Portuguese Temporal Expressions Recognition: from TE characterization to an effective TER module implementation

Caroline Hagège[1], **Jorge Baptista**[2], **Nuno Mamede**[3]

[1]Xerox Research Centre Europe – XRCE
6 Chemin de Maupertuis – Meylan – France

[2]Universidade do Algarve, FCHS
Campus de Gambelas – Faro – Portugal

[3]L2F, INESC-ID Lisboa
Lisboa – Portugal

Caroline.Hagege@xrce.xerox.com, jbaptis@ualg.pt,Nuno.Mamede@inesc-id.pt

Abstract. *Taking into account the temporal dimension conveyed in texts is a challenge to natural language processing. At the same time this task is of great importance for a wide range of natural language processing applications. The goal of this paper is twofold. First a characterization of Portuguese temporal expressions as they appear in texts is presented. This classification is intended to meet the requirements of high inter-agreement between annotators of temporal expressions. Second, relying on this characterization, an effective temporal expression annotation tool is described. Results from its evaluation are reported.*

Resumo. *A dimensão temporal é um elemento estruturante fundamental para a informação veiculada em textos e constitui um desafio para o processamento de língua natural, sendo igualmente importante para muitas aplicações do processamento das línguas. Este artigo apresenta uma classificação das expressões temporais do português que permite esclarecer algumas incertezas relativas ao estatuto das diferentes expressões temporais e constitui uma base para a anotação intersubjectiva destas expressões. Utilizando esta classificação, foi desenvolvida uma ferramenta de anotação automática das expressões temporais do português, cujo desempenho foi avaliado.*

1. Introduction

This paper focuses on the temporal expressions (TE) appearing in Portuguese written text. Usually, dealing with temporal information involves the following steps:

- Identification and tagging of temporal expressions appearing in texts
- resolution of referential temporal expressions in order to be able to perform normalization
- identification of processes associated to the temporal expressions
- characterization of the relation between processes and temporal expressions (this include also to take into account tense and aspect)
- temporal inference

These steps are however closely interdependent for a proper treatment of temporality. This paper addresses some of the issues mentioned above. Namely, a set of guidelines is proposed to deal with the identification and tagging of temporal expressions appearing in Portuguese texts. In this characterization, the differences of referential status of these expressions are taken into account because they lead to different methods for normalization of those temporal expressions. A tool has also been developed that automatically tags temporal expressions according to those guidelines and performs a first step of temporal normalization.

Over the last years, there has been a renewed interest in temporal processing within the NLP community. It relies on the fact that a proper treatment of the temporal component in texts enables to perform better in a wide ranges of NLP tasks like question-answering, multi-document summarization, information extraction, etc. One of the references in this domain is TimeML [Saurí et al. 2006], which provides annotation guidelines of English temporal expressions and associated processes. These guidelines have currently been adapted for French (see [Bittar 2008]). Other approaches for the normalization of temporal expression are presented in [Battistelli et al. 2006]. In this approach a subset of temporal expressions (calendar expressions) are described in terms of composition of basic predefined operators. Some systems dedicated to temporal annotation have been developed and recently a competition has been organized to evaluate the accuracy of automatic temporal processors for English [Verhagen et al. 2007]. Annotations are mainly based on the TimeML guidelines mentioned before and most of systems are learning systems relying on already annotated corpora for training. Unfortunately, temporally anotated corpora are only available for English (namely TimeBank which is provided by the TimeML [Saurí et al. 2006] effort, and they are very time-consuming to build. More recently, [Parent et al. 2008] and [Hagège and Tannier 2008] presented rule-based systems annotating and normalizing temporal expressions for French and English. For Portuguese, a first step towards temporal annotation has been performed in the context of the Second HAREM competition [Mota and Santos 2008].

