Making Sense of Ubiquitous Media

Max Mühlhäuser
Technische Universität Darmstadt, Telecooperation
max@informatik.tu-darmstadt.de

Abstract

In the emerging Post-PC era, more and more computers 'in the net' can see, hear, or feel. Since these computers are networked, they can cooperate in the interpretation of their 'sensation'. Cameras, camcorders, etc. will soon be wirelessly connected, doubling as mobile phones. In other words: multimedia goes ubiquitous. On the other hand, users leverage off the wealth of text-based information present in the global Internet. However, the potential that lies in the 'cooperative sensation' and in the use of global textual information is by far not leveraged: it is the past, present, and future grand challenge to enable computers to 'make more sense' of all this information. The talk will provide a unified model for both multimedia sense-making and textual-information sense-making, and propose fostering the confluence of these two threads. Based on this unified view, it will suggest steps towards improved sense-making in the world of ubiquitous computers.

1. Introduction

Networked computers become ubiquitous, be it in the form of companion devices or as co-located parts of everyday objects. As the cost of sensors and of recording and storage devices continues to decline, these computers are increasingly capable of sensing all kinds of real-world signals such as classical media (sound, vision), interaction modalities (handwriting, voice, gestures), and sensor data known as *context* (temperature, location etc.). As to classical media, cell phones start to comprise full-fledged camera functionality, making instant networked media a commodity that influences social networks. As to multimodality, the mobile use of 'networked media computers' boosts multimodal interaction as a means for coping with both restricted screen/keypad real-estate and restricted attention during usage. As to context, the use of environmental sensor data (location, acceleration etc.) has developed into an area of intensive research called context-aware computing. Altogether, it seems that the distinction between the areas multimedia, multimodal interaction and context-awareness becomes blurred; we will use *media* as the encompassing term for all three categories of related data in the remainder. As media are generated and stored, transmitted, processed, and output, the need for *media sense-making*[1] becomes ever more urging – in face of a broad range of users with usually very little technical background

On the other hand, many computers on the net serve as storage for all kinds of human-generated information, most of which is not machine-readable but based on natural language. Zillions of Web pages, 'PDF documents' and the like, e-Mail threads, Forums, Blogs, and Wikis contain a wealth of valuable information that we have much trouble finding and mapping to actual user demands. This leads us to a second issue of key importance that we call 'text sense-making'.

Just as each *media-prone computer* (sensors, audio and video recorders, etc. as mentioned above) holds only a tiny fraction of the 'sensed' world, each *text-prone computer* (mail servers, blog/wiki/news-group servers, etc.) holds only a tiny fraction of the globally available textual information. Drawing an analogy to human (biological) information processing, these computers resemble (small clusters of) neurons in our senses or in the memory of our brain, respectively: in both the biological and the computer world, the power lies in the interconnection and distributed processing of the zillions of sensory and memory-bearing nodes – and it is the power of sense-making. It is obvious that – in comparison to the brain – we are only just beginning to exploit this 'power of interconnection' in the computer world.

Of course, all of the above arguments represent an extremely simplistic view while the true attempts to learn from human brains for improving computers have been the lifetime endeavor of geniuses like Marvin Minsky [8]. Our intension here is much simpler, namely two-fold: i) we want to look broad enough for

[1] We deliberately use *meaning* and *sense* as convivial terms instead of *semantics* in order not to confine the discussion to a narrow definition

0-7695-3058-3/07 $25.00 © 2007 IEEE
DOI 10.1109/ISM.2007.43

understanding the effects of the ever more ubiquitous nature of media-prone and text-prone computers; ii) we want to 'zoom out' to a view the helps us understand the fundamental differences *and similarities* among general approaches taken by two rather distinct research communities, one gathered around the problem of *multimedia information retrieval (MMIR),* the other one around *natural language processing (NLP)*. This comparison of differences and similarities leads to the proposal of a combined approach, in which mutual benefits can be exploited; this combined approach is mapped onto the challenges of 'ubiquitous media'. Given the panoramic view of the keynote paper at hand, the proposed combined approach is much more the outline of future research to be undertaken than a ready-made solution. It remains to be referred back onto the wealth of global research activities in the field, which try to make concrete small steps towards working solutions.

