompletion feature was improved to support n-grams of arbitrary length. A different set of sentences (C1′, C2′, C3′) was randomly extracted from the EU corpus.

Procedure To avoid possible bias regarding which system was being used, sentences were presented in random order, and engine type was hidden to participants. As a consequence, users could not evaluate each system independently. Therefore, a reduced questionnaire with just two questions was shown on a per-sentence basis. **Q1** asked if system suggestions were useful. **Q2** asked if the system was cumbersome to use overall. A text area for submitting free-form comments was also included in the UI.

Results Still with no statistical significance, we found that IMT was perceived now better than PE. First, interacting with IMT-SV was more efficient than with PE-SV on average (55 s vs. 69 s). The number of interactions was also lower (79 vs. 94). Concerning user satisfaction, IMT-SV was perceived as more helpful (3.5 vs. 3.1) but also slightly more cumbersome (3.1 vs. 2.9). However, in this case differences were narrower. On the other hand, IMT-SV received 16 positive comments whereas PE received only 5. Regarding negative comments, IMT-SV accounted for 35 items and PE-SV 31 items. While the number of negative comments is similar, there was an important difference regarding positive ones. Finally, user complaints of IMT-SV can be summarized in the following items: *a)* system suggestions changed too often, offering very different solutions on each keystroke; *b)* while correcting one mistake, subsequent words that were correct were changed by a worse suggestion; *c)* system suggestions did not keep gender, number, and tense concordance; *d)* if the user goes back in the sentence and performs a correction, some parts of the sentence already corrected were not preserved on subsequent system suggestions.

Evaluation Discussion

Our initial UI performed poorly when tested with real users. However, when the UI design was adapted to user expectations, results were encouraging. Note that in both cases the same IMT engine was evaluated under the hood. This

	PE-SV	**IMT-SV**
Avg. time (s)	69 (SD=42)	55 (SD=37)
No. interactions	94 (60)	79 (55)
Q1 (Likert scale)	3.1 (1.2)	3.5 (1.1)
Q2 (Likert scale)	2.9 (1.2)	3.1 (1.3)

Table 6.3: Summary of the results for the second test.

fact remarks the importance of an adequate UI design when evaluating a highly interactive system as IMT.

In sum, the following issues should be addressed in IMT: *1)* user corrections should not be modified, since that causes frustration; *2)* system hypotheses should not change dramatically between interactions, in order to avoid confusing the user; *3)* the system should produce a new hypothesis only when it is sure that it improves the previous one.

6.3.3 Interactive Image Retrieval

In this scenario, it is desirable to retrieve as much precise images as possible in a few feedback iterations. To this end, Paredes et al. [2008] demonstrated that implicitly validating non-selected images as non-relevant is a safe and convenient assumption. Experiments on the well-known Corel/Wang dataset revealed that this method was able to retrieve 94.5% of the relevant images in just 2 iterations.

However, in an image retrieval system, there are generally available many different types of image features, and also there are textual features, such as metadata, annotations provided by users, or text surrounding the images from where they appear. Adequately leveraging all this available information is a major goal in order to obtain the best performance possible. In this section we study two approaches to achieve this goal: *1*) how to combine textual and visual information by using relevance feedback, and *2*) how to present this information to the user in a way that it may improve retrieval results.

Evaluation of Multimodal Fusion

We opted for *late fusion* as a fusion method of visual and textual features, since it is simple and easy to integrate in our previously developed prototype, a Relevant Image Search Engine (RISE) [Segarra et al., 2011]. Since only two modalities are considered, an $\alpha \in [0, 1]$ parameter is set to assign an importance weight to visual image descriptors. This allowed us to implement a linear combination of both features and let the system decide the best ranking of images according to:

$$R_\alpha(x) = \alpha R_v + (1 - \alpha) R_t \tag{6.3}$$

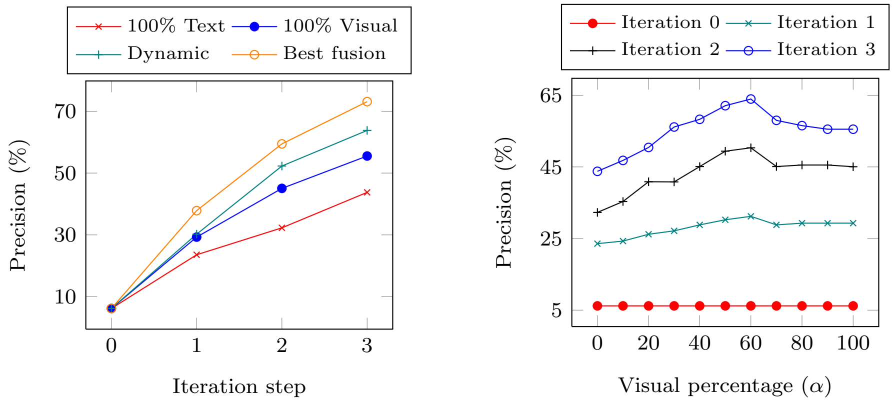
This way, when $\alpha = 0$, only textual features are considered (i.e., textual modality, R_t); while $\alpha = 1$ means that only visual features are considered (i.e., visual modality, R_v). Clearly, α should not be kept fixed for a given system, since it is known that in general some queries will perform better with visual information, or the other way around, and leaving this task to the user is too much burden. Hence, to deal with this dynamically variable weighting, we propose to take advantage of information derived from relevance feedback and solve an optimization problem: the system will try to rank relevant and non-relevant images as far as possible, also placing the relevant images in the top positions. We named this approach *dynamic linear fusion*.

To evaluate this approach, we manually labeled a subset of 21 queries with 200 images each from the RISE image database [Villegas and Paredes, 2012]. The reader may consult [Toselli et al., 2011] for a brief description of each query, together with their respective images.

Instead of recruiting users, as usual, we decided to do a preliminary evaluation first, which would eventually lead to a lab study in case results were promising. For consistency with the default RISE UI (Figure 6.6), we simulated a user who wants to retrieve $N = 10$ images, which were shown at a time. So, in each iteration, the user would see 10 images and judge which were relevant. Visual features were comprised of color histograms, while textual features were comprised of automatic image annotations (extracted from the web pages where images were located). Results are shown in Figure 6.8.

Evaluation Discussion Figure 6.8a shows the evolution of retrieval accuracy with the successive interaction steps for different retrieval strategies. As we suspected, both pure text and visual retrieval alone are worse performers. After one interaction step, the dynamic linear fusion approach performs better on average. The best fusion combination is just an upper bound, and therefore in practice it is unreachable.

It can be observed in Figure 6.8b that the system quickly gains accuracy with the progression of user interaction steps. That is, the more the information known about what is considered relevant in previous steps, the better it can predict the best fusion parameter α for the current step. In the first step, there is a clearly ascendant slope toward the visual strategy, achieving high precision when full visual search is used. However, in the following iterations the best precision is not obtained on the extremes, which shows the importance of having a dynamic user/query-adaptative α to achieve always the best precision.



(a) Precision vs. Iterations

(b) Precision vs. Fusion

Figure 6.8: Dynamic linear fusion results, for $N = 10$ images to be seen at a time. [6.8a] Comparison of image retrieval techniques. [6.8b] Precision as a function of α (visual percentage), for several feedback iteration steps.

Query Refinement Evaluation

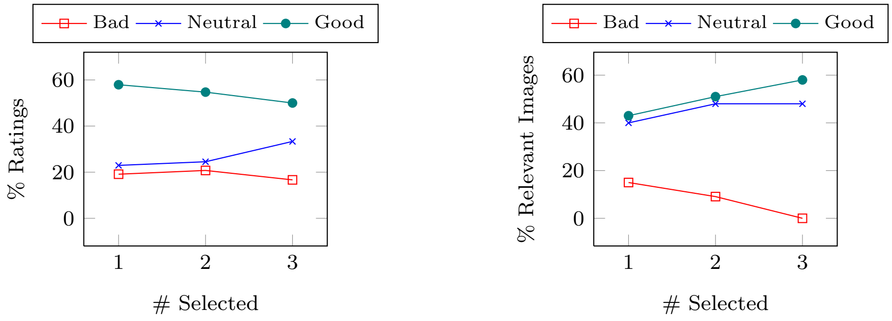
The image database used in RISE prototype was built from real data gathered from the Internet with completely unsupervised annotations, so there is no ground truth available, i.e., labeled samples. Furthermore, labeling a subset of the images in order to evaluate query refinement suggestions is rather challenging. The labeling would require to have a list of sample queries, and for each query, several subsets of selected relevant images corresponding to different subclasses of the original query. Moreover, for each of these subsets we would require a list of possibly correct query refinements. Thus, in order to evaluate the proposed approach, we opted to conduct an informal field study. The procedure was simple: to measure the user's subjectivity toward the query suggestion technique.

