It Is the Time for Portuguese Texts!

Olga Craveiro[1,2], Joaquim Macedo[3], and Henrique Madeira[2]

[1] School of Technology and Management, Polytechnic Institute of Leiria, Portugal
[2] CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
{marine,henrique}@dei.uc.pt
[3] Department of Informatics, University of Minho, Portugal
macedo@di.uminho.pt

Abstract. In this work, we introduce a software testbed for temporal processing of Portuguese texts, composed by several building blocks: identification, classification and resolution of temporal expressions and temporal text segmentation. Starting from a simple document, we can reach a set of temporally annotated segments, which enables the establishment of relationships between words and time. This temporally enriched information is then placed into an Information Retrieval system. This work represents a step forward for Portuguese language processing, with notorious lack of tools. Its main novelty is temporal segmentation of texts. Even with target application in temporal aware Information Retrieval, the described software tools can be used in other application scenarios.

1 Introduction

Time is an important dimension in understanding the text information. However, there is still much to do to achieve its full integration in the most popular retrieval models [1]. Our focus is the design, implementation and evaluation of a temporal aware Information Retrieval (IR) models.

As we work also with Portuguese text collections, our first concern was to find the available tools that allow us to reach as soon as possible the focused subject related tasks. However, with almost an inexistence of temporal information extraction tools from Portuguese texts, we decided to develop a system from scratch.

We had to start by identifying the temporal expressions in the text. Later it was necessary classify them, in order to facilitate their resolution into standard dates, where possible. Thus, we have a series of *chronons* associated with each text. A *chronon* is a normalized date which is anchored in a calendar/clock system. Finally, and as our aim is to associate the time with the words that describe entities and events or facts, we decided to do the temporal segmentation of the text. The result is a set of text segments, each with one or more dates, used to create time enriched indexes.

In the following, we present related work, with special emphasis on Portuguese language processing. Subsequently, we present the testbed software architecture, with a concise description of the components and their relationships. At the end, we have a more detailed evaluation and the conclusions and further work.

H. Caseli et al. (Eds.): PROPOR 2012, LNAI 7243, pp. 106–112, 2012.

2 Related Work

There are several tools for extraction of temporal expressions from English texts. Most of them process term by term, using linguistic features for their identification [2]. An annotation scheme, using finite state automata, represents dates and times based on a diverse set of manual defined and automatically discovered rules [3]. In [4], it is described a different approach based on Context-Scanning Strategy. The TARSQUI is also a very popular toolkit [5]. There are several temporal expressions resolution strategies [3,6]. Unfortunately, these strategies or tools are not directly applicable to the Portuguese language. For Portuguese, the temporal information extraction area is not yet focus of significant amount of work. We found only the XTM tool, recently developed by Hagège et al. [7]. It is rule-based and processes word by word, using a deep analysis approach to extract temporal information from texts. XTM is not available in the public domain.

In terms of topic segmentation of texts, the literature is quite extensive [8-9]. But, about the temporal one it is very limited. To the best of our knowledge, there is no research work focused on Portuguese text temporal segmentation. Bramsen [10] also worked on the temporal segmentation of texts. However, despite being mainly interested in the chronological order of the segments, the application domain is the clinical narratives.

In temporal extraction, the novelty of our proposal is to be supported by the existence of lexical patterns generated automatically from Portuguese texts. It also uses an inductive approach that starts from the data to the knowledge, unlike above referenced work. The distinction of our segmentation approach is the division of the text into temporally coherent units. It is not concerned with segment chronological order, but with association between segments and accurate dates, for later use of this information. The use of a rule based algorithm, instead of machine learning one, is justified by the lack of suitable Portuguese training collections.

3 Testbed System Architecture

The objective of our work is the temporal enrichment of the IR system. So, we want to establish a relationship between words and time. We are building up a testbed system which extracts temporal information from Portuguese texts. Despite we are used the four modules together, each one can be used individually. It can be used for instance for temporal characterization of a text collection. In testbed system design, we are searching for a tradeoff between simplicity and efficiency, while maintaining a suitable effectiveness level.

In the Fig. 1 is shown a tool used to extract relevant temporal information from Portuguese documents which is composed by three modules: Co-Occurrence Processor (henceforth COP), Annotator and Resolver. The recognition task is carried out by the COP and the Annotator modules. The interpretation and normalization of the temporal expressions into *chronons*, normalized dates which are anchored in a calendar/clock system, are performed by the Resolver module (see section 4).

Fig. 2 shows the module for the temporal segmentation of the Portuguese texts. Based on content and metadata temporal information of documents, this single module partitions texts into temporally coherent segments, like in topical segmentation. These segments are also tagged with timestamps which are the *chronons* collected from the text segment (see section 5).

