Aspectual Type and Temporal Relation Classification

Francisco Costa
Universidade de Lisboa
fcosta@di.fc.ul.pt

António Branco
Universidade de Lisboa
Antonio.Branco@di.fc.ul.pt

Abstract

In this paper we investigate the relevance of aspectual type for the problem of temporal information processing, i.e. the problems of the recent TempEval challenges.

For a large list of verbs, we obtain several indicators about their lexical aspect by querying the web for expressions where these verbs occur in contexts associated with specific aspectual types.

We then proceed to extend existing solutions for the problem of temporal information processing with the information extracted this way. The improved performance of the resulting models shows that (i) aspectual type can be data-mined with unsupervised methods with a level of noise that does not prevent this information from being useful and that (ii) temporal information processing can profit from information about aspectual type.

1 Introduction

Extracting the temporal information present in a text is relevant to many natural language processing applications, including question-answering, information extraction, and even document summarization, as summaries may be more readable if they follow a chronological order.

Recent evaluation campaigns have focused on the extraction of temporal information from written text. TempEval (Verhagen et al., 2007), in 2007, and more recently TempEval-2 (Verhagen et al., 2010), in 2010, were concerned with this problem. Additionally, they provided data that can be used to develop and evaluate systems that can automatically temporally tag natural language text. These data are annotated according to the TimeML (Pustejovsky et al., 2003) scheme.

Figure 1 shows a small and slightly simplified fragment of the data from TempEval, with TimeML annotations. There, event terms, such as the term referring to the event of releasing the tapes, are annotated using `EVENT` tags. States (such as the situations denoted by verbs like *want* or *love*) are also considered events. Temporal expressions, such as *today*, are enclosed in `TIMEX3` tags. The attribute `value` of time expressions holds a normalized representation of the date or time they refer to (e.g. the word *today* denotes the date `1998-01-14` in this example). The `TLINK` elements at the end describe temporal relations between events and temporal expressions. For instance, the event of the plane going down is annotated as temporally preceding the date denoted by the temporal expression *today*.

The major tasks of these two TempEval evaluation challenges were about guessing the type of temporal relations, i.e. the value of the `relType` attribute of the `TLINK` elements in Figure 1, all other annotations being given. Temporal relation classification is also the most interesting problem in temporal information processing. The other relevant tasks (identifying and normalizing temporal expressions and events) have a longer research history and show better evaluation results.

TempEval was organized in three tasks (TempEval-2 has four additional ones, that are not relevant to this work): task A was concerned with classifying temporal relations holding between an event and a time mentioned in the same sentence (although they could be syntactically unrelated, as the temporal relation represented by the `TLINK` with the `lid` with the value `l1` in Figure 1); task

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 266–275,
Avignon, France, April 23 - 27 2012.

```
<s>In Washington <TIMEX3 tid="t53" type="DATE"
value="1998-01-14">today</TIMEX3>, the Federal
Aviation Administration <EVENT eid="e1"
class="OCCURRENCE" stem="release"
aspect="NONE" tense="PAST" polarity="POS"
pos="VERB">released</EVENT> air traffic control tapes from
<TIMEX3 tid="t54" type="TIME"
value="1998-XX-XXTNI">the night</TIMEX3> the TWA
Flight eight hundred <EVENT eid="e2"
class="OCCURRENCE" stem="go" aspect="NONE"
tense="PAST" polarity="POS"
pos="VERB">went</EVENT> down.</s>
<TLINK lid="l1" relType="BEFORE" eventID="e2"
relatedToTime="t53"/>
<TLINK lid="l2" relType="OVERLAP"
eventID="e2" relatedToTime="t54"/>
```

Figure 1: Sample of the data annotated for TempEval, corresponding to the fragment: *In Washington today, the Federal Aviation Administration released air traffic control tapes from the night the TWA Flight eight hundred went down.*

	Task		
	A	B	C
Best system	0.62	0.80	0.55
Average of all participants	0.56	0.74	0.51
Majority class baseline	0.57	0.56	0.47

Table 1: Results for English in TempEval (F-measure), from Verhagen et al. (2009)

B focused on the temporal relation between events and the document's creation time, which is also annotated in TimeML (not shown in that Figure); and task C was about classifying the temporal relation between the main events of two consecutive sentences. The possible values for the type of temporal relation are BEFORE, AFTER and OVERLAP.[1]

Table 1 shows the results of the first TempEval evaluation. The results of TempEval-2 are fairly similar (Verhagen et al., 2010), but the data used are similar but not identical.

The best system in TempEval for tasks A and B (Puşcaşu, 2007) combined statistical and knowledge based methods to propagate temporal constraints along parse trees coming from a syntactic parser. The best system for task C (Min et al., 2007) also combined rule-based and machine learning approaches. It employed sophisticated NLP to compute some of the features used; more specifically it used syntactic features.

