TimeBankPT: A TimeML Annotated Corpus of Portuguese

Francisco Costa, António Branco

University of Lisbon
fcosta@di.fc.ul.pt, Antonio.Branco@di.fc.ul.pt

Abstract

In this paper, we introduce TimeBankPT, a TimeML annotated corpus of Portuguese. It has been produced by adapting an existing resource for English, namely the data used in the first TempEval challenge. TimeBankPT is the first corpus of Portuguese with rich temporal annotations (i.e. it includes annotations not only of temporal expressions but also about events and temporal relations). In addition, it was subjected to an automated error mining procedure that checks the consistency of the annotated temporal relations based on their logical properties. This procedure allowed for the detection of some errors in the annotations, that also affect the original English corpus. The Portuguese language is currently undergoing a spelling reform, and several countries where Portuguese is official are in a transitional period where old and new orthographies are valid. TimeBankPT adopts the recent spelling reform. This decision is to preserve its usefulness in the future. TimeBankPT is freely available for download.

Keywords: Corpora, Temporal Information Extraction, Portuguese

1. Introduction

There has been recent interest in the extraction of temporal information from natural language text. The TERN 2004 (`http://timex2.mitre.org/tern.html`) evaluation campaign focused on the automated identification of dates and times in text, and their normalization (i.e. converting natural language expressions like *yesterday* or *last year* into a number-based representation).

The recent TempEval (Verhagen et al., 2007) and TempEval-2 (Verhagen et al., 2010) challenges, part of SemEval-2007 (Agirre et al., 2007) and SemEval-2010 (Erk and Strapparava, 2010) respectively, went farther and explored the automated classification of temporal relations holding between the events and the times and dates mentioned in a text. An annotation scheme called TimeML (Pustejovsky et al., 2003a) has gained prominence, and several corpora have been released with TimeML annotations, mostly in association with the TempEval campaigns. TimeML is a rich annotation scheme in so far as it allows for the annotation of several phenomena related to time: the times, dates and periods denoted by temporal expressions, events, temporal relations, etc.

TempEval-2 released TimeML annotated data for Chinese, English, French, Italian, Korean and Spanish. The previous TempEval had made available data for English. Both English data sets (that used in TempEval and that used in TempEval-2) are based on a previous English corpus annotated with TimeML, the TimeBank (Pustejovsky et al., 2003b). These are all the languages so far with data featuring rich annotations about time.

The TimeBank and the data used in the two TempEval challenges are important, as they have annotations describing not just dates and times, but also events and temporal relations between these entities. The TempEval challenges also changed the focus of temporal information processing to the temporal relations (previously it had been mostly on temporal expressions). This is an important problem, as the task of automatically classifying temporal relations (e.g. identifying whether the temporal relation holding between two specific events mentioned in a text is one of temporal overlap or temporal precedence) is hard, and there is still much improvement to be made.

In this paper, we describe TimeBankPT, a corpus of Portuguese annotated with TimeML. The corpus has around 70,000 words, a size similar to what is available for other languages with this sort of annotation. TimeBankPT is based on the English data used in the first TempEval.

2. Methodology

In order to create TimeBankPT, we adapted the existing English data that were used in the first TempEval, which are annotated with TimeML.

In a first step, all TimeML markup in the English TempEval data was removed. The result was then given as input to the Google Translator Toolkit,[1] which combines machine translation with a translation memory. A human translator corrected the proposed translations manually.

After that, there are three collections of documents (the original TimeML annotated data, the English unannotated data and the Portuguese unannotated data). These three collections are aligned by paragraphs: the line breaks in the the original collection are simply maintained in the other collections. Therefore, for each paragraph in the Portuguese data all the corresponding TimeML tags in the original English paragraph are known.

A small script was developed to place all relevant TimeML markup at the end of each paragraph in the Portuguese text. Each TimeML markup element was then manually placed in the correct place in that paragraph. At this point some necessary changes to the annotations were also done manually. These are motivated by language differences (e.g. the inventory of verb tenses is language specific; see Section 3). This approach involving manual steps is feasible because the original TempEval corpus is not very large. It must be noted that the TimeML annotations that describe the temporal relations (the `<TLINK>` elements; see Section 3) always occur at the end of each document, each in a separate line, separate from the text: therefore they do not need to be repositioned.