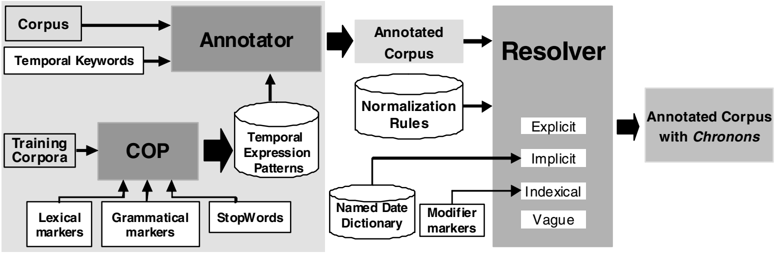

Fig. 1. System Architecture for Temporal Information Extraction

4 Temporal Expressions Extraction

Due to space limitations, we could not include the details of our extraction tool which are reported in [11,12]. A brief description of their modules is presented below.

The temporal information extraction task is divided into recognition and resolution of temporal expressions. The identification and semantic classification of temporal expressions are in the recognition part. Temporal expressions are classified as date, hour, interval, duration, and frequency, following the guidelines defined in [13].

Our recognition method is based on a two-stage approach, each stage being carried out by a different module (see Fig. 1). Firstly, the COP module produces semantically classified temporal patterns, based on regular expressions. The patterns are created using word co-occurrences, determined from training corpora and a pre-defined set of seed keywords derived from the used language temporal references. These reference words are divided in lexical [1] and grammatical [2] markers. An example of the COP output (pattern, classification) is: (*In the (last | following)* ***year***, DATE).

Then, these patterns are used by the Annotator module to annotate Portuguese temporal expressions. The Annotator processes each sentence to determine whether it matches any of the temporal patterns, and, if so, the sentence is annotated with semantic classification corresponding to the matching pattern in the original text. An example of the Annotator:

```
I run <EM ID="2" CATEG="TIME" TYPE="FREQUENCY">every day</EM>.
```

The resolution comprises interpretation and normalization of the temporal expressions. The interpretation of temporal expressions consists in the inference of a new date, using information from document. The normalization is the transformation of the dates into a standard format. Our definition of temporal expressions as explicit, implicit, relative and vague is supported by Schilder et al. [6]. Relative references are expressions that need one point in time to be completely resolved and anchored in calendar/clock system and can be classified as *deictic timexes* or *anaphoric timexes* [14].

The Resolver module maps temporal expressions found in documents' content into a discrete representation of time, denoted *chronons* by Alonso [15]. A *chronon* is a normalized date which is anchored in a calendar/clock system by timelines defined by points.

[1] Some examples: months, seasons, weekdays, Christmas/*Natal*, day/*dia,* today/*hoje, …*

[2] Some examples: prepositions {in/*em,* since/*desde, …*}, ordinal adjectives {next/*próximo, …*}.

Each timeline has a different level of granules, such as year, month, week, day and hour. This module relies on a set of rules, used to interpret temporal expressions previously annotated by the Annotator module. It starts with the document timestamp normalization, a time related metadata, such as the creation or publication data of the document. This date is used in *deictic timexes* resolution. In the *anaphoric timexes* resolution is used a date evoked in the text. Since indexical references can mention a past, present or future event, it is also required the modifiers of these references. The correct rule to be applied is chosen by these modifiers, such as *next*, *after*. For example, *next year* is resolved with the rule *"document timestamp + 1 time_Unit(y)"*.

Our system has also a named date dictionary used to resolve the implicit references. For example, *Christmas Day 2011* is normalized as *2011-12-25*.

5 Temporal Segmentation of Text

Our segmentation algorithm makes use of temporal information extracted from the text to divide text into temporally coherent segments. These segments are tagged with a timestamp to obtain an association between time and document terms. A temporal segment is defined as a set of adjacent sentences that shares the same temporal focus. The segment length ranges from a single sentence to a multi-paragraph text. Thus, adjacent sentences with the same *chronons* must belong to the same segment. Each segment is tagged with all the different *chronons* found in their sentences. The document timestamp is also associated to each segment.

Some examples are presented below. The segment of the first example is composed by two sentences. The second sentence will also belong to this segment, since the topic is the same. The other example shows a segment tagged with two *chronons*, because these two normalized date are in the same sentence.

```
(1) <SEGMENT DN="2011-10-31">Sunday's storm caused some problems
    in electricity networks. The company of electricity received
    about 31,000 calls.</SEGMENT>
(2) <SEGMENT DN="2011-11-10 2011-11-11">It rained on Friday and
    Saturday.</SEGMENT>
```

Fig. 2 shows the temporal segmentation module. The segmentation process begins at the sentence level. Each sentence is a candidate to start a new segment and it is compared with the *current segment*, composed by the previous adjacent sentences of the text. There are two approaches defined by the temporal information analysis of the sentence: sentence with *chronons* and sentence without *chronons*.

