Temporal Expressions Extraction in SMS messages

Stéphanie Weiser **Louise-Amélie Cougnon** **Patrick Watrin**
CENTAL - Institut Langage et Communication - UCLouvain
1348 Louvain-la-Neuve, Belgium
{stephanie.weiser, louise-amelie.cougnon, patrick.watrin}@uclouvain.be

Abstract

This paper presents a tool for extracting and normalizing temporal expressions in SMS messages in order to automatically fill in an electronic calendar. The extraction process is based on a library of finite-state transducers that identify temporal structures and annotate the components needed for the time normalization task. An initial evaluation puts recall at 0.659 and precision at 0.795.

1 Introduction

In this study, the extraction of temporal information in SMS messages is regarded as a precondition to the development of an application for the construction of a calendar. This application includes the automatic analysis of meetings and the pre-filling of calendar events. We consider the language of temporal expression in SMS messages as a sublanguage which forms a finite subset of the whole language at the syntactic and lexical levels (Harris, 1968).

Most of the recent studies (Beaufort et al., 2010; Kobus et al., 2008; Aw et al., 2006) do not process SMS messages directly. They use a heavy preprocessing step in order to standardize the SMS script. We do not deny the relevance of the transliteration process for complex applications such as SMS messages vocalisation. However, within the framework of our project, we show that, for the extraction of temporal expressions, a normalization phase is not needed, as we tend to simply identify the boundaries of precise and particular surface structures.

Before exploring the extraction task (Section 3), we briefly introduce the corpus used (Section 2). The results of the evaluation we performed are outlined in Section 4, while Section 5 shows the prospects that emerge from this preliminary work.

2 SMS Corpus

2.1 Corpus-based study

The data used for this study is a corpus of 30,000 SMS messages (Fairon et al., 2006a) that were gathered following the strict *sms4science* collection methodology (Fairon et al., 2006b). *sms4science* is an international project that promotes the study of a substantial corpus of spontaneous text messages: users are asked to send a copy of text messages that they have already sent to a real addressee in a genuine communication situation. The 30,000 SMS messages corpus that constitutes the raw material for this study was collected in 2004 in the French-speaking part of Belgium; it was semi automatically anonymized and manually normalized[1] at the Université catholique de Louvain.

2.2 SMS Script Characteristics

We prefer not to talk about *SMS language* but about *SMS script* as it is not a new type of language but a new written practice through a new communication medium (Cougnon and Ledegen, 2010). This new practice shows various specificities, notably it seems to inhibit fear-related behaviour in writing — it erases traditional social, professional and academic demands. The addressee's physical absence, in addition to the delayed character of the media, encourages SMS users to play with language and to move away from standard language[2]. At a syntactical level, one would identify some similarities with French oral syntax such as the recurrent lack of *ne* negation marker and the absence of pronouns at the beginning of sentences. We follow a more nuanced path: it appears that these characteristics

[1] "SMS normalization consists in rewriting an SMS text using a more conventional spelling, in order to make it more readable for a human or for a machine" (Yvon, 2008).

[2] Standard language can be understood as a graphic and syntactic demand and/or as a register standard.

Proceedings of the Workshop on Information Extraction and Knowledge Acquisition, pages 41–44,
Hissar, Bulgaria, 16 September 2011.

are not related to the communication medium (oral/written) but to the communication situation (formal/casual) and are more related to register, in a Koch and Österreicher (1985) manner. Inspired by the theory of these authors, we consider there is a continuum between intimacy (*Nähesprache*) and distance (*Distanzsprache*) in the SMS communication model (Cougnon and Ledegen, 2010).

In addition to these variations to the norm, SMS script is also strongly influenced by social and regional features which increase the linguistic disparity (as in *On va au ciné* à soir/au soir/ce soir (in Canada/Belgium/France) - "We're going to the movies tonight"). Even though these variations are versatile, they form a finite set which can be formalised within a grammar.

3 Extraction of Temporal Expressions

In natural language processing, and particularly in the field of information retrieval and extraction, temporal expressions have been widely studied. The annotation (Pustejovsky et al., 2010; Bittar, 2010) as well as the extraction process (Weiser, 2010; Kevers, 2011) have often been addressed. Indeed, both are needed if we want to compute a date representation from the textual information (Pan and Hobbs, 2006).

These studies offer a wide range of solutions to automatically process temporal information, although they limit their experiments to standard language and don't take into account language variation. Nevertheless, some texts that could benefit from temporal analysis do not follow the norms of a standard language, notably an important part of CMC (Computer Mediated Communication) like e-mails, chats, social networks... In this study, we intend to determine if the methods used for standard language can be applied to more informal languages with specific characteristics, such as SMS script.

3.1 Typology of Temporal Expression

For an information extraction system, the typology of the data to be extracted is very important. We based this study on the typology developed by Kevers (2011) on standard language in which we selected the categories that are useful for our SMS temporal extraction purpose.