This paper first explains the motivations for this work, which took a stronger impetus from the participation in the HAREM evaluation campaign [Mota and Santos 2008], showing that a proper characterization of temporal expressions is not a trivial task and that it needs a wide context to be well performed. It will then introduce a set of guidelines for TE annotation and finally present the temporal annotator that has been developed and its results. The paper will conclude by summarizing what has been done so far, and what is going to be done in the future.

2. Motivation

In order to introduce motivation, the difficulty of the task will be illustrated with the following examples:

(1) *Banana **de manhã** emagrece. Será ?*
(to eat) banana in the morning helps getting slimmer. Is that right ?
(2) *Partiu **esta manhã***
(He) left this morning
(3) ***A manhã** é um momento mágico do dia*
Morning is a magical moment of the day

(4) ***Numa bela manhã**, resolveu partir*
One fine morning, he decided to leave

These examples illustrate the difficulty in dealing with temporal expressions. All these expressions are noun phrases or prepositional phrases that have the same lexical head (*manhã (morning)*). But, each expression has to be interpreted differently. As stated in [Ehrmann and Hagège 2009], any TE interpretation cannot be perform properly without taking into account the process to each it is attached.

In the first case the expression *de manhã* has to be interpreted as a frequency (i.e. a repetitive temporal expression), equivalent to *todas as manhãs (every morning)*. The second case constitutes a referential TE whose antecedent is the moment of enunciation (i.e. he left the morning of the day this utterance was produced). In the third case, one is dealing with a generic temporal expression. This means that the expression does not provide any temporal anchoring to the associate predicate. Finally, the last expression is an underspecified temporal expression in the sense that, there is no clear anchoring point to the time line for the associated process.

These examples illustrate the fact that a simple pattern-matching scheme is not enough to perform a proper TE characterization. And this may advocate against considering TE recognition as a subtask which can be included within the more general task of Named Entities Recognition (as it has been done, for instance, in [Mota and Santos 2008]).

3. Guidelines for identification and classification of Portuguese TE

One of the key points in the guidelines is that TE only can be properly classified and annotated when considered in relation to the processes they modify. This remark seems to be straightforward, however, even in TimeML guidelines [Saurí et al. 2006], some uncertainty concerning the status of TE remains, especially when they are cited without any context.

3.1. Identification

In order to objectively identify time expressions, several criteria are provided. A TE is defined by obeying both criteria 1 and 2 or, else, it is considered a generic TE, defined by criterion 3:

1. **Criterion 1** - A TE must constitute, in its context, an adequate answer to the question-answer paired sentences with interrogative forms *quando (when)*, *quanto tempo (how much time)*, eventually preceded by a preposition, or *com que frequência (how frequently)*,
2. **Criterion 2** - A TE must involve one or more of the following types of lexical items (or numeric formats). For lack of space full enumeration can not be provided here, but only some few exemples. Enumeration of the lexical items involved are reported in [Baptista et al. 2009].
    (a) numerical and alphanumeric date expressions (22-Maio-2009), both calendar dates and hour formats (12:30), including abbreviations of months, and certain adverbial-like expressions (e.g. AM, GMT, a.C.)