For the evaluation, we selected 81 out of the 99 concepts from the ImageCLEF 2011 dataset[7], and used these as the initial text search queries. The reason to remove 18 concepts was because they were related to specific image properties rather than high-level concepts, e.g., "Neutral Illumination", "No Blur", etc.

The evaluation task consisted of two stages [Leiva et al., 2011b]. First, users were presented with the first 10 ranked images for a given text query, e.g., "cat". Then the user would select a subset of images which had a common concept or relation among them, e.g., "all are black cats". If the system was able to derive a query refinement, the UI would show it and let the user rate whether the suggestion was either good, bad, or neutral. The number of times there was no query suggested (NQ) was also recorded (Table 6.4). In the second

[7] http://imageclef.org/system/files/concepts_2011.txt

stage of the evaluation, users were presented with the images after following the query suggestion, and they had to mark all of the images considered relevant to the concept they had in mind when selecting the images in the first stage of the evaluation. This two-stage process was repeated for all subsets of related images the user could identify. Three people from our department took part in the evaluation. Results are presented in Figure 6.9 and Tables 6.4 and 6.5, respectively.



(a) Stage 1: query ratings.

(b) Stage 2: retrieved images.

Figure 6.9: Query refinement evaluation results. [6.9a] Average rating of suggested queries against number of initially selected images. [6.9b] Percentage of images considered as relevant after following the query suggestions against number of initially selected images.

# selected	# samples	Bad	Neutral	Good	NQ
1	194	35	42	106	11
2	74	11	13	29	21
3	24	2	4	6	12
>3	30	0	0	3	27
Overall	322	48	59	144	71

Table 6.4: Results for stage 1 of query refinement evaluation, showing absolute ratings for suggested query refinements.

Evaluation Discussion Regarding the first stage of the evaluation, the first thing to note is that, as more images are selected, it is less probable that the system will suggest a query (see Table 6.4). This is understandable, since it is less likely that there will be common terms to all selected images. Moreover, terms associated to each image completely depend on the web pages where the image appears, thus not all images will be well annotated. Nonetheless, most of the suggested queries were rated as being good, which indicates that this approach of deriving suggestions based on selected (relevant) images can be quite useful.

# selected	# ratings	Bad	Neutral	Good
1	183	1.5 (1.5)	4 (3)	4.3 (3)
2	53	0.9 (1.4)	4.8 (3.2)	5.1 (3.3)
3	12	0 (0)	4.8 (4.8)	5.8 (2.8)
>3	3	0 (0)	0 (0)	9.6 (0.5)
Overall	226	1.3 (1.5)	4.3 (3.2)	4.7 (3.2)

Table 6.5: Results for stage 2 of query refinement evaluation, showing mean (and standard deviation) values of the number of relevant images retrieved after following suggested queries.

Regarding the second stage of the evaluation, as expected, query suggestions which were rated as being good or neutral retrieved more relevant images than bad query suggestions (see Figure 6.9b). This is convenient, since it is unlikely that a user will use a suggestion considered to be bad. A particular behavior that was also observed is that performance tends to be better for suggestions that were derived using more selected (relevant) images. Then, overall, as more images are selected, it is less likely that the system will suggest a query; however if there is a suggestion it tends to be a better one.

Another observation from the evaluation was that suggestion quality depends highly on the particular query. There are some queries where images presented to the user clearly belong to different subgroups, which, if selected, most of the time a query will be suggested that relates to that subgroup. An example of a query that provides good suggestions was shown in Figure 6.6.

Tag Cloud Evaluation

In the same way as in query refinement evaluation, obtaining labeled data to be able to assess tag cloud suggestions is rather difficult. Thus, to perform the evaluation, we conducted again an informal field study, using the same database used in RISE prototype.

Fourteen users aged 31.42 (SD=5.34) were recruited via email advertising to participate in the evaluation study. They were told to assess the relevance of the $N = 10$ top scored tags suggested in the cloud for a series of queries (12 queries per person on average).

The list of queries was compiled by merging two lists from ImageCLEF 2012: Photo Annotation and Retrieval. Concretely, we merged concepts from 'Large-scale annotation using general Web data' subtask[8] and queries used in 'Visual concept detection, annotation, and retrieval using Flickr photos' subtask[9]. The final list comprised 164 search queries in total.