3.1.1 Existing Typology

Kevers (2011) classifies temporal information following four criteria that combine to give 16 categories: punctual or durative, absolute or relative, precise or fuzzy, unique or repetitive. For example, *Le 22 octobre 2010* is an expression which is punctual, absolute, precise and unique, whereas *Le 22 octobre 2010 vers 20h* (that has a different granularity) is punctual, absolute, fuzzy and unique.

This typology is very rich as it includes all types of temporal expressions found in standard (French) written language like dates, durations (*du 20 juin au 30 juillet* - "from June the 20th to July the 30th "), relative data (*le jour d'avant* "the previous day"), etc. However, not everything is useful for SMS temporal extraction.

3.1.2 Temporal Expressions to be Extracted for our Application

Our aim is to build an application to identify temporal information related to meetings or events in SMS messages. We do not need to extract past information (like *hier* - "yesterday" or *la semaine dernière* - "last week" or other information like *dors bien* ***cette nuit*** - "sleep well **tonight**"). More than that, as these expressions will serve as triggers for event extraction, the recognition of irrelevant sequences could lead to the identification of "false" candidates.

The information fundamental to this research and this application concerns meetings or events that can take place in an agenda. This is the only criterion that we used to determine the temporal expressions to extract (we call it the "calendar criterion"). The temporal expressions to be extracted can be:

- **a time:** *à 18h* - "at 6:00"; *de 14h à 18h* - "from 2 to 6"
- **a date:** *le 22 octobre* - "on October the 22nd"
- **a relative moment (day, part of day):** *aujourd'hui* - "today"; *maintenant* - "now"; *mardi* - "Tuesday"; *mardi prochain* - "next Tuesday"; *ce soir* - "tonight"; *dans 5 minutes* - "in 5 minutes"
- **an implicit expression:** *à mardi* - "see you on Tuesday"; *à demain* - "see you tomorrow".

According to the Kevers (2011) classification, the categories that are concerned by SMS messages events planning are PRPU (punctual, relative, precise, unique), DRPU (durative, relative, precise, unique) and PRFU (punctual, relative, fuzzy, unique). 13 categories from the original typology are not taken into account. We created a new category to deal with expressions such as *à demain* - "see you tomorrow" which imply that "something" will happen the next day. These expressions, which are typical of the dialogues found

in SMS messages, were not dealt with by Kevers as the corpus he studied did not contain dialogues.

3.2 Sublanguage of Temporal Expressions in SMS

The study of Temporal Expressions in SMS messages has lead us to the observation that grammars which have been created for standard language can be applied to a specific sublanguage, at least for the temporal expressions in SMS messages.

3.2.1 Comparison with SMS Script

In order to compare temporal expressions in standard language with those in SMS script, we applied the temporal grammars developed by Kevers (2011) for standard French to the normalized version of an extract of the SMS corpus (1,000 SMS messages) and compared the original SMS form and the normalized form. We found that the syntax remains the same and that only the lexicon changes. A lot of variations are introduced in SMS script, but, concerning the sublanguage of temporal expressions, they only affect the form of the words and not the word order, the syntax or the semantics.

3.2.2 Adaptation of Existing Grammars - Lexical Characteristics

As we have just mentioned, the adaptation of existing grammars to extract temporal information in SMS concerns the lexical level. As SMS messages are well known for their lexical productivity, most of the common words are subject to variation. For example *demain* (tomorrow) is usually invariable but can take many forms in SMS: *2m1, dem1, dm1, dmain* ... In order to solve this problem we built a specialized lexicon in which each variation (*2m1*) is linked to a standard lemma (*demain*), a POS tag (*ADV* for adverb) and, in some cases, some semantic features (*Time*): {2m1,demain.ADV+Time}.

One may expect the lexicon to require constant updating, as it is intended to capture phenomena that rely on human linguistic creativity, which is potentially boundless. However, this theoretical assumption is refuted by our experiments which show that even if these forms vary consequently, they form a finite lexical set, respecting the closure property of sublanguages (Harris, 1968).

3.2.3 Resources Creation

Using the extracted expressions in normalized SMS messages, we have listed all the forms for all the words that appear in a temporal expression. This has lead to a preliminary dictionary composed of 177 forms, for 55 lemmas. This dictionary still needs to be extended but covers the main temporal expressions variants.

The grammar developed for standard French has been adapted: the invariable words have been lemmatized in order to match the variations listed in our dictionary, the sub-graphs that need to be applied have been selected and new sub-graphs have been created to cover the temporal expressions that are specific to SMS and do not appear in the original grammar (*à demain* - "see you tomorrow").

4 Evaluation

We performed an evaluation for the task of temporal expression extraction. We built an evaluation corpus and manually annotated the temporal expressions. Results in terms of precison and recall are provided in Section 4.2.

4.1 Evaluation Corpus

The evaluation corpus is composed of 442 SMS messages containing temporal expressions, following the "calendar criterion". Some SMS messages contain more than one temporal expression so the total number of temporal expressions is 666.