(b) a time-unit (*Segundo* (second)); this set also includes non-standard time-units such as *fim-de-semana (weekend)*.
(c) the nouns corresponding to the designation of some of this time-units, such as the names of the months (*janeiro (January))*, days of the week (*segunda-feira (Monday))*, adverbs derived from time-units (*diariamente (daily))*.
(d) nouns designating different kinds of holidays, of religious, political, historical, or cultural origin; name of seasons and different festivities, which may or may not include the noun *dia (day)*.
(e) simple and compound non-ambiguous time adverbs (e.g. *ontem* (yesterday) or *depois de amanhã* (after tomorrow) together with time adverbs with suffix *-mente (-ly)*.
(f) a prepositional noun phrase (PP) whose head is a generic time-related noun (e.g. *altura* (time, moment), *data* (date), *instante* (instant), *momento* (moment), *vez* (time)). The generic time-related nouns are usually accompanied by: demonstrative determinants (e.g. *nessa altura (at that time)*), other determinative adjectives, as well as quantifiers of different types, and determinative pronouns, including possessive pronouns (*no meu tempo (in my time)*); relative clauses (*na altura em que ela vivia em Lisboa (at the time that she was living in Lisbon)*); n.b.: the TE does not include the relative clause; an adjective (usually capitalized) designating a historical period (*durante o período Barroco (during the Baroque period))*; the adjective is included in the named entity.
(g) the determinative prepositional phrases involving numerals and time-unit that complement/modify predicative (event) nouns (*uma viagem de 5 dias (a 5 days trip)*); preposition *de (of)* must be included in the TE.
(h) prepositional phrases with time-units and a relative clause with verbs *passar (pass)*, *vir (come)*, or the like (these verbs constitute a closed set): *no ano que passou (last year)*, *para o mês que vem (next month)*; time-units can also present adjectival modifiers: *no ano passado (last year)*, *no próximo mês (next month)*, *durante o corrente ano (during the current year)*, *nos séculos vindouros (in the coming centuries)*.
(i) expressions with verbs *fazer (do)* or *haver (there be)* and time-units: *há três anos (three years ago)*, *faz duas semanas (two weeks ago)*;

3. Criterion 3 - The expression involves one or more of the lexical items (or numeric formats) described in Criterion 2, but it does not comply to Criterion 1. For example: *A primavera é a mais bela estação do ano (Spring is the most beautiful season)*.

3.2. Segmentation

TE include the preposition, if the TE is a prepositional phrase (PP, *no ano passado (last year)*), or the determinant if the expression is a noun phrase (NP, *dois dias depois (two days latter)*). In the case of complex, eventually ambiguous sequences, the following segmentation criteria, defined in [Hagège and Tannier 2008] are adopted:

A complex temporal expression is to be split in smaller units iff both of the following conditions are true:

1. Each component expression is syntactically valid if combined with the process that it modifies

2. Each component expression is logically implied in the complex expression; in other words, if the truth-value of the complex expression is judged as true, then the truth-value of each component expression must also be true.

For example, in *Visitei o Pedro **dois dias nesta semana** (I visited Peter two day this week)*, the complex time expression in this sentence is to be split in two since each component expression can combine with the event: *Visitei o Pedro **dois dias** / Visitei o Pedro **nesta semana** (I visited Peter two days / I visited Peter this week)*, and each partial expression is as true as the truth-value of the longer expression.

On the contrary, in the following sentence, only one TE is to be considered: *Visitei o Pedro **dois dias depois** (I visited Peter two days after(=two days later))*, since, even if each smaller expression can be syntactically combined with the event: *Visitei o Pedro **dois dias**/ Visitei o Pedro **depois** (visited Peter two days / I visited Peter later)*, the meaning of each individual combination becomes different from the global meaning of the complex expression.

3.3. Classification

The classification is proposed together with a set of criteria. This classification is inspired from previous work [Saurí et al. 2006] but it is also influenced by the result of the experience on temporal annotation that took place in the second HAREM Campaign [Baptista et al. 2008]. Finally, it is closely related to the classification made in [Ehrmann and Hagège 2009].

The main criterion used to classify TE consists on the kind of anchoring of temporal processes that they operate. Four main types are thus construed:

1. DATE – the TE corresponds to a unique anchoring of the process onto the timeline;
2. DURATION – the TE does not anchor the process onto the timeline;
3. FREQUENCY – the TE relates the process to the timeline by way of multiple anchoring instances;
4. GENERIC – the expression does not anchor any event onto the timeline. It is not really a temporal expression in the sense that no temporal information is associated to any process, but keeps a time-related meaning that may be important for the resolution of temporal references.

While the three first and main TE types can constitute an adequate answer to interrogatives[1] with *quando? (when?), (Prep) quanto tempo? ((Prep) how much time?)* or *com que frequência? (how frequently?)*, respectively, the GENERIC type can not.