If the *chronons* of the sentence are equal to the *chronons* of the *current segment*, the sentence must belong to this segment; otherwise, this sentence starts a new segment.

If the sentence has no *chronon,* a temporal mark in the three first words of the sentence determine if the sentence starts or not a segment. These marks are used to express temporal relation between successive actions or events and can signal the topic continuity or discontinuity [16]. Thus, if the sentence has a continuity marker, e. g. *and*, it remains within the *current segment*. If the marker expresses discontinuity, e.g. *next*, this sentence starts a new segment. If there is not any marker in the sentence, the sentence must be processed with the similarity calculation. Before this calculation, the sentence and the *current segment* are pre-processed for removing punctuation marks and stopwords. Based on the vector space model, the topic cohesion of the sentence and the *current segment* is computed using the cosine measure, according to the approach used in some topic segmentation methods [17]. A threshold value is also defined to decide if the sentence starts or not a new segment.

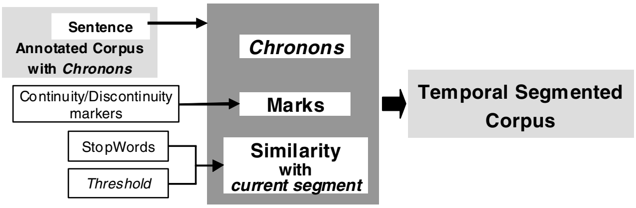
Sentence
Annotated Corpus with *Chronons*
Continuity/Discontinuity markers
StopWords
Threshold
Chronons
Marks
Similarity with *current segment*
Temporal Segmented Corpus

Fig. 2. System Architecture for Temporal Segmentation

6 Evaluation and Results

In this evaluation, we intended to verify the effectiveness of the system. Indeed, such evaluation has become an hard task, namely for the Temporal Segmentation module as we explain below. Each module was evaluated using a Portuguese text collection. Unfortunately, due to space limitations, we could not include the details of temporal information extraction modules, namely COP, Annotator and Resolver, which are reported in [11-12].

The collections used in our experiments are subsets of the Second HAREM Collection[3] (henceforth HC), properly detailed in [13]. Since there is no Portuguese collection for temporal segmentation we created two HC subsets: Temporal Segmentation Training Set, composed by 4 documents with 82 sentences and 1,195 words, and, Temporal Segmentation Evaluation Set which has 28 documents with 401 sentences and 6,678 words.

The evaluation of the Temporal Segmentation module was a complex task. The selection of the reference segmentation is difficult since the detection of the boundaries of topics involves an inherent subjectivity. In order to solve this difficulty we compared our algorithm against a manual segmentation corpus based on human judgments. Since human judges do not always agree, we measured the agreement between judges removing the probability of chance agreement using a commonly used measure, *Kappa coefficient*.

The documents of the corpus were manually annotated and segmented by two human judges. We observed an agreement of 0.91 and the *Kappa* value was equal to 0.82. So, the corpus is appropriated for the evaluation [18].

For this evaluation, our algorithm was implemented in Perl Language and named as *Time4Word* (henceforth *T4W*). The *T4W* input are an annotated and normalized corpus, a list of continuity[4] and discontinuity[5] marks, a stopwords list composed by prepositions, conjunctions, articles and pronouns of the Portuguese language, and, a threshold value for the similarity measure, as is shown in Fig. 2.

[3] Available at Linguateca site: http://www.linguateca.pt/HAREM.

[4] Some examples: e/*and*, também/*also*, (n)este/*(in) this*, (n)esse/*(in) that*, eles/*they*, *(…)*.

[5] Some examples: após/*after*, antes/*before*, depois/*next*, mais tarde/*later*, então/*then*, *(...)*.

Besides the traditional measure of accuracy, we decided to use also the WindowDiff (*WD*) measure. This measure, a variation of the P_k metric, was proposed in [19] as the suitable measure for the evaluation of the text segmentation tasks. *T4W* was evaluated considering not only the boundaries of the segment but also the timestamp of the segment. As the segment length is variable, the width of window (*k*) used in the calculation of *WD* was set to the average of the segment length in the reference segmentation, using the sentence as the unit. The average of the segment length in the evaluation set is 1.18, *k* was set to 1. We made several evaluations varying the threshold of the cosine similarity between 0.01 and 0.35. Table 1 shows the minimum and the maximum values of *WD*. The best result (0.15) was obtained when the threshold was set to 0.04. As the threshold increases, the number of false positives increases as well.