[1] `http://translate.google.com/toolkit`

This methodology is also described in (Costa and Branco, 2010; Costa, to appear). The main difference between the data obtained by the methodology that is reported there and the data being described in the present paper is the application of automated error detection, which is explained below in Section 4. The cited work also contains a more detailed description of the points where the annotations employed in TimeBankPT differ from the ones in the English TempEval data due to language differences, and the motivation for some of the annotation decisions that had to be made because of such differences.

3. Annotations

Figure 1 contains a fragment of the data in TimeBankPT, together with the corresponding original English data, which is part of the annotated corpus used in the first TempEval.

In the annotation scheme that is employed, words that denote events are enclosed in `<EVENT>` elements. The attributes that are appropriate for these elements are `tense`, `aspect`, `class`, `polarity`, `pos`, `stem`. The `stem` is the term's lemma, and `pos` is its part-of-speech. Grammatical tense and aspect are encoded in the features `tense` and `aspect`. The attribute `polarity` takes the value `NEG` if the event term is in a negative syntactic context, and `POS` otherwise. The attribute `class` contains several levels of information. It makes a distinction between terms that denote actions of speaking, which take the value `REPORTING` and those that do not. For these, it distinguishes between states (value `STATE`) and non-states (value `OCCURRENCE`), and it also encodes whether they create an intensional context (value `I_STATE` for states and value `I_ACTION` for non-states).

Temporal expressions (timexes) are inside `<TIMEX3>` elements. The most important features for these elements are `value`, `type` and `mod`. The timex's `value` encodes a normalized representation of this temporal entity, its `type` can be e.g. `DATE`, `TIME` or `DURATION`. The `mod` attribute is optional. It is used for expressions like *early this year*, which are annotated with `mod="START"`. As can be seen in Figure 1 there are other attributes for timexes that encode whether it is the document's creation time (`functionInDocument`) and whether its value can be determined from the expression alone or requires other sources of information (`temporalFunction` and `anchorTimeID`). It is important to note that the corpus is divided in several documents, and every document contains the annotation of one special time—the document's creation time—that is the time when the document was created.

The `<TLINK>` elements encode temporal relations. The attribute `relType` of these elements represents the type of relation. The attribute `eventID` is a reference to the first argument of that relation. The second argument is given by the attribute `relatedToTime` (if it is a time, a date or a duration) or `relatedToEvent` (if it is another event).

The temporal relations are divided in three groups: in group A one finds the relations holding between an event and a temporal expression occurring in the same sentence; in group B, temporal relations between events and the document's creation time; and in group C there are relations holding between the main events of two adjacent sentences. These groups reflect the three tasks of the first TempEval. The `task` attribute of `TLINK` elements is the name of the TempEval task to which this temporal relation pertains (`A`, `B` or `C`).

4. Automated Error Mining

These data have been used extensively for over a year now for experimental work. During this time, a few errors, already present in the English source, have been detected and corrected.

It is possible to automatically detect errors in temporal annotation. For instance, if an event *A* is annotated as temporally preceding another event *B*, and *B* is annotated as preceding *C*, *A* must precede *C* as well, because temporal precedence is a transitive relation. If we then find an annotation according to which *C* precedes *A*, we have a temporal loop, and something is wrong. We have run a temporal reasoning system on the adapted data, which enabled us to detect this kind of error.

The original TempEval data had been similarly checked for consistency (Verhagen, 2005). However, our reasoning component performs one extra step, that allowed us to identify more errors: before applying any temporal reasoning rules, it first orders annotated temporal expressions according to their normalized value (e.g. the date `1989-09-29` is ordered as preceding `1989-10-02`). That is, we exploit the `TIMEX3` annotations in order to enrich the set of temporal relations that we work with, and more specifically we make use of the `value` attribute of `TIMEX3` elements. In this way, we end up working with many more temporal relations than those explicitly annotated. All temporal relations that are explicitly annotated are binary and involve at least one event. Our approach further adds a large number of temporal relations between dates or times.

The corpus distribution contains a file where each error that was discovered with the help of temporal reasoning is described. This file serves as documentation about the changes introduced during the adaptation process, but from these descriptions it is also easy to identify the corresponding data in the original English corpus.

The inference procedure allowed for the detection of around 80 problems in the corpus (affecting both the train and the test sets), that were then manually corrected. These corrections result in some differences between TimeBankPT and the original TempEval English data. Since they affect the type of the annotated temporal relations, they cause differences in the distribution of temporal relations. In Section 6, we quantify the effect of these corrections on the data, by comparing the distribution of temporal relations in TimeBankPT with that in the English TempEval data set.