[8] http://imageclef.org/2012/photo-flickr

[9] http://imageclef.org/2012/photo-web

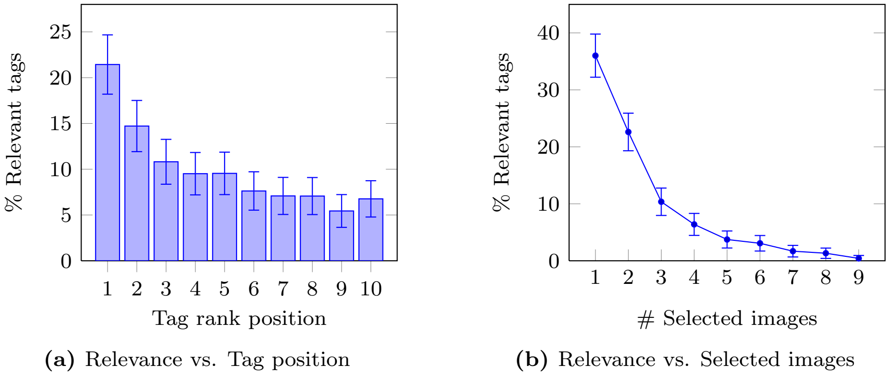


Figure 6.10: Evaluation results of the tag cloud, with 95% confidence intervals.

While evaluating each query, participants had to follow their given list of queries and select a subset of images for different subtopics from the presented set of 10 images. Participants had no restrictions on subtopic selection, e.g., a subtopic could have an arbitrary number of images, no minimum or maximum subtopics per query were imposed, etc.

Whenever a relevant image was selected from the presented set, a list of 10 tags was displayed in order of relevance (most relevant tags at the beginning of the list, in a left-to-right order). A check box was attached to each tag, so that users could mark whether the tag was considered relevant to the subtopic or not.

It is worth pointing out that no tag cloud in the strict sense was displayed, but a text-only tag list sorted by relevance, since we wanted to avoid any possible visual bias in the study. Figure 6.10 shows the evaluation results.

Evaluation Discussion Users reported that sometimes tags were found to be really useful and beneficial for the current query, but also sometimes they were found to be meaningless. This fact is explained by the noise due to the image indexing procedure, which was completely unsupervised and therefore the cloud may contain irrelevant tags for a particular query. This can be observed in Figure 6.10a, where each bar represents the average percentage of relevant tags (normalized by the number of selected relevant tags) given the rank position of each tag. Nonetheless, as expected, tags in the first positions of the cloud tended to be perceived more often as relevant. Differences between the first ranked tag and the other tags are statistically significant.

A study by Bateman et al. [2008] reported that tags with a larger number of characters tended to be selected less often. We investigated whether this could be observed in our study as well. We computed the tag length ratio as the division of the average length of selected tags by the average length of all suggested tags, and obtained 1.03 (SD=0.16), which means that selected tags were around the average tag length overall. Furthermore, only 10% of the time a user chose a tag that had more than a 1.1 of tag length ratio. This suggested that the length of a tag was not determinant to assess its relevance toward a particular query, but also that users did not choose neither shorter or longer tags overall.

Figure 6.10b depicts the proportion of relevant tags according to the number of selected images. As observed, relevance differences between tags presented when selecting #1 or #2 images with respect to the rest of selections were found to be statistically significant. Similar conclusions to those observed in the query suggestion evaluation were derived: 1) as more images are selected, the overview the tag cloud provides about such a set of images tends to be more general; and 2) the quality of the tags depends highly on the particular query.

As observed, therefore, when a single image is selected, nearly half of the tags are considered as relevant, since the tag cloud is specifically tailored to such a single selection. Then, this proportion falls dramatically as more images are selected. This suggests that when many images are selected, a new strategy for generating tag clouds should be devised. Nonetheless, on average, 21.49% (SD=10) of the presented tags were considered as relevant at any time.

All in all, our study indicates that the tag cloud approach supports its intended goal, i.e., impression formation about a particular set of relevant images. Furthermore, the tag cloud provides the user with more options to refine the initial (textual) query. As such, we believe that a tag cloud has more potential than a query refinement suggestion, at least in an interactive image retrieval scenario.

6.4 Conclusions and Future Work

The IPR framework proposes a radically different approach to correct the errors committed by a PR system. This approach is characterized by human and machine being tied up in a much closer loop than usually. This way, errors can be avoided beforehand and correction costs can be dramatically reduced. We have characterized the interaction protocol that rules the IPR framework, and have introduced a series of prototypes that successfully illustrate it.