Subclassification of these main types depends next on the simple or complex structure of the TE. Therefore, the DATE type is further structured in

- simple DATEs, including not only calendar dates proper, but also hour TE (e.g. *20/05/2009 11:45 TMG*)
- INTERVALs, TE involving two explicit dates (*de 5 a 15 de Maio (from May 5 to May 15)*; and
- a COMPLEX subtype corresponding to TE involving both DATE and DURATION.

[1]In order to capture all relevant types, other interrogative forms are also used, but their full listing is given in [Baptista et al. 2009]

In much the same way, the DURATION type includes a SIMPLE subtype (e.g. *A reunião durará **2 horas** (The meeting will last 2 hours)*), and an INTERVAL subtype; the later involves two quantifying expressions (*A reunião durará **entre 1 e 2 horas** (The meeting will last from 1 to 2 hours)*).

Furthermore, the DATE type is also classified based on the temporal reference of the TE and/or its indeterminacy regarding its anchoring in the timeline. In this sense, the following subtypes are distinguish:

- ABSOLUTE dates, directly computable from the TE (e.g. *em Maio de 2009 (on May 2009)*;
- RELATIVE dates, involving some temporal reference calculation; these TE are further split depending on whether they refer to the moment of ENUNCIATION (e.g. *ontem (yesterday)*) or to some other TEXTUAL element, somewhere in the text (e.g. *no dia seguinte (the following day)*).

A special feature, called *fuzzy*[2] in [Baptista et al. 2009] is marked on different types of TE. For example, as DATEs TE, they provide an anchoring point of the associated process to the time line. However, this anchoring point is not specified. For instance in:
***Numa bela manhã**, resolveu partir (One fine morning, [he] decide to leave)*
the process is anchored to the timeline, but nothing in the expression and in a broader context enables to state the precise anchoring point. The same kind of indeterminacy can be found in TE of the DURATION and FREQUENCY types.

4. A System for Temporal Expressions Recognition

4.1. XIP Temporal Annotation Module

We have developed a module for TE annotation and typing. The development was initiated in 2007 [Loureiro 2007] and deeply revised for the HAREM campaign in which we proposed a special track on temporal annotation. Our temporal processor is an extension of XIP [Aït-Mokhtar et al. 2002]. XIP is a rule-based syntactic analyzer and its architecture can be divided into the three following parts:

- a pre-processing stage handling tokenization, morphological analysis and POS tagging;
- a surface syntactic analysis stage consisting in chunking the input; a Named Entity Recognition module (NER) is included; and
- a dependency analyzer which links lexical items with labelled ordered syntactic relations.

Temporal processing is intertwinned in the general linguistic processing. TE recognition and typing is mostly performed at the surface syntactic analysis stage (local grammar rules) but a better typing needs to consider the dependency analysis in the sense that links between processes and TE are present at this stage (i.e. these links are a special kind of MODIFIER links holding between predicates and adjuncts).

At this stage our temporal module is active on the local level. Temporal expressions are recognized using ordered rewrite rules taking into consideration if necessary a right and a left context.

[2]This kind of expressions are also called *undetermined dates* in [Ehrmann and Hagège 2009] and [Gosselin 1996]

The next rule (a simplification of the *real* rule) builds a new noun node when a noun with lemma *meia-noite* or *meio-dia* is preceeded by the preposition *antes de*. The preposition is not included in the new noun node and the just created node is tagged with the feature *time* and the feature *hour*:

```
noun[hour=+,time=+] @= |?[lemma:"antes de"]|
                       ?[lemma:meia-noite]; ?[lemma:meio-dia].
```

Together with the rules, actions are associated which enable to perform normalization. These actions are calls to Python functions that can be executed directly from the parser [Roux 2006].