The remainder of the paper is structured as follows: in chapter 2, we resume the differences between MMIR and NLP, and in chapter 3, we attempt to look at the similarities. Chapter 4 roughly sketches an integrated approach that combines the two fields. Chapter 5 looks at important sub issues, where the two fields start to meet, trying to give at least a few hints about how to map the big picture onto intermediate steps to be undertaken in particular research areas.

2. Differences

As pointed out in the introduction, media-prone computers take in information from the real world via cameras, microphones, and sensors. In contrast, the text-prone computers take in information basically via keyboards from users. These differences become blurred as user-generated media on the net come into fashion and get increasingly mixed with text and accross media (cf. Flickr and YouTube as just a start) and as alternative input methods arise. However, a fundamental difference remains: the result of the 'intake' is a digital representation of physical *signals* in the case of media-prone computers, as opposed to a *symbolic* representation (natural language: letters, words, phrases …) in the case of text-prone ones. In the latter case, concepts are articulated using the means which humans have developed to express and convey 'meaning' or 'sense'. For our considerations, it is sufficient to stick to this major difference, we can abstract from most other valid considerations in the comparison.

The statements about media-prone and text-prone computers made in the last two chapters are summarized in the table below. The reader may observe the challenge of deriving 'sense as in sense-making' from 'sense as in sensation' (the latter is meant in the table, the former will occupy us in the rest of this article).

Table 1. Comparison of node categories in the future ubiquitous information network

Class	primary function	represen-tation	brain analogy	sense-making challenge
media prone	sense	signal	sense	sensing → perception
text prone	store	symbol	memory	knowledge → deduction

3. Similarities

We will now look at how users are supported today in their attempts to access the wealth of information available in the 'global digital brain' as we dare to call the Internet in our simplified analogy. Depending on the information class (media vs. text), two disjoint research domains come into play, namely MMIR and NLP as mentioned.

As figure 1 shows, the very coarse-level architecture of typical systems is astonishingly similar in both cases despite the differences listed in table 1: the user is supposed to formulate her demand by providing either a sample (in the case of similarity search) or a query. At the core of system design, there is a model of the set of candidates (media, text) in store. Let us assume a machine learning approach: in this case, the model basically consists of assumptions about which features to consider, how to arrange them as input to a machine learning algorithm, etc. The model is tuned based on a candidate training set, and the match between queries or samples and the candidates in store is computed based on the correlation among features.

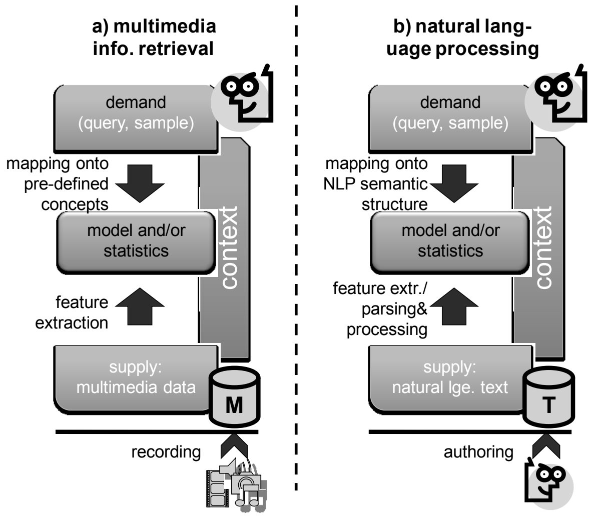
a) multimedia info. retrieval
demand (query, sample)
mapping onto pre-defined concepts
model and/or statistics
context
feature extraction
supply: multimedia data
M
recording
b) natural language processing
demand (query, sample)
mapping onto NLP semantic structure
model and/or statistics
context
feature extr./ parsing& processing
supply: natural lge. text
T
authoring

