

# Exploring Key Phrases for Browsing an Online News Feed in a Mobile Context

Anette Hulth<sup>1</sup>, Fredrik Olsson<sup>2</sup>, and Mark Tierney<sup>3</sup>

<sup>1</sup> Machine Learning Group,  
Dept. of Computer and Systems Sciences,  
Stockholm University,  
Electrum 230, SE-164 40 Kista, Sweden  
[hulth@dsv.su.se](mailto:hulth@dsv.su.se)

<sup>2</sup> Information Access and Refinement Theme,  
Swedish Institute of Computer Science,  
Box 1263, SE-164 29 Kista, Sweden  
[fredrik.olsson@sics.se](mailto:fredrik.olsson@sics.se)

<sup>3</sup> Chief Technology Officer,  
room33 AB,  
Gustavslundsvägen 139, 8th floor,  
SE-167 51 Stockholm, Sweden  
[mark@hq.room33.com](mailto:mark@hq.room33.com)

**Abstract.** This paper describes ongoing work on how to automatically identify and use key phrases extracted from items of a news feed available on the Internet. These phrases are used for two different tasks: users of mobile devices (e.g., cellular phones and personal digital assistants) will be able to subscribe to news in different categories, where the categorisation of the news is based on the extracted phrases; and by browsing through small portions of the news items — the phrases — a user can decide whether an item is interesting without having to download the whole text.

## 1 Introduction

Mobile Internet is claimed by leading suppliers and service providers to become widely spread and used within the next few years. Still, accessing the Internet via wireless mobile devices such as cellular phones and personal digital assistants presents a range of obstacles. Firstly, the connection between a mobile device and an Internet access point is usually very slow. Secondly, a mobile device typically has a small display that is not suitable for rendering graphical items. Thirdly, a mobile device is usually equipped with a processor and memory far below the standards of any stationary PC.

We believe that by focusing on the contents of the information to be presented, rather than on the accessing technology itself, browsing and retrieving relevant pieces of information is made possible without forcing the user to withstand long download times and slow processing.

## 2 The challenge

The outset is a set of daily news telegrams, originally provided by RDSL<sup>1</sup>, dealing with aspects of information technology and entertainment, and a large group of users wishing not to be exposed to every single item. The news feed is made available by the Swedish mobile software and portal technology company *room33*<sup>2</sup>, and the news service is one of many that are accessible both on the WWW and via mobile devices that have access to the Internet by WAP, microHTML, web clipping and other wireless standards.

Our goal is to design a system that presents an informative enough portion of a news item, so that the user can decide whether or not it is worthwhile to download the whole item to the mobile device.

### 2.1 Characteristics of the domain

Based on the news items we have examined (in total 141, from four different dates: October 27th 2000, January 16th 2001, February 5th 2001 and February 6th 2001), the characteristics of the present domain can be sketched along at least two dimensions: the contents of the news items; and the extra-content features such as length and news through-put rate.

Along the content dimension, the news items include a highly specialised vocabulary for expressing important information about the topic at hand. This is because the items originate from multiple sources and are edited by RSDL, where the editors create what they call *info bytes*. For instance abbreviations of monetary expression are frequent, e.g., *USDlr1.5 bil* and *Y11.3 bil* (1.5 billion US dollars and 11.3 billion Yens, respectively), as are abbreviations denoting time periods, e.g., *3rd-qtr* (third quarter of the current year). Further, names of persons, products, companies, places, and organisations are important, e.g., *David Finkelstein*, *Nextel Communications*, *Reston, VA*.

As for the extra-content characteristics, the following sums up the most important features: each news item is fairly short and the frequency of each potential key phrase or content descriptor<sup>3</sup> is low; since the heading of each news item is assigned by a professional editor, it conveys much of what the item is about; the news are intended for consumption rather than anything else; and the domain is ever-changing, which means that new descriptors may emerge at any time, while others may quickly lose their significance.

---

<sup>1</sup> Available on the Internet at <http://www.rdsl.co.uk>

<sup>2</sup> <http://room33.com>

<sup>3</sup> A *content descriptor* (descriptor for short) is an entity derived from the actual content that is to be described, e.g., the index in the back of a book contains descriptors describing, in some sense, what the book is about. We will use key phrases as content descriptors, and thus use the terms interchangeably.