4.2. Results

Evaluation of the Temporal Annotation Module has taken place at the Second HAREM campaign (temporal expression annotation track). Seven systems participated on this track with different degrees of granularity (which shows the interest of the Portuguese NLP community to this topic) and only one system participated in the full task (including absolute date normalization). The results obtained by the system are quite encouraging. Considering identification and classification of temporal expressions[3], the system reached a 0.85 precision and a 0.76 recall. Some errors are due to the fact that the identification and classification procedure is performed only at the local level, and so the particular semantics of the process associated to the TE was not taken into account. Other errors were due to missing codification of temporal lexical triggers. Normalization of absolute dates and partial normalization of referential dates also produced encouraging results as the system achieved a 0.74 f-measure. However, it is the belief of the authors that only the consideration of a broader context involving the TE can improve these results.

5. Conclusion

Temporal processing is a difficult but important aspect in information extraction from texts. This line of research has been developed for some time already for different languages. For Portuguese, however, this research is still at its beginning.

One of the difficulties is to first properly characterize what is meant here by temporal expressions and to properly characterize them taking into account their referential properties, keeping in mind that the ultimate main goal is to be able to anchor the processes described by the texts in a timeline.

Precise guidelines have been developed to determine, segment and characterize temporal expressions. The development of a temporal module has also been started, which automatically recognize and classifies temporal expressions appearing in texts according to those guidelines. At this stage, the temporal module only operates at surface syntactic level, but encouraging results have already been obtained. It is, however, quite obvious that surface level is not sufficient to a precise temporal partial ordering, reference resolution or event temporal anchoring. In the near feature, the information (including temporal and aspectual information) about processes linked to these ET, together with discourse information, must be exploited, in order to perform a better temporal processing aiming at the partial ordering of processes over a timeline.

[3]Temporal characterization in HAREM guidelines is slightly different from the present classification, however they are compatible

References

Aït-Mokhtar, S., Chanod, J.-P., and Roux, C. (2002). Robustness beyond shallowness: Incremental deep parsing. In *Natural Language Engineering, 8*, pages 121–144.

Baptista, J., Hagège, C., and Mamede, N. (2008). Identificação, classificação e normalização de expressões temporais do português: A experiência do segundo harem e o futuro. In Mota, C. and Santos, D., editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*.

Baptista, J., Mamede, N., and Hagège, C. (2009). *Time Expressions in Portuguese. Guidelines for Indentification, Classification and Normalization (Internal Report Inesc-ID)*.

Battistelli, D., Minel, J.-L., and Schwer, S. (2006). Représentation des expressions calendaires dans les textes: vers une application à la lecture assistée de biographies. *Traitement Automatique des Langues*, pages 11–37.

Bittar, A. (2008). Annotation des informations temporelles dans des textes en français. In *Actes de RECITAL 2008*, Avignon, France.

Ehrmann, M. and Hagège, C. (2009). Proposition de caractérisation et de typage des expressions temporelles en contexte. In *Actes de TALN 2009*, Senlis, France.

Gosselin, L. (1996). *Sémantique de la temporalité en français. Un modèle calculatoire et cognitif du temps et de l'aspect*. Duculot.

Hagège, C. and Tannier, X. (2008). Xtm: A robust temporal text processor. In *Proceedings of CICLing 2008*, Haïfa, Israël.

Loureiro, J. M. S. (2007). Reconhecimento de Entidades Mencionadas (Obra, Valor, Relações de Parentesco e Tempo) e Normalização de Expressões Temporais. Master's thesis, Technical University of Lisbon, Instituto Superior Técnico, Lisboa, Portugal.

Mota, C. and Santos, D., editors (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Aveiro, Portugal.

Parent, G., Gagnon, M., and Muller, P. (2008). Annotation d'expressions temporelles et d'événements en français. In *Actes de TALN 2008*, Avignon, France.

Roux, C. (2006). Coupling a linguistic formalism and a script language. In *Proceedings of CSLP-06 - Coling-ACL*, Sydney, Australia.

Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). *TimeML Annotation Guidelines Version 1.2.1*.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). In *SemEval-2007 - Task 15 TempEval Temporal Relation Identification*.