Figure 1. MMIR vs. NLP based information retrieval

In NLP, linguistic models (grammers etc.) and machine learning approaches are applied. In case a grammar-based approach is used, the text is parsed (and cleansed), and linguistically relevant concepts are identified. The transition from linguistic structure to 'meaning' is attempted with two different scopes: narrow-but-deep scope is provided by (manually generated) ontologies describing 'canonical' knowledge about particular application domains, broad-but-shallower scope is provided via 'world knowledge'. The latter is referred to via electronic lexical databases like WordNet for English [3], knowledge representations like OpenCyc [10], and – maybe most notably! – via properly processed 'folksonomies' like Wikipedia [11]. The latter approach copes for the highly dynamic evaluation of knowledge and for the power of social networking, hence of the globally interconnected 'computer brain' i.e. Internet; in combination with detailed ontologies of the application domains, a 'best of both worlds' solution can be generated – an important hint for future directions in MMIR!

Figure 1 indicates that both text- and media-prone retrieval consider context. It must be noted, though, that the kind of context considered in the two cases is rather different: NLP systems consider linguistic context, i.e. (roughly speaking) the text before and after the clauses currently processed; MMIR systems consider situational context i.e. auxiliary information about the situation in which media were recorded – as far as helpful. In particular, situational context comprises sensor data that can be used to annotate media data, such as the time, location, and camera settings used when taking a photograph. The concepts of context seem to be very different at first sight. At second sight, however, it is obvious that the two research domains might 'learn' a lot from one another in this respect: for instance, the situational context used in MMIR systems is likely to improve the quality of NLP algorithms. In addition, situational context can not only be derived from sensors, but also from machine readable data and from text (as an example, a calendar entry object may provide a time and date, and the associated text may indicate a location). Looking at the performance of known NLP and MMIR concepts and systems, the following can be stated as a summary:

- MMIR systems are currently bound to the visual concepts (picture of a human vs. animal, landscape, etc.) that were integrated i.e. modeled and/or trained before; the fact that top of the line image retrieval systems support hundreds of concepts (such as blue sky vs. cloudy sky, indoor vs. outdoor shots, etc.) should not be confounded with the ability to formulate arbitrary queries; in other words: 'common knowledge' is not supported in queries but 'approached' by increasing the number of supported concepts
- NLP systems suffer from a breadth-vs.-depth dilemma: they work well and provide sophisticated results as long as the application domain remains small (e.g., train tables, hotel bookings), but support only a limited level of sophistication if domain independence is required.

However, despite the shortcomings listed above, we dare to claim that the state of the art in NLP related systems is more advanced than that of MMIR in two respects: i) they are based on the evolutionary agreed-upon ground for expressing and conveying 'meaning': natural language[2], as opposed to signals in the case of MMIR; ii) they have managed to take at least an inroad to considering world knowledge in addition to restricted-domain knowledge (note that NLP was at a comparable status as MMIR years ago, when many researchers believed that the consideration of world knowledge would remain fiction for a long time).

The considerations above should be a sufficient motivation for the proposed a tighter integration of NLP and MMIR approaches as described below

3. Consolidated Approach

It is safe to assume that users trying to retrieve text or images have 'something in mind' regarding what they look for (called demand in our figures). The query or sample that they provide is supposed to represent this demand. It is therefore a promising attempt to make computers understand the multi-facetted 'meaning' or 'sense' associated with the query or sample. Since human language is the most sophisticated common way to convey meaning or sense, it can be considered the user friendly candidate for exchanging meaning between humans and machines.

In fact, natural language dialogues are a high priority challenge as, e.g., a look at IT centric science fiction can tell: most pertinent books and movies let humans *talk* to computer in unconstraint language. Ssuch interfaces are still restricted to application domains since present NLP approaches are still far from human performance in 'sense-making'. Lately however, NLP based techniques for 'question answering' (as opposed to formal queries) improved considerably [6].