### 3 Towards a solution

Our attempt at meeting the goal outlined in Section 2 is divided into two parts; firstly, a news item could be assigned to one or more of a small number of adequate categories that, taken together, cover the contents of the news feed. As far as the user goes, the categories will serve as a first step in deciding what news could be interesting. Secondly, each news item will be described by its heading, together with a number of descriptors from the body of the text. When browsing through this information, the user can make the final decision as to whether the news item is worth downloading to the mobile device. Further, all descriptors of a category can be used to characterise that category in more detail.

In the remainder of this section, we will first describe the descriptors, as they form the basis of the categorisation, then continue with the actual categorisation.

#### 3.1 What descriptors to use

Humans in general seem to have little or no problem in deciding what is significant in a document or, for example, what makes it belong to a certain category. Automating these processes is, however, all but trivial, as humans make use of so much more knowledge than the terms in the actual text. At this stage of our work, we rely only on the terms present in the collection at hand, and have to find ways to make the most out of them.

According to Carroll and Roeloffs, “The concepts an author is trying to convey is reflected in the number of times non-common adjectives and nouns are used” [2]. Justeson and Katz have shown that technical terminology consists mostly of multi-word terms being noun phrases with nouns, adjectives or the preposition *of* (more than 99 percent), and single-term words are rarely terminological terms [5]. Lahtinen has performed experiments with an automatic back-of-book indexer using linguistic information, and has come to similar conclusions: the five most common term combinations were all noun phrases, containing nouns, adjectives, or the preposition *of*. 80 percent of the unknown nouns were found to be relevant index terms in the training corpus used [6].

Based on these findings, as well as the characteristics of the domain, we will — at least as an outset — consider noun phrases consisting of more than one word. The noun phrases may contain, apart from nouns, determiners, adjectives, and the preposition *of*.

The above mentioned experiments were all performed on English texts. However, much of it could probably be generalised to be valid for other languages as well, although for Swedish (two of the authors’ native tongue), excluding single-word terms would be unwise — as Swedish is rich in compounding.

In Figure 1, a news item from February 6, 2001 is displayed together with the extracted content descriptors. The structure of the descriptors is simple, i.e., zero or more determiners followed by zero or more adjectives followed by one or more nouns or abbreviations (corresponding to  $DET^* A^* (N+|ABBR+)$  in regular notation). The phrases matched by the rule are then filtered according to their length.

---

#### US - FLEXTRONICS LAUNCHES ALMOST USDLR1 BIL SHARE OFFER

Flextronics International, contract electronics manufacturer, has launched an almost USDLr1 bil share offer, with the proceeds expected to be used to finance the company's deal to take over mobile phone production for Ericsson (Sweden), industrial group. The cost of transferring the Ericsson business to Flextronics is estimated at between USDLr300 mil and USDLr800 mil. Flextronics will issue 25 mil ordinary shares, and will also use some of the proceeds from the deal to finance the further expansion of its business. Flextronics plans to relocate Ericsson's mobile phone production to Malaysia and China from Sweden and other high-cost locations used by Ericsson.

---

contract electronics manufacturer — share offer — the company's deal — mobile phone production — industrial group — the Ericsson business — USDLr300 mil — ordinary shares — the further expansion — mobile phone production — other high-cost locations

---

**Fig. 1.** An example of a news item together with its extracted content descriptors.

### 3.2 News categories

When a user states an interest in a particular category, it can be thought of as though the user expresses some kind of information need — without having to know exactly what he/she is looking for. In doing so, the number of news items eventually presented to the user is likely to be smaller than that of the whole collection.

What categories to form for the current domain (or any domain, for the matter) is not evident. The division should somehow answer the question “What is this news about?” It is also of importance that the categories are intuitive: that a user easily can see how a category is characterised. Based on the characteristics of the domain, we suggest the following nine categories (with examples in brackets):

- Acquisitions (“company A buys company B”)
- Computers and Internet (broadband)
- Entertainment (cinema, computer games, and music)
- Financial Reports (“company A has made a net profit of. . .”)
- Fixed Telephony (companies, and regulations)
- Investments (“company A plans to invest. . .”)
- Media (television, radio, and newspapers)
- Mobile Phone Services (portals, WAP, and subscribers)
- Mobile Phone Technology (networks, companies, and phones)

The task of assigning one or more categories to the items has so far been done manually by keeping simple production rules in mind, such as:

- IF *cellular products* THEN Category = Mobile Phone Technology
- IF *mobile subscriber* THEN Category = Mobile Phone Services

The main drawback with hand-written rules (apart from the time-consuming and tedious work) is that they generalise poorly. In order to overcome this, we plan to add a machine learning component to the system to handle the categorisation. We will let the phrases constitute attributes, and each category a class. What other attributes to use will be decided on an experimental basis. These could, for example, be term frequency, inverse document frequency, or presence in a document. Manually categorised news items will be used for training the classifier. Plenty of work has been done on automatic text categorisation. For example, experiments with, and evaluations of, five different methods have been presented by Dumais, Platt, Heckerman, and Sahami [3].