Several authors have used reasoning as a means to aid temporal annotation. Katz and Arosio (2001) used a temporal reasoning system to compare the temporal annotations of two annotators. In a similar spirit, Setzer and Gaizauskas (2001) first compute the deductive closure of annotated temporal relations so that they can then assess annotator agreement with standard precision and recall measures.

```
<s><TIMEX3 tid="t18" type="DATE" value="1998-01-11" temporalFunction="true"
functionInDocument="NONE" anchorTimeID="t14">Hoje</TIMEX3> há helicópteros a <EVENT
eid="e7" class="OCCURRENCE" stem="sobrevoar" aspect="NONE" tense="INF"
polarity="POS" pos="VERB">sobrevoar</EVENT> o norte de Nova Iorque a <EVENT eid="e8"
class="I_ACTION" stem="tentar" aspect="NONE" tense="INF" polarity="POS"
pos="VERB">tentar</EVENT> <EVENT eid="e9" class="OCCURRENCE" stem="localizar"
aspect="NONE" tense="INF" polarity="POS" pos="VERB">localizar</EVENT> pessoas <EVENT
eid="e10" class="OCCURRENCE" stem="isolar" aspect="NONE" tense="PPA"
polarity="POS" pos="VERB">isoladas</EVENT> sem alimentos, aquecimento ou medicamentos.</s>
<TLINK lid="l3" relType="OVERLAP" eventID="e8" relatedToTime="t18" task="A"/>
```

```
<s>Helicopters are <EVENT eid="e7" class="OCCURRENCE" stem="fly" aspect="PROGRESSIVE"
tense="PRESENT" polarity="POS" pos="VERB">flying</EVENT> over northern New York <TIMEX3
tid="t18" type="DATE" value="1998-01-11" temporalFunction="true"
functionInDocument="NONE" anchorTimeID="t14">today</TIMEX3> <EVENT eid="e8"
class="I_ACTION" stem="try" aspect="NONE" tense="PRESPART" polarity="POS"
pos="VERB">trying</EVENT> to <EVENT eid="e9" class="OCCURRENCE" stem="locate"
aspect="NONE" tense="INFINITIVE" polarity="POS" pos="VERB">locate</EVENT> people
<EVENT eid="e10" class="OCCURRENCE" stem="strand" aspect="NONE" tense="PASTPART"
polarity="POS" pos="VERB">stranded</EVENT> without food, heat or medicine.</s>
<TLINK lid="l3" relType="OVERLAP" eventID="e8" relatedToTime="t18" task="A"/>
```

Figure 1: Example fragment taken from TimeBankPT, in the upper box. The original English annotation is shown in the lower box. The raw text is *Helicopters are flying over northern New York today trying to locate people stranded without food, heat or medicine.* The event *e8*, denoted by the term *tentar* "trying", overlaps the date *t18*, denoted by the term *hoje* "today".

Verhagen (2005) uses temporal closure as a means to aid TimeML annotation. He reports that closing a set of manually annotated temporal relations more than quadruples the number of temporal relations in TimeBank (Pustejovsky et al., 2003b), a corpus that is the source of the data used for the TempEval challenges.

A considerable amount of work in the area of temporal information processing—for which this sort of data is useful—has used reasoning components in the proposed solutions. One recent example is the work of Ha et al. (2010), a participant of the second TempEval, but there are several others.

4.1. Ordering of Dates and Times

As mentioned already, temporal expressions are ordered according to their normalized value. For instance, the date `2000-01-03` is ordered as preceding the date `2010-03-04`. Since all temporal expressions are normalized in the annotated data, we order temporal expressions before applying any temporal reasoning. This increases the number of temporal relations we start with, and the potential number of relations we end up with after applying temporal reasoning.

To this end, we used Joda-Time 2.0 (`http://joda-time.sourceforge.net`). Each normalized date or time is converted to an interval.

In many cases it is possible to specify the start and end points of this interval. For instance, the date `2000-01-03` is represented internally by an interval with its start point at `2000-01-03T00:00:00.000` and ending at `2000-01-03T23:59:59.999`. Many different kinds of normalized expressions require many rules. For instance, an expression like *last Winter* could be annotated in the data as `2010-WI`, and dedicated rules are used to get its start and end points.