Since a segment timestamp can have more than one *chronon*, we calculated the agreement, overlapping and disagreement, analyzing the matching between the timestamps of the segments obtained by *T4W* and the reference segmentation. Table 1 shows such values considering the same variation of the cosine similarity threshold. Indeed, the difference between the minimum and maximum values is not very significant.

The results obtained show a good effectiveness, even with some limitations. Although the accuracy was about 78%, the *WD* was not so penalized (0.*15*). This means that some incorrect boundaries of the segment are within the *k*-sentence window used by *WD*. The incorrect boundaries of the segments have been determined in sentences without dates and where it was applied the similarity calculation. The use of synonyms and a stemmer will certainly improve these results.

Table 1. Agreement, overlapping and disagreement of the segment timestamp

WD, *k=1*	Agreement	Overlapping	Disagreement
0.15 – 0.193	76% - 79%	1.75% - 2%	19.5% - 22.5%

7 Conclusion and Future Work

We introduced a software testbed for temporal processing of Portuguese texts. Even with a set of limitations and simplifications, our system has shown promising results in the effectiveness evaluation of each module (a precision in the range of 78%-84% and a recall in the range of 64%-75%).

The main contribution of this paper is the original text segmentation method which uses temporal information of documents (metadata and contents), for divide the text into temporal coherent parts, allowing a relationship between words and time which will be used to enrich a temporal aware IR index. Supported by a simple rule-based algorithm, despite the auspicious result of 0.15 for *WD*, the method can be improved using phrases as minimal segment size and, for instance, thesaurus and stemming for topic change detection.

References

1. Alonso, O., Strötgen, J., Baeza-Yates, R., Gertz, M.: Temporal Information Retrieval: Challenges and Opportunities. In: 1st International Temporal Web Analytics Workshop (TWA-WWW 2011), pp. 1–8 (2011)

2. Mani, I.: Recent developments in temporal information extraction. In: RANLP, Borovets, Bulgaria, pp. 45–60 (2003)
3. Mani, I., Wilson, G.: Robust temporal processing of news. In: 38th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, pp. 69–76 (2000)
4. Vazov, N.: A system for extraction of temporal expressions from French texts based on syntactic and semantic constraints. In: ACL 2001 Workshop on Temporal and Spatial Information Processing, Toulouse, France (2001)
5. Verhagen, M., Pustejovsky, J.: Temporal processing with the TARSQI toolkit. In: COLING, ACL, Morristown, USA, pp. 189–192 (2008)
6. Schilder, F., Habel, C.: From temporal expressions to temporal information: Semantic tagging of news messages. In: ACL 2001 Workshop on Temporal and Spatial Information Processing, Toulouse, France, pp. 65–72 (2001)
7. Hagège, C., Baptista, J., Mamede, N.J.: Caracterização e processamento de expressões temporais em português. Linguamática 2(1), 63–76 (2010)
8. Misra, H., Yvon, F., Jose, J.M., Cappe, O.: Text segmentation via topic modeling: an analytical study. In: CIKM 2009, pp. 1553–1556. ACM, New York (2009)
9. Misra, H., Yvon, F., Cappé, O., Jose, J.: Text segmentation: a topic modeling perspective. Information Processing and Management 47(4), 528–544 (2011)
10. Bramsen, P., Deshpande, P., Lee, Y.K., Barzilay, R.: Finding temporal order in discharge summaries. In: AMIA 2006, Washington DC, USA, pp. 81–85 (2006)
11. Craveiro, O., Macedo, J., Madeira, H.: Use of Co-occurrences for Temporal Expressions Annotation. In: Karlgren, J., Tarhio, J., Hyyrö, H. (eds.) SPIRE 2009. LNCS, vol. 5721, pp. 156–164. Springer, Heidelberg (2009)
12. Craveiro, O., Macedo, J., Madeira, H.: Leveraging temporal expressions for segmented-based information retrieval. In: ISDA, pp. 754–759. IEEE (2010)
13. Mota, C., Santos, D. (eds.): Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca (2008)
14. Ahn, D., Adafre, S.F., de Rijke, M.: Extracting temporal information from open domain text: A comparative exploration. In: DIR 2005, pp. 3–10 (2005)
15. Alonso, O., Gertz, M., Baeza-Yates, R.: Clustering and exploring search results using timeline constructions. In: CIKM 2009, pp. 97–106. ACM, New York (2009)
16. Bestgen, Y., Vonk, W.: The role of temporal segmentation markers in discourse processing. Discourse Processes 19, 385–406 (1995)
17. Hearst, M.A.: Multi-paragraph segmentation of expository text. In: 32nd Annual Meeting on Association for Computational Linguistics, pp. 9–16. ACL (1994)
18. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 22, 249–254 (1996)
19. Pevzner, L., Hearst, M.A.: A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics 28, 19–36 (2002)