The literature had reported good experimental results in simulated-user scenarios, where IPR is focused on optimizing some automatic metric. However, user productivity is strongly related to how users interact with the IPR system and other UI concerns. For instance, in the NLP applications introduced in this chapter, a hypothesis that changes on every keystroke might obtain better automatic results, whereas user productivity may decrease because of the cognitive effort needed to process those changes. Therefore, the current IPR framework should be revised in order to optimize further these NLP systems toward the user. In this regard, we have suggested some approaches, such as avoid modifying any user-submitted correction by any means, or deriving a new hypothesis only when the system is sure that it will improve the previous one.

Regarding IPR systems that deal with desultory user input, we have shown that implicit interaction can notably improve system performance. We have presented a series of image retrieval strategies to illustrate this fact, such as *1*) rethinking the classical retrieval protocol, in which users must indicate which images are relevant *and* non-relevant, to a much simpler one in which the system can assume that non-selected images are not relevant; *2*) combining multimodal information from selected images to provide better results; *3*) using this multimodal information to provide the user with optional suggestions, either in the form of a refined query suggestion or a tag cloud.

We have demonstrated that the techniques presented so far are both suitable and convenient. Each technique is based on a probabilistic model to handle user interaction, which allows IPR systems to take the lead in coordinating different user feedback signals. We hope these considerations will guide researchers to future developments that can have a significant impact both on academia and industry.

Bibliography of Chapter 6

V. Alabau, L. A. Leiva, D. Ortiz-Martínez, and F. Casacuberta. User evaluation of interactive machine translation systems. In *Proceedings of the European Association for Machine Translation (EAMT)*, pp. 20–23, 2012.

S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. L. Lagarda, H. Ney, J. Tomás, E. Vidal, and J. M. Vilar. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28, 2009.

S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia (HT)*, pp. 193–202, 2008.

M. Bisani and H. Ney. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 409–12, 2004.

J. Brooke. SUS: A “quick and dirty” usability scale. In P. Jordan, B. Thomas, B. Weerdmeester, and A. McClelland, editors, *Usability Evaluation in Industry.* Taylor and Francis, 1996.

A. Culotta, T. Kristjansson, A. McCallum, and P. Viola. Corrective feedback and persistent learning for information extraction. *Artificial Intelligence*, 170(14–15):1101–1122, 2006.

L. A. Leiva, V. Romero, A. H. Toselli, and E. Vidal. Evaluating an interactive-predictive paradigm on handwriting transcription: A case study and lessons learned. In *Proceedings of the 35th Annual IEEE Computer Software and Applications Conference (COMPSAC)*, pp. 610–617, 2011a.

L. A. Leiva, M. Villegas, and R. Paredes. Query refinement suggestion in multimodal interactive image retrieval. In *Proceedings of the 13th International Conference on Multimodal Interaction (ICMI)*, pp. 311–314, 2011b.

D. Ortiz-Martínez, L. A. Leiva, V. Alabau, and F. Casacuberta. Interactive machine translation using a web-based architecture. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI)*, pp. 423–425, 2010.

R. Paredes, T. Deselaers, and E. Vidal. A probabilistic model for user relevance feedback on image retrieval. In *Proceedings of the 5th international workshop on Machine Learning for Multimodal Interaction (MLMI)*, pp. 260–271, 2008.

G. Pask. *Conversation, cognition and learning: A cybernetic theory and methodology.* Elsevier Science, 1975.

V. Romero, L. A. Leiva, A. H. Toselli, and E. Vidal. Interactive multimodal transcription of text images using a web-based demo system. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI)*, pp. 477–478, 2009.

F. M. Segarra, L. A. Leiva, and R. Paredes. A relevant image search engine with late fusion: Mixing the roles of textual and visual descriptors. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI)*, pp. 455–456, 2011.

A. H. Toselli, V. Romero, M. Pastor, and E. Vidal. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1814–1825, 2009.

A. H. Toselli, E. Vidal, and F. Casacuberta, editors. *Multimodal Interactive Pattern Recognition and Applications.* Springer, 1st edition, 2011.

TT2. TransType2 - computer assisted translation. project technical annex, 2001. Information Society Technologies (IST) Programme, IST-2001-32091.

E. Vidal, L. Rodríguez, F. Casacuberta, and I. García-Varea. Interactive pattern recognition. In *Proceedings of the 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, pp. 60–71, 2007.

M. Villegas and R. Paredes. Image-text dataset generation for image annotation and retrieval. In *II Congreso Español de Recuperación de Información (CERI)*, pp. 115–120, 2012.