Some time expressions are normalized as `PRESENT_REF` (e.g. *now*), `PAST_REF` (*the past*) or `FUTURE_REF` (*the future*). These cases are not represented by any Joda-Time object. Instead we need to account for them in a special way. They can be temporally ordered among themselves (e.g. `PRESENT_REF` precedes `FUTURE_REF`), but not with other temporal expressions. We further stipulate that `PRESENT_REF` includes each document's creation time (which therefore precedes `FUTURE_REF`, etc.). So, in additional to the representation of times and dates as time intervals, we employ a layer of *ad-hoc* rules.

The variety of temporal expressions makes it impossible to provide a full account of the implemented rules in this paper; more details are in (Costa, to appear).

Chambers and Jurafsky (2008) also order dates symbolically before applying reasoning to increase the number of explicit temporal relations. Their work is, however, more limited: they only order dates (we also order times); when doing so they only look at the year, month and day of the month (the normalized value of temporal expressions can be represented by resorting to other fields, such as the season of the year, which we explore). In addition, our work uses a richer set of temporal relations (we allow for inclusion relations between dates/times) and a richer set of reasoning rules.

4.2. Deduction Procedure

The rules implemented in our reasoning component are:

- Temporal precedence is transitive, irreflexive and antisymmetric;
- Temporal overlap is reflexive and symmetric;
- If A does not precede B, then either B precedes A or A and B overlap;
- If A overlaps B and B precedes C, then C does not precede A.

Because we also consider temporal relations between times and dates, we also deal with temporal inclusion, a type of temporal relation that is not part of the annotations used in the TempEval data but that is still useful for reasoning. We make use of the following additional rules, dealing with temporal inclusion:

- Temporal inclusion is transitive, reflexive and antisymmetric;
- If A includes B, then A and B overlap;
- If A includes B and C overlaps B, then C overlaps A;
- If A includes B and C precedes A, then C precedes B;
- If A includes B and A precedes C, then B precedes C;
- If A includes B and C precedes B, then either C precedes A or A and C overlap (A cannot precede C).
- If A includes B and B precedes C, then either A precedes C or A and C overlap (C cannot precede A).

5. Description of the Corpus

The original English data for TempEval are organized in two data sets: one for training and development and another one for evaluation. The full data are organized in 182 documents (162 documents in the training data and another 20 in the test data). Each document is a news report from television broadcasts, newswire or newspapers. A large amount of the documents (123 in the training set and 12 in the test data) are taken from several issues of the Wall Street Journal dating from 1989. These texts are usually smaller than the other ones, and contain a large amount of jargon and stock market data. Therefore, the corpus is mostly quite domain specific.

TimeBankPT contains the same translated documents.

5.1. Qualitative Description

As mentioned before, TimeBankPT has annotations similar to those of the data prepared for the first TempEval, which follow an annotation scheme similar to TimeML. Event terms are identified and marked with `<EVENT>` tags. Temporal expressions are annotated inside `<TIMEX3>` elements. Temporal relations holding between these events and the dates, times or durations denoted by the temporal expressions are represented with `TLINK` elements. These elements have attributes that refer to the two arguments of the relation, and the `relType` attribute encodes the relation type. Its possible values are `BEFORE`, `AFTER`, `OVERLAP`, `BEFORE-OR-OVERLAP`, `OVERLAP-OR-AFTER` and `VAGUE`, but the last three values occur rarely.

	Train	Test
Sentences	2,281	351
Words	60,782	8,920
Annotations		
Events	6,790	1,097
Temporal expressions	1,244	165
Temporal relations	5,781	758
Task A	1490	169
Task B	2556	331
Task C	1735	258
Words/events	8.95	8.13
Words/temporal expressions	48.86	54.06

Table 1: Size of the corpus and number of annotations

5.2. Quantitative Description

The major difference between the original TempEval corpus in English and TimeBankPT is the number of words, which is due to language differences, with Portuguese being more verbose than English (the English data consist of 52,740 words for training and 8,107 words for evaluation). Table 1 presents some quantitative information about TimeBankPT.

6. Evaluation

We ran several off-the-shelf machine learning algorithms on the adapted data and compared the results with those published for the English data, using the same classifiers.