[2] more appropriate formalisms exist for certain purposes, such as mathematical or logical ones, but they are restricted in purpose and bound to humans with corresponding education

The move from formal queries to free text questions in MMIR may be considered to raise the scientific challenge too much. The following arguments make us stick to this challenge in the present article: i) if we want MMIR to move from the signal level to the 'meaning' level, we have to find a means for expressing it anyway; ii) natural language questions tend to be more extended than formal queries, hence contain more context, and hence provide better chances for computers to capture the 'meaning' associated; iii) most importantly, natural language questions can be considered as elements of both an HCI *dialogue* and a (information handling) *process;* using a dialogue in order to elicit more of the users intended 'meaning' of what (media, text) she is looking for, and understanding the process in which the search is embedded, can considerably improve the pertinence of the retrieval.

Let us assume (simplistically formulated again) that both MMIR and NLP systems were integrated and both 'computed' the meaning, or sense, from an expressed user demand, and match it to the 'meaning' of stored candidates. To take the example of image search: not only could attributes of objects in the scene be reflected (a search for a 'door' would not yield an image of something that has a high visual resemblance with a door but cannot be opened), but rather, relationships between objects could be exploited. More far reaching, combined search across multiple media *and* across media and text could be performed.

Returning to the issue of context, two important observations regarding MMIR context must be mentioned:

i) modern digital cameras record much more than physical signals; in fact, the data captured by their 'other' sensors is partly translated into a first level of symbolic representation already in the camera, such as date and time; further context can be retrieved via auxiliary i.e. non integrated 'sensors' such as GPS receivers

ii) the further such context data can be moved from a physical level (GPS coordinates) to a semantic level ('place'), the more they can help to match textual queries.

At this time, we want to summarize the key statements related to the proposed integrated NLP/MMIR approach:

1. Just like text retrieval is successfully augmented by state of the art 'sense-making' i.e. semantic technologies, in particular NLP including ontology-based and world-knowledge based reasoning, MMIR should be augmented by putting more emphasis on semantic representations of images
2. Text based queries play an important role in MMIR today already; this fact and requirement (1) together call for a fusion of text based and media based technologies: the handling of text-prone and media-prone nodes (cf. chapter 1) should become confluent in a common collaborative distributed processing (cf. the 'brain analogy')
3. Natural language should be the common ground on which more natural formulation of demand (from queries to question answering) and sense-making in both text and media should be based
4. Beyond the above, *interactive* search dialogues and consideration of the encompassing *process* of information handling should be considered, as a key for better understanding of the 'meaning' associated with the interactive search.
5. Besides linguistic context, the important role of situational context should be leveraged for both media understanding (more than in the past) and text understanding (at all).

Item (4) above leads to the requirement that, in addition to the data discussed (text, media, meta data, and situational context), interaction and process knowledge should be *captured and stored:* this approach is promising since the interaction-and-process observed during search can be compared to the interaction-and-process during generation of the media and text in store.

All in one, the above listed arguments lead to a system design as depicted in figure 2, where the search process – integrating both MMIR and NLP – is considered a variant of the generation process: both are treated equally, and data in store include information about interaction and process.

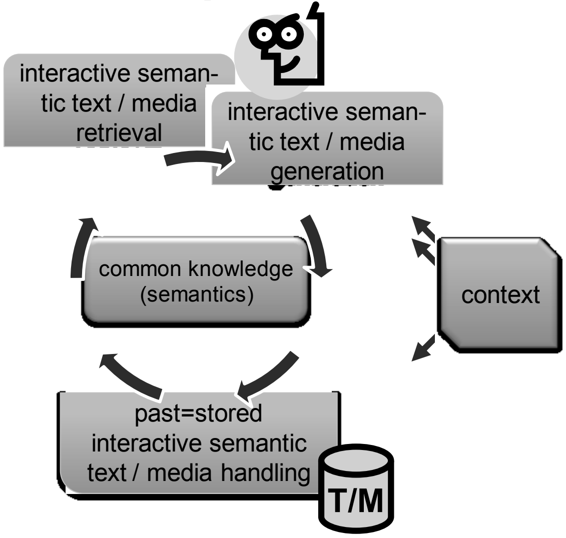
interactive semantic text / media retrieval
interactive semantic text / media generation
common knowledge (semantics)
context
past=stored interactive semantic text / media handling
T/M

Figure 2. Integrated and consolidated approach

4. Steps towards the vision

Of course it is crucial to devise reasonable steps towards the grand challenge laid out in the last chapter. The oral keynote will provide more in-depth coverage of this issue, while the present article, for the sake of space, will only scratch at the surface of a few important such steps as follows.

