

Doctoral studies and research proposal  
ADAPTING TO SITUATIONS INFERRED FROM CHAT DISCUSSIONS  
Taneli Vähäkangas  
December 16, 2009

# 1 Proposed research

In context-aware computing we are concerned with what is a person's situation and how a computer system may adapt to changes in it.

This is a proposal for researching how a computer system may adapt to the context inferred from a person's online textual communication. There are three major parts to this: understanding which features can be extracted from informal chat discussions, what statistical models best capture the essence of these features, and which new functionalities can be built based on these that are meaningful to users.

The next chapters provide an overview to state of the art research in the relevant fields of computer science, the research question and finally methods to answer the question and expected outcomes. The proposal is concluded with a plan of studies, timeline and financing.

## 1.1 State of the art

The research proposed here belongs to the **ubiquitous computing** sub-field. More specifically, in **context-aware computing** we can identify two distinct approaches to defining a person's situation, or context. It is treated as a representational problem in [DA99], placing emphasis on information properties of context, as gathered from sensors. Criticizing this, [Dou04] treats context as primarily an interactional problem, arguing that context is inseparable from the activity within which it happens.

In **social network analysis** (SNA) of particular interest is research which not only considers the social network graph, but also attaches attributes inferred from communication to the edges of the graph. In [TT04], topic models are learned from group chat discussion with multinomial principal component analysis. The results are further improved with parameters obtained from social network analysis. In [BdR07b] online communities are discovered not only by analysing social network topology, but also by contents of blog postings. In [MWCE07] the topic models are learned from communication with latent Dirichlet allocation. In an improved model the roles of authors are identified by extending topic models with author and recipient data. In [ZM07] events are detected from communication. Interestingly, the authors acknowledge that "... text streams are sensors of the real world."

**Sentiment detection** has recently gained popularity as a research topic, for example in [KH04] word and sentence level sentiments are identified. In [WWH04] authors discuss the strength of opinion sentences.

A large body of work is associated with **mood and emotion detection or identification**, further associated with the term *affective computing* in-

troduced by [Pic97]. For a slightly more recent research, for example [ARS05] and [LLS03] offer opposing views on use of statistical methods in emotion identification. Of particular interest is research among the lines of [BdR07a], on associations between blog post topics and author's mood.

While not covered in this proposal, research may explore and draw insights from outside analysis of text. Two specific areas can be identified. Emotion in **spoken communication** has been studied quite extensively, for a synopsis, see [Sch03]. **Non-verbal communication**, or so called body language, is seen as an emerging area of study in [VPBP08].

## 1.2 Research question

State of the art research has addressed many relevant problems, but also leaves many open. Thus, the following research questions emerge

- what are the natural language features in chat discussions that define or describe the user's situation,
- which of those can be computed efficiently,
- how to adapt computer systems for the benefit of users in such identified situations.

The hypothesis is that

using suitable natural language processing and machine learning methods on communication content and parameters, it is possible to build computer systems that better adapt to users' needs.

## 1.3 Methods

The theoretical part of the work will be to analyse communication data for patterns and features that explain or capture the situation a person is in. Using data collected from actual chat discussions, the initial approach will be to employ supervised learning algorithms for classification. The classes correspond to features relevant for the proposed research. Building on the understanding gained in this initial step, further research will also explore other algorithms.

One intuition into the research is the following. Often learning a model from text requires filtering out some perceived noise. It is likely that contextual information is most prominently present in the part that was left out, and nothing of it can be learned.

The practical part of the work will be to extend existing software applications to adapt to the situations found in theoretical analysis, or to implement completely new applications with novel user interface features. The new adaptation techniques can be evaluated through increased/decreased user acceptance, surveillance/observation and interviews or diary study.

The following application examples illuminate the proposed research. If we can sufficiently accurately extract the sentiment or mood of personal communication, and further automatically learn which dependencies it has to other information, can we also use such models to predict outcomes of team work, or assist in managing group dynamics at the work place? At a more basic level, is there contextual information in features often considered noise, or impediments, e.g. typing mistakes, style, register, use of interjections and grunts? For example, if it is possible to show that typing mistakes happen more frequently for some people only when in a hurry, how could this be detected and how could the software system adapt to such situations? Can we tell hurry apart from mistakes due to relative subjective unimportance of message topic or recipient?

## **2 Expected research results and time schedule**

I expect that the theoretical work of data analysis provides excellent basis for novel models of situational data. Further expectation is that this context information can be successfully used to adapt different applications to everyday situations. Evaluating the effectiveness of such adaptation requires experiments with human subjects. This introduces a considerable risk to meeting the expectations. A low participation rate would (1) dilute the relevance of results, and (2) cause delays in meeting the deadlines.