(Hepple et al., 2007) present the results of training Weka (Witten and Frank, 2005) classifiers on the English TempEval data. The problem here is to guess the relation type (the value of the `relType` attribute mentioned earlier) of the annotated temporal relations: `BEFORE`, `AFTER`, `OVERLAP`, etc. The temporal relations themselves are already identified, as well as their arguments, and all other annotations (about events, times, etc.) are given. Many of these annotations are used as classifier attributes.

All classifier features are based on the textual string and on the remaining TimeML annotations. We used the same set of features and the same algorithms as the ones that were used for English by Hepple et al. (2007). Table 2 shows the classifier features employed by Hepple et al. (2007) and also by us for this experiment.

In this table, the features grouped under EVENT are based on the attributes of TimeML `EVENT` elements with the same name. The ones grouped under TIMEX3 are taken from `TIMEX3` elements. The ORDER attributes are computed by simple string manipulation of the TimeML annotated documents: `event-first` encodes whether the event term appears in the document

Type	Attribute	Task A	B	C
EVENT	aspect	✓	✓	✓
	polarity	✓	✓	×
	POS	✓	✓	✓
	stem	✓	×	×
	string	×	×	×
	class	×	✓	✓
	tense	×	✓	✓
ORDER	adjacent	✓	N/A	N/A
	event-first	✓	N/A	N/A
	event-between	×	N/A	N/A
	timex-between	×	N/A	N/A
TIMEX3	mod	✓	×	N/A
	type	✓	×	N/A
TLINK	relType	✓	✓	✓

Table 2: Features used by Hepple et al. (2007) in TempEval.

		F-Measure	
Group	Algorithm	English	Portuguese
A	`KStar`	0.59	0.58
	Baseline	0.57	0.59
B	`DecisionTable`	0.73	0.77
	Baseline	0.56	0.56
C	`SMO`	0.54	0.54
	Baseline	0.47	0.47

Table 3: Performance of some classifiers on the test data

before the timex; `event-between` whether there is an annotated event term in the text between the two entities; `timex-between` is similar, but considers temporal expressions; and `adjacent` is true if and only if both `event-between` and `timex-between` are false (even if some textual material actually occurs between the two annotated elements). The last feature is the class attribute (what the classifiers are supposed to guess).

We also used the same implementation of the algorithms—namely Weka's (Witten and Frank, 2005): `KStar`, a nearest neighbors algorithm for task A (Cleary and Trigg, 1995); a decision table for task B (Kohavi, 1995); and `SMO`, a support vector algorithm for task C (Platt, 1998). These classifier and feature combinations are optimized for English, but they serve our purpose of comparing the two data sets.

Table 3 shows their results alongside ours. The results presented are for classifiers trained on the entire training set and evaluated on the test set.

The results in Table 3 show that, despite language differences and the additional corrections performed on the Portuguese data, the results on the two data sets are nevertheless quite comparable. From these results we conclude that the adaptation was not lossy.

The most salient difference, when it comes to classifier performance, is for task B, with a 4% difference between the English data and the Portuguese data. After inspecting the models produced by classification algorithms that produce human readable models (such as the RIPPER algorithm (Cohen, 1995) or decision trees (Quinlan, 1993)) for this task, we see that verb tense is the most important feature used by them. Because verb tense is language specific, we hypothesize that it is the differences in the tense system of the two languages that are behind the differences in the results for task B (i.e. they are due to language differences).

The other tasks do not seem to be so sensitive to tense. It makes sense that it is precisely task B that is affected the most by it, as task B is about temporal relations holding between events and the document's creation time, and verb tense is primarily an indicator of the temporal relation between the event denoted by the verb and the speech time.

In Table 3 we also present the majority class baselines for each task. The differences in the baselines between TimeBankPT and the TempEval corpus of English are due to the corrections to the data resulting from the automated error mining procedure described in Section 4. Table 4 shows the class distributions for the three TempEval tasks, both for the English data used in TempEval and for TimeBankPT, in full detail. As can be seen from that table, the differences are very small.

6.1. Size of the Corpus

A corpus of approximately 70,000 words is small for many natural language processing tasks. In order to check whether the size of TimeBankPT is adequate for the tasks that it is meant to address (automatic temporal relation classification), one can measure the effect of the size of the data on classifier performance.

Figure 2 shows the performance of classifiers similar to the ones in Table 3 but trained with subsets of the training data. They were evaluated on the whole test set.

The machine learning algorithms employed to get the values shown there are the same as the ones in Table 3. The models were produced using the same feature set, too. Each value used to plot that graph is the average of ten samplings of the training data that differ only in as much as they use different seeds for the random number generator involved in the sampling process.