Regarding *situational context*, it is important to note that the use of NLP techniques is already considered important for advancements in context aware computing, fostering the confluence demanded in the present article. For instance, NLP approaches were proposed as a means for closing the gap between the 'physical' world of sensor updates and the 'information processing' world of software context models. The idea of a 'text sensor' introduced in [2] was used in [4] to discuss further opportunities and challenges associated with NLP-prone context processing.

Regarding *natural language* interfaces and multiple media, advancements in the combination of NLP and multimodal interaction is a viable intermediate step; for an excellent example, the reader may refer to [5].

The *process and interaction aspect* is probably most advanced (with respect to the holistic approach proposed in this article) in the area of *meeting recording*, summarization, and search. . In this area, the communities of NLP, machine learning, and multimedia have probably come closest. Meetings are a good example for a structured process of interactive text and media generation. It gives rise to the assumption that appropriate process models should be based on sophisticated i.e. hierarchical and richly described *events*From the wealth of pertinent publications in all three fields, [1,7,9] present just a selection for the interested reader to start with.

5. Conclusion

The conclusion of this article is just an invitation to embark on the road that was merely opened in the chapters above. The confluence of natural language, multimedia, multimodality, and context-aware computing, is a promising endeavor, and first results from the application domain 'meeting recordings' provide confidence that the challenge is not too demanding to be met in the foreseeable future. It was argued that pertinent solutions should comprise interactive search and take into account the process in which generation and retrieval are embedded. Considerable impact on the NLP and MMIR communities can be hoped for.

6. References

[1] S. Bengio, H. Bourlard (Ed.) 1st Intl. Workshop on Machine Learning for Multimodal Interaction, MLMI 2004, Springer LNCS 3361, 2004.

[2] A. Blessing, S. Klatt, D. Nicklas, S. Volz, H. Schütze, *Language-Derived Information and Context Models*. Proc. 3rd IEEE PerCom Workshop on Context Modeling and Reasoning (CoMoRea) at 4th IEEE Int. Conf. on Pervasive Computing and Communication (PerCom'06), January 2006.

[3] C. Fellbaum (Ed.), *WordNet - An Electronic Lexical Database,* MIT Press 1998

[4] I. Gurevych, M. Mühlhäuser: *Natural Language Processing for Ambient Intelligence*. In: Special Issue of KI-Zeitschrift "Ambient Intelligence und Künstliche Intelligenz". pp. 10-16, 2007.

[5] M. Johnston, L. D'Haro, M. Levine, B. Renger, *A Multimodal Interface for Access to Content in the Home,* Proc. 45th Annual Meeting of the Assoc. of Computational Linguistics, Prague, June 2007, ACL, pp. 376-383

[6] J. Lin, B. Katz, *Question answering from the web using knowledge annotation and knowledge mining techniques.* In Proc. 12th Intl. Conf. Information and Knowledge Management (New Orleans, LA, USA, November 03 - 08, 2003). CIKM '03. ACM Press, pp. 116-123.

[7] D. McColgin, M. Gregory, E. Hetzler, A. Turner, *From Question Answering to Visual Exploration*, Proceedings of the ACM SIGIR workshop on Evaluating Exploratory Search Systems, EESS 2006 Workshop., 47-50, Microsoft Research 2006.

[8] M. Minsky: *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, Simon & Schuster 2006

[9] P. Wellner , M. Flynn , M. Guillemot (2004) Browsing recorded meetings with Ferret. In: Bengio S., Bourlard H. (eds) Proceedings of machine learning for multimodal interaction: first international workshop, MLMI 2004, vol. 3361. Springer, Berlin Heidelberg New York, pp. 12–21

[10] http://www.opencyc.org (last seen on Sept. 21, 2007)

[11] T. Zesch, I. Gurevych, M. Mühlhäuser, *Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets.* In: Proc. Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007). pp. 205-208, Accociation for Computational Linguistics, 2007.