4.2 Results

For the task of temporal expression extraction, we obtained a recall of 0.659 and a precision of 0.795. Examples of well recognized expressions are, following the classification presented in Section 3.1.2 : *N'oubliez pas: ciné Pi* ***{ce soir,.ADV+Time+PRPU} {à 20H,.ADV+Time+PRPU}*** *aux locaux!* - "Don't forget: movie Pi tonight at 8:00 at the office!" (PRPU), *cela arrangeait pierre de venir voir asseliane {demain,.ADV+Time+PRPU}* ***{entre 11h et midi,.ADV+Time+DRPU}*** - "it would suit pierre to come and see asseliane tomorrow between 11:00 and noon" (DRPU), *on sera à la maison* ***{vers cinq h trente,.ADV+Time+PRFU}*** - "we'll be home around 5:30" (PRFU), *à demain* - "see you tomorrow"(new category). The reasons behind missing expressions or incomplete annotations are of three types. (i) The format of the expressions was not predicted and is not taken into account by the grammar, e.g. *à 8.30 - 9.00*; (ii) the variant of a word is missing from the dictionary,

e.g. *dimanci* for *dimanche* - "Sunday"; (iii) there is a "mistake" in the SMS, e.g. ***un*** *peu près 15 minutes* instead of ***à*** *peu près 15 minutes* - "about 15 minutes". The results can easily be improved by working on the first two sources of errors (by extending grammars and dictionaries), while the third source of errors is more problematic, because they are really unpredictable.

5 Conclusion and Future Work

This preliminary study shows that the linguistic specificities of the SMS sublanguage of temporal expressions can be structured in order to eliminate the need for a transliteration process which can lead to errors that are difficult to deal with during the extraction process itself. This study points to numerous opportunities for future work as informal texts, such as informal texts such as SMS but also Tweets, chats, e-mails and Facebook status updates, become increasingly present and contain a lot of information that could be automatically processed.

We intend to apply this research to a calendar application that would find in an SMS all the data about events and time in order to open the calendar on the right date and help the user to fill it in. This approach suggests two complementary steps that we are currently working on:

- **Extracting the event itself**: it implies finding the subject (activity, event), the actants (in SMS, it is mostly the sender and the addressee), the time and place. On a linguistic level, we will try to find out if the properties of the sublanguage (a finite list of graphic and syntactic variations that can be formalized) can also be applied to the different items of events (place, subject, actants).
- **Importing the event in a calendar**: the important task in filling a calendar is to open it on the right date (and time). In order to do this, temporal expressions extracted from the SMS needs to be standardized and formalized in "calendar information" format.

References

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING/ACL 2006*, COLING-ACL '06, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, and Cédrick Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of ACL 2010*, pages 770–779.

André Bittar. 2010. *Building a TimeBank for French: A Reference Corpus Annotated According to the ISO-TimeML Standard*. Ph.D. thesis, Université Paris Diderot (Paris 7).

Louise-Amélie Cougnon and Gudrun Ledegen. 2010. C'est écrire comme je parle. Une étude comparatiste de variétés de français dans l'écrit sms. In M. Abecassis et G. Ledegen, editor, *Les voix des Français*, volume 2, Modern French Identities, 94, pages 39–57. Peter Lang.

Cédrick Fairon, Jean-René Klein, and Sébastien Paumier. 2006a. Le corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation. CD-Rom, Louvain-la-Neuve, P.U.Louvain, Cahiers du Cental, 3.2.

Cédrick Fairon, Jean-René Klein, and Sébastien Paumier. 2006b. Le langage SMS. *Louvain-la-Neuve, P.U.Louvain, Cahiers du Cental*, 3.1.

Zellig S. Harris. 1968. *Mathematical Structures of Language*. Wiley-Interscience.

Laurent Kevers. 2011. *Accès sémantique aux bases de données documentaires. Techniques symboliques de traitement automatique du langage pour l'indexation thématique et l'extraction d'informations temporelles*. Ph.D. thesis, Université Catholique de Louvain.

C. Kobus, F. Yvon, and G. Damnati. 2008. Normalizing SMS : are two metaphors better than one ? In *Proceedings of COLING 2008*, pages 441–448.

Peter Koch and Wulf Österreicher. 1985. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43.

Feng Pan and Jerry R. Hobbs. 2006. Temporal arithmetic mixing months and days. In *In Proceedings of the 13th International Symposium on Temporal Representation and Reasoning (TIME)*, pages 212–217. IEEE Computer Society.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *LREC 2010*, Malta.

Stéphanie Weiser. 2010. *Repérage et typage d'expressions temporelles pour l'annotation sémantique automatique de pages Web - Application au e-tourisme*. Ph.D. thesis, Université Paris Ouest Nanterre la Défense.

François Yvon. 2008. Reorthography of SMS messages. Technical report, IMSI/CNRS, Orsay, France.