The performance of the classifiers for the three sorts of temporal relations appears quite stable across many sizes of training data. Classifier performance does go up with more training data, but it does so very slowly. Therefore, more data would likely not increase classifier performance very quickly.

Figure 3 shows similar data, this time using subsets of the test data. That is, the classifiers trained with the full training set were tested with subsets of the test data of different sizes. Each data point is also the average of ten runs that used the same amount of test data but different seeds to the random number generator used to sample the data. Once again, it can be seen that the curves are rather stable after an initial range of very short test data sizes, where, precisely because of the small size of the test data, the curves are a bit erratic and variation is high (not visible in that graph). This problem is more obvious in Figure 3 than in Figure 2 be-

		Task A		Task B		Task C	
Set	Class	EN	PT	EN	PT	EN	PT
Train	BEFORE	19%	19%	62%	62%	25%	25%
	AFTER	25%	25%	14%	14%	18%	17%
	OVERLAP	50%	49%	19%	19%	42%	42%
	BEFORE-OR-OVERLAP	2%	2%	2%	2%	4%	4%
	OVERLAP-OR-AFTER	2%	2%	1%	1%	3%	3%
	VAGUE	2%	2%	2%	1%	9%	9%
Test	BEFORE	12%	11%	56%	56%	23%	23%
	AFTER	18%	18%	15%	15%	16%	16%
	OVERLAP	57%	59%	24%	25%	47%	47%
	BEFORE-OR-OVERLAP	1%	2%	2%	3%	5%	5%
	OVERLAP-OR-AFTER	3%	3%	1%	0%	3%	3%
	VAGUE	8%	7%	2%	2%	6%	6%

Table 4: Class distributions for the three tasks, the two data sets in each corpus (train and test) and the two corpora (English, EN, and Portuguese, PT).

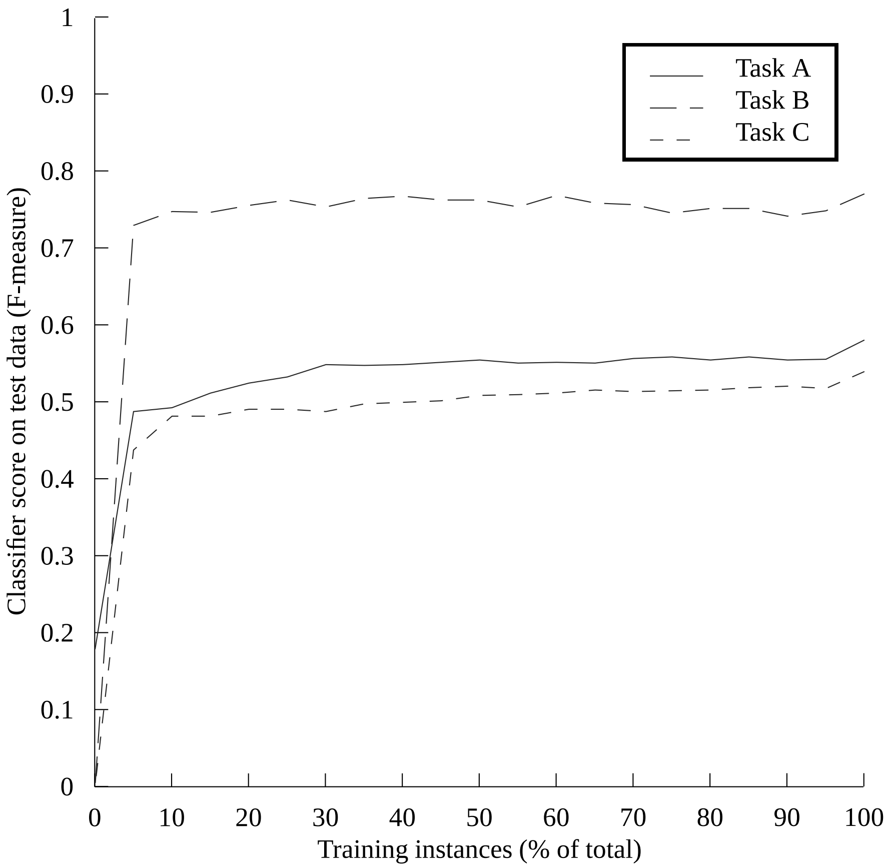

Figure 2: Classifier performance by size of training data

cause the test data set is considerably smaller than the train data set (see Table 1).