## 4 Implementation

We have implemented a prototype of a content descriptor extractor taking the kind of noun phrases outlined in Section 3.1 into account. The prototype was implemented utilising an open architecture for information refinement, KABA, which is currently under development at SICS. Essentially, KABA is a set of Java-packages centred around an implementation of a light version of the TIPSTER document model defined by Grishman *et al.* [4], and a fragment of the Common Pattern Specification Language (as described by, e.g., Appelt [1]) which allows for specifying regular rules over TIPSTER annotations. As a basic linguistic analyser, we used the Functional Dependency Grammar for English (EN-FDG) from Conexor Oy [7].

## 5 Remaining challenges

There is a number of challenges that remains to be tackled. The first one is the problem of how the use of terms changes over time. How can we decide whether to discard an existing descriptor as being significant for a particular news category, or whether to adopt a new descriptor as being significant as a feature for classifying news items?

Another challenge is that of making sure that the user understands what characterises a particular news category, in order for him/her to rest assured that a chosen category corresponds to the kind of news subjects the user expects to see.

As for future work, we intend to make use of information extraction techniques for, e.g., identifying monetary expressions, as well as identifying and categorising names of, e.g., persons, places, companies, organisations, and products. Also, in conjunction with named entity recognition, it would be desirable to assign attributes to the names, e.g., John Doe, *Vice president of Tacit Inc.*

An important aspect of the continuation of the project will be the evaluation. For the described approach, we can see two possibilities: we may evaluate in a qualitative manner, by interrogating a small group of experienced users that are familiar with the domain, and compare their judgements for getting portions of selected items to getting the whole news directly. The second possibility would

be to evaluate without human assessors, thus enabling a quantitative evaluation. However, to our knowledge, no such method exists for keyword indexing, and we would therefore be obliged to develop the required method ourselves.

## 6 Conclusions

In this paper, we have presented ongoing work on the extraction of content descriptors used for categorisation and for browsing the content. Apart from these two tasks, the content descriptors may also constitute the basis for user defined profiles. In addition, they can serve a purpose similar to that of a back-of-book index. I.e., all content descriptors are displayed in alphabetic order (for the whole domain, or by category), through which users may access relevant items.

The intended user is one who makes use of a mobile device for accessing information on the Internet. This implies that as much information as possible must be conveyed with as few words as possible for the user to be able to decide whether it is worthwhile to download a particular news item. For English, multi-word noun phrases seem to form an excellent base for this task. We believe that pure statistical methods are not enough — explicit knowledge about the language is also needed. In our work, we are inspired by ideas from different areas in natural language processing, such as information retrieval, information extraction, and summarisation, and we believe that mixing approaches is beneficial for the tasks such as those described in this paper.

## Acknowledgements

We would like to thank Jussi Karlgren, and Kristofer Franzén for valuable input.

## References

1. D. E. Appelt. *The Complete TextPro Reference Manual*, June 1999. Available on the Internet at <http://www.ai.sri.com/~appelt/TextPro>.
2. J. M. Carroll and R. Roeloffs. Computer selection of keywords using word-frequency analysis. *American Documentation*, July 1969.
3. S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *CIKM-98: Proceedings of the Seventh International Conference on Information and Knowledge Management*, 1998.
4. R. Grishman, Ted Dunning, J. Callan, B. Caid, J. Cowie, L. Guthrie, J. Hobbs, P. Jacobs, M. Mettler, B. Ogden, B. Schwartz, I. Sider, and R. Weischedel. *TIP-STER Text Phase II Architecture Design. Version 2.3*. New York, New York, Jan. 1997.
5. J. S. Justeson and S. M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27, 1995.

6. T. Lahtinen. *Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods*. PhD thesis, Department of general linguistics, University of Helsinki, 2000.
7. P. Tapanainen and T. Järvinen. A non-projective dependency parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing*, Washington, D.C. U.S.A., April 1997. ACL.