

# Automated Email Integration with Personal Information Management Applications

Angelo Dalli  
NLP Research Group  
Department of Computer Science  
University of Sheffield  
*angelo@dcs.shef.ac.uk*

## Abstract

An email analysis system that extracts calendar information automatically from text is presented. Appointment and meeting information is extracted using a summariser and named entity recogniser and presented to a PIM system as a structured record. Examples and evaluation results are presented.

Email is one of the most ubiquitous applications used on a daily basis by millions of people worldwide. This work is focused on creating better ways of integrating Personal Information Management (PIM) applications = such as those found in mainstream email and scheduling applications, PDAs and mobiles with email and possibly text/SMS messages. The Email-PIM integration application has been done as part of the EU FASiL project, which aims to construct a conversationally intelligent Virtual Personal Assistant (VPA) designed to manage the users personal and business information through a voice-based interface accessible over mobile phones. In an increasingly information dominated society, the need for easy and pervasive access to information is paramount. To achieve effectiveness, two fundamental issues must be addressed: information access and information overload, especially in a voice-based application (Whittaker and Sidner, 1996)

A common problem with PIM applications are that they require significant effort to schedule appointments and meetings when collaborating with a sizeable group of people. Widely supported standards for personal data interchange such as vCalendar and iCalendar have enabled calendar information to be interchanged across most PIM applications (IMC, 1997; Dawson and Stenerson, 1998; Silverberg et al., 1998; Mansour et al., 1998). In the case of voice-based applications, it would be very convenient to have structured calendar information rather than a potentially ambiguous and long piece of text as this can enable efficient scheduling dialogues that optimise user interaction.

Although reasonably effective screen-based solutions requiring all intended attendees to use the same software exist, many web-based email applications unfortunately do not have tightly integrated automated calendar facilities. It is still difficult to send structured meeting information to different recipients using different email applications. When various scheduling options need to be presented to groups of people, this often results in some kind of discussion about the best options to choose from. Consequently, it is almost impossible to use completely automated calendar systems without involving some kind of textual communication. This work aims to close this gap by automatically identifying possible meetings, events and to-do items and presenting the

extracted information as a structured vCalendar/iCalendar record. The structured calendar information can also be embedded into an XML reply that is forwarded by the email analyser to other components, services and applications.

Four main components are used by the email analyser to extract calendar information automatically from emails: a named entity recogniser, threaded email filter, email signature filter and an email summariser. The named entity recogniser is a gazetteer based recogniser that has been trained on over half a million person names extracted from the Internet. Additional heuristics are also used to identify named entities. Named entities are further classified into the following categories: male and female proper nouns, organisation names, place names (such as towns, cities, etc.), location names (such as common building areas hall, room, floor, etc. and other common locations street corner, junction, bus stop, etc.) together with identified anaphora. Proper nouns are automatically assigned probabilities to whether they refer to males or females. Simple anaphora resolution is performed using this information together with additional heuristics (Mitkov, 2002; Mitkov, 1999). Anaphora resolution permits correct associations between named entities and meetings to be made even when there is an indirect reference between named entities and a meeting or event.

The named entity recogniser is also used to extract date and time references from emails using a large multilingual list of dates and times used by the IBM International Components for Unicode (ICU) project.

Many emails are composed by replying to an original email, often including part or whole of the original email together with new content, thus creating a thread or chain of emails. The first email in the thread will potentially be repeated many times over, which would mislead a simple analysis process into over-generating calendar entries. A thread-detection filtering tool is used to identify unoriginal content in the email by comparing the contents of the current email with the content of previous emails. Simple stylistic features like the presence of a greater than symbol in the beginning of a line are also used to determine whether a particular piece of email should be filtered out or not. A cascading logical hierarchy of meeting and event references can be extracted from the result thread structure, enabling partial information to be resolved. As borne out through experiments performed for this work, the majority of emails that contain partial, incomplete utterances related to meetings or events, almost always included the body of the original email in the reply, thus providing enough context to resolve the information successfully.

Another problem that cropped up during system testing was the information that is often included in email signatures at the end of every email reply. More than half of the email signatures detected in this work contained some form of named entity information (apart from the sender's name and/or surname), and of these most additionally contained some kind of address information. While this information can be used to resolve incomplete information in particular cases, in general it is better to ignore these boiler plate texts appended to emails irrelevantly of context.

The last component used by the analyser is word and sentence relevance information that is obtained from an email summarisation system forming part of the FASiL adaptive email handling system. The different rank-importance values are used to extract the most salient features that fall near the date and time references together with any location references. A simple heuristic was

used to determine whether an email is about a meeting or an event by looking at whether place names and locations are available together with the presence of proper nouns for meetings (and absence of proper nouns for events). To-do items linked to some particular date and time were distinguished by the absence of place name and location references.

After the type of calendar information is determined, an appropriate structured record is created from the resolved information and sent to the receiving PIM application. For this works purpose, the record was also sent to a dialogue manager that handles the speech output of the calendar information and also asks for confirmation of the automatically extracted information.=

An evaluation study was performed using a corpus of email that had its pertinent calendar information pre-noted by a human judge. The system's suggested calendar information was then compared to this gold standard to compute precision and recall values for every email. Experimental results and examples included in the full paper.

## **References**

Dawson, F. Stenerson, D. 1998. Internet Calendaring and Scheduling Core Object Specification (iCalendar), RFC2445, Internet Society.

Silverberg, S. Mansour, S. Dawson, F. Hopson, R. 1998. iCalendar Transport-Independent Interoperability Protocol (iTIP), RFC2446, Internet Society.

Mansour, S. Dawson, F. Silverberg, S. 1998. iCalendar Message-Based Interoperability Protocol (iMIP), RFC2447, Internet Society.

Mitkov, Ruslan. 1999. Multilingual anaphora resolution, *Machine Translation*, 14 (3-4), 281-299.  
Mitkov, Ruslan. 2002. *Anaphora Resolution*. London, Longman.

Whittaker, Steve. Sidner, Candace. 1996. *Email Overload: Exploring Personal Information Management of Email*, Lotus Development Corporation. Proc. ACM SIG-CHI 1996, Vancouver, Canada.