From these two results we conclude that it appears that increasing the size of the corpus would not rapidly increase classifier performance.

7. A Note on Spelling

The spelling of the Portuguese language is currently in the middle of a reform. The new spelling (Houaiss, 1991) is known as the 1990 spelling agreement but its coming into effect is quite recent.[2] It unifies the two official orthographies that existed for Portuguese: the Brazilian spelling, followed by Brazil, and the European spelling, followed by the remaining Portuguese speaking countries.

The new orthography has already been ratified in five countries (Brazil, Cape Verde, East Timor, Guinea-Bissau, Portugal and São Tomé and Príncipe). Only two countries

[2] An official document with the spelling agreement can be found at `http://www.dre.pt/pdf1s/1991/08/193A00/43704388.pdf`.

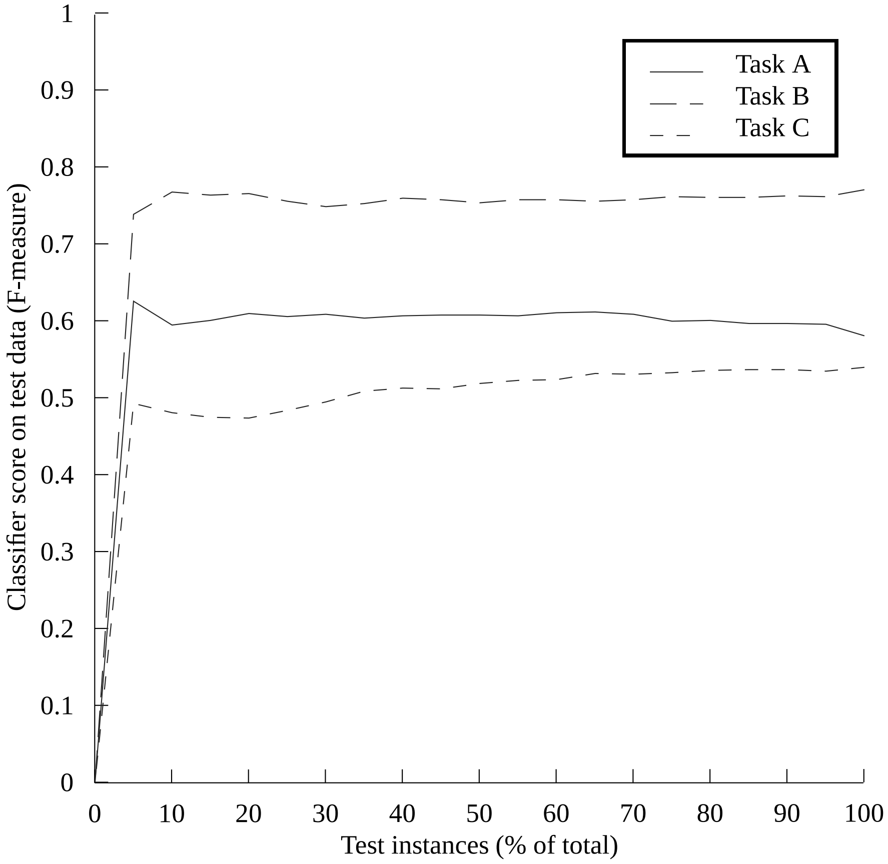

Figure 3: Classifier performance by size of test data

where Portuguese is official (Angola and Mozambique) have yet to ratify it. In 2009, several countries, including Brazil and Portugal, initiated a transitional period in which the old spellings are still acceptable, in parallel with the new ones.

The most noticeable change to the spelling is, from the Brazilian point of view, the deletion of diacritic marks in some words. In many cases the European spelling did not use them already. So for instance, *ideia* ("idea") is now written like that by all speakers, whereas the old Brazilian spelling is *idéia*, and similarly for the word *frequente* ("frequent"), with the older spelling *freqüente*. The most striking change to the European orthography is the removal of silent consonants (consonants that were written solely because of etymology but had no phonological basis), that had already been abandoned in the Brazilian spelling. One example is the word *ótimo* ("great"), which has the old European spelling *óptimo*, with a silent *p*.