As an end result of evaluation, it is expected that we will gain a robust model for explaining a users' situations in a novel way, based on their communication characteristics.

It is entirely possible that the proposed research benefits and advances the state of the art of not just context-aware computing, but possibly also natural language processing, machine learning and data mining sub-fields.

A particularly desirable secondary outcome, albeit rather difficult to promise or predict, would be new understanding of linguistic or social behaviour of people.

Expected time schedule is from spring 2010 to spring 2015, i.e. five years from start to successful defense of the thesis.

### 3 Concrete goals for the first two years of study

The following studies support the topic of thesis.

#### Spring 2010

- 4 credits, Probabilistic Models (582636)
- 2 credits, Project in Probabilistic Models (582637)
- 6 credits, Unsupervised machine learning (582638)

#### Autumn 2010

- 10 credits, Statistics
- 4 credits, Information-Theoretic Modeling (582650)
- 3 credits, Seminar

#### Spring 2011

- 8 credits, Natural Language Processing (582602), if available

#### Autumn 2011

- 3 credits, Seminar
- 4 credits, additional computer science courses

#### During 2010-2013

- 6 credits, PhD student seminar (582710)

The following support general postgraduate studies.

582720 philosophy of science and 582721 research ethics fulfilled by:

- 5 credits, 50048 Tieteenfilosofia ja tieteen etiikka (Philosophy and Ethics of Science / if available), spring 2011

582722 general expertise studies fulfilled by:

- 5 credits, Yliopistopedagogiikan perusteet I (Basics of Academic Pedagogics / if available), autumn 2010

582723 international scientific activities fulfilled by participation in international scientific conferences

In total, 60 credits, excluding the thesis.

## 4 Financing plan

PhD student position at HIIT, 75 % employment, from 1 Jan 2010.

## 5 References

### References

- [ARS05] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat, *Emotions from text: machine learning for text-based emotion prediction*, HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (Morristown, NJ, USA), Association for Computational Linguistics, 2005, pp. 579–586.
- [BdR07a] Krisztian Balog and Maarten de Rijke, *How to overcome tiredness: Estimating topic-mood associations*, Proceedings Int. Conf. on Weblogs and Social Media (ICWSM-2007), 2007, pp. 199–202.
- [BdR07b] Jeroen Bulters and Maarten de Rijke, *Discovering weblog communities: A content- and topology-based approach*, International Conference on Weblogs and Social Media, 2007.
- [DA99] Anind K. Dey and Gregory D. Abowd, *Towards a better understanding of context and context-awareness*, Tech. Report GVU Technical Report GIT-GVU-99-22, College of Computing, Georgia Institute of Technology, 1999.
- [Dou04] Paul Dourish, *What we talk about when we talk about context*, Personal Ubiquitous Comput. **8** (2004), no. 1, 19–30.
- [KH04] Soo-Min Kim and Eduard Hovy, *Determining the sentiment of opinions*, COLING '04: Proceedings of the 20th international conference on Computational Linguistics (Morristown, NJ, USA), Association for Computational Linguistics, 2004, p. 1367.
- [LLS03] Hugo Liu, Henry Lieberman, and Ted Selker, *A model of textual affect sensing using real-world knowledge*, IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces (New York, NY, USA), ACM, 2003, pp. 125–132.

- [MWCE07] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel, *Topic and role discovery in social networks with experiments on enron and academic email*, Journal of Artificial Intelligence Research **30** (2007), 249–272.
- [Pic97] Rosalind W. Picard, *Affective computing*, MIT Press, Cambridge, MA, USA, 1997.
- [Sch03] Klaus R. Scherer, *Vocal communication of emotion: A review of research paradigms*, Speech Communication **40** (2003), 227–256.
- [TT04] Ville H. Tuulos and Henry Tirri, *Combining topic models and social networks for chat data mining*, Web Intelligence, 2004, pp. 206–213.
- [VPBP08] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland, *Social signal processing: state-of-the-art and future perspectives of an emerging domain*, MM '08: Proceeding of the 16th ACM international conference on Multimedia (New York, NY, USA), ACM, 2008, pp. 1061–1070.
- [WWH04] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa, *Just how mad are you? finding strong and weak opinion clauses*, AAAI, 2004, pp. 761–769.
- [ZM07] Qiankun Zhao and Prasenjit Mitra, *Event detection and visualization for social text streams*, International Conference on Weblogs and Social Media, 2007.