TimeBankPT features the unified orthography, so that the corpus remains useful for future research on the long run. This decision has, however, negative short term consequences, as the typical existing natural language processing tools, developed for the old spellings, may not have been updated yet. Error rates may be higher currently when processing data with the new spelling, as some frequent words are now out-of-vocabulary (because they have a different spelling) for the natural language processing tools not yet updated.

8. Concluding Remarks

In this paper we presented TimeBankPT, a corpus of Portuguese with rich temporal annotations that is available for free.
This is a novel resource for Portuguese. Although there is data for this language containing annotated temporal expressions, full temporal annotation—with events and temporal relations—had not been released before. In addition, by increasing the set of languages for which this kind of annotated data are available, we hope to stimulate research on temporal information processing, where a lot of progress can still be made.
Furthermore, we reported on a sophisticated method to automatically detect errors in the corpus. It is based on existing methodology that was also employed in the creation of the original data on which TimeBankPT is based, i.e. the English data used in the first TempEval. However, we expanded this automated error mining procedure in such a way that more errors were detected.
Finally, we also tried to check whether the resulting data would be useful, by replicating the results obtained for English for the problem of temporal information extraction (more specifically the classification of temporal relations), and whether the size of TimeBankPT, which is a small corpus, is adequate for this task, which it is intended to serve.
TimeBankPT has already been used to develop a temporal annotation tool for Portuguese (Costa and Branco, 2012a; Costa and Branco, 2012b).
The data are freely available at `http://nlx.di.fc.ul.pt/~fcosta/TimeBankPT`.

9. References

Eneko Agirre, Lluís Márquez, and Richard Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, June.

Nathanael Chambers and Daniel Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 698–706, Honolulu, Hawaii. Association for Computational Linguistics.

John G. Cleary and Leonard E. Trigg. 1995. K*: An instance-based learner using an entropic distance measure. In *12th International Conference on Machine Learning*, pages 108–114.

William W. Cohen. 1995. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123.

Francisco Costa and António Branco. 2010. Temporal information processing of a new language: Fast porting with minimal resources. In *ACL2010—Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Francisco Costa and António Branco. 2012a. Extracting temporal information from Portuguese texts. In *Proceedings of PROPOR2012*.

Francisco Costa and António Branco. 2012b. LX-TimeAnalyzer: A temporal information processing system for Portuguese. Technical report, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática.

Francisco Costa. to appear. *Processing Temporal Information in Unstructured Documents*. Ph.D. thesis, Universidade de Lisboa, Lisbon.

Katrin Erk and Carlo Strapparava, editors. 2010. *SemEval 2010—5th International Workshop on Semantic Evaluation—Proceedings of the Workshop*. Uppsala University, Uppsala, Sweden, July.

Eun Young Ha, Alok Baikadi, Carlyle Licata, and James C. Lester. 2010. NCSU: Modeling temporal relations with Markov logic and lexical ontology. In Katrin Erk and Carlo Strapparava, editors, *SemEval 2010—5th International Workshop on Semantic Evaluation—Proceedings of the Workshop*, pages 341–344, Uppsala, Sweden. Uppsala University.

Mark Hepple, Andrea Setzer, and Rob Gaizauskas. 2007. USFD: Preliminary exploration of features and classifiers for the TempEval-2007 tasks. In *Proceedings of SemEval-2007*, pages 484–487, Prague, Czech Republic. Association for Computational Linguistics.

Antônio Houaiss. 1991. *A Nova Ortografia da Língua Portuguesa*. Ática, São Paulo.

Graham Katz and Fabrizio Arosio. 2001. The annotation of temporal information in natural language sentences. In *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, Toulouse, France. Association for Computational Linguistics.

Ron Kohavi. 1995. The power of decision tables. In *8th European Conference on Machine Learning*, pages 174–189.

John Platt. 1998. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Chris Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods—Support Vector Learning*.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5, Fifth International Workshop on Computational Semantics*.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656.

John Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Andreas Setzer and Robert Gaizauskas. 2001. A pilot study on annotating temporal relations in text. In *ACL 2001 Workshop on Temporal and Spatial Information Processing*.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, and James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval temporal relation identification. In *Proceedings of SemEval-2007*.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In Katrin Erk and Carlo Strapparava, editors, *SemEval 2010—5th International Workshop on Semantic Evaluation—Proceedings of the Workshop*, pages 51–62, Uppsala, Sweden. Uppsala University.

Marc Verhagen. 2005. Temporal closure in an annotation environment. In *Language Resources and Evaluation*, number 39, pages 211–241.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco. second edition.