Our goal with this work is to evaluate the impact of information about aspectual type on these tasks. The TimeML annotations include an attribute `class` for EVENTs that encodes some aspectual information, distinguishing between stative (annotated with the value STATE) and non-stative events (value OCCURRENCE). This attribute is relevant to the classification problem at hand, i.e. it is a useful feature for machine learned classifiers for the TempEval tasks (although this `class` attribute encodes other kinds of information as well). However, aspectual distinctions can be more fine-grained than a mere binary distinction, and so far no system has explored this sort of information to help improve the solutions to temporal relation classification.

In this paper we work with Portuguese, but in principle there is no reason to believe that our findings would not apply to other languages that display similar aspectual phenomena, such as English. Some of the details, such as the material in Section 4.2, are however language specific and would need adaptation.

2 Aspectual Type

Distinctions of aspectual type (also referred to as situation type, lexical aspect or *Aktionsart*) of the sort of Vendler (1967) and Dowty (1979) are expected to improve the existing solutions to the problem of temporal relation classification. The major aspectual distinctions are between (i) states (e.g. *to hate beer*, *to know the answer*, *to own a car*, *to stink*), (ii) processes, also called activities (*to work*, *to eat ice cream*, *to grow*, *to play the piano*), (iii) culminated processes, also called accomplishments (*to paint a picture*, *to burn down*, *to deliver a sermon*) and (iv) culminations, also called achievements (*to explode*, *to win the game*, *to find the key*). States and processes are atelic situations in that they do not make salient a specific instant in time. Culminated processes and culminations are telic situations: they have an intrinsic, instantaneous endpoint, called the culmination (e.g. in the case of *to paint a picture*, it is the moment when the picture is ready; in the case of *to explode*, it is the moment of the explosion).

There are several reasons to think aspectual

[1] There are the additional disjunctive values BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER and VAGUE, employed when the annotators could not make a more specific decision, but these affect a small number of instances.

type is relevant to temporal information processing. First, these distinctions are related to how long events last: culminations are punctual, whereas states can be very prolonged in time. States are thus more likely to temporally overlap other temporal entities than culminations, for instance.

Second, there are grammatical consequences on how events are anchored in time. Consider the following examples, from Ritchie (1979) and Moens and Steedman (1988):

(1) When they built the 59^{th} Street bridge, they used the best materials.

(2) When they built that bridge, I was still a young lad.

The situation of building the bridge is a culminated processed, composed by the process of actively building a bridge followed by the culmination of the bridge being finished. In sentence (1), the event described in the main clause (that of using the best materials) is a process, but in sentence (2) it is a state (the state of being a young lad). Even though the two clauses in each sentence are connected by *when*, the temporal relations holding between the events of each clause are different. On the one hand, in sentence (1) the event of using the best materials (a process) overlaps with the process of actively building the bridge and precedes the culmination of finishing the bridge. On the other hand, in sentence (2) the event of being a young lad (which is a state) overlaps with both the process of actively building the bridge and the culmination of the bridge being built. This difference is arguably caused by the different aspectual types of the main events of each sentence.

As another example, states overlap with temporal location adverbials, as in (3), while culminations are included in them, as in (4).

(3) He was happy last Monday.

(4) He reached the top of Mount Everest last Monday.

In other cases, differences in aspectual type can disambiguate ambiguous linguistic material. For instance, the preposition *in* is ambiguous as it can be used to locate events in the future but also to measure the duration of culminated processes; it is thus ambiguous with culminated processes, as in *he will read the book in three days* but not with other aspectual types, as in *he will be living there in three days*.

A factor related to aspectual class, that is not trivial to account for, is the phenomenon of aspectual shift, or aspectual coercion (Moens and Steedman, 1988; de Swart, 1998; de Swart, 2000). Many linguistic contexts pose constraints on aspectual type. This does not mean, however, that clashes of aspectual type cause ungrammaticality. What often happens is that phrases associated with an incompatible aspectual type get their type changed in order to be of the required type, causing a change in meaning.

For instance, the progressive construction combines with processes. When it combines with e.g. a culminated process, the culmination is stripped off from this culminated process, which is thus converted into a process. The result is that a sentence like (5) does not say that the bridge was finished (the event has no culmination), whereas one such as (6) does say this (the event has a culmination).

(5) They were building that bridge.

(6) They built that bridge.

Aspectual type is not a property of just words, but phrases as well. For example, while the progressive construction just mentioned combines with processes, the resulting phrase behaves as a state (cf. the sentence *When they built the 59^{th} Street bridge, they were using the best materials* and what was mentioned above about *when* clauses).

3 Strategy

Aspectual type is hard to annotate. This is partly because of what was just mentioned: it is not a property of just words, but rather phrases, and different phrases with the same head word can have different aspectual types; however annotation schemes like TimeML annotate the head word as denoting events, not full phrases or clauses.

For this reason, our strategy is to obtain aspectual type information from unannotated data. Because these data are gradient—an event-denoting word can be associated with different aspectual types, depending on word sense—we do not aim to extract categorical information, but rather nu-

meric values for each event term that reflect associations to aspectual types. These may be seen as values that are indicative of the frequencies in which an event term denotes a state, or a process, etc.

In order to extract these indicators, we resort to a methodology sometimes referred to as Google Hits: large amounts of queries are sent to a web search engine (not necessarily Google), and the number of search results (the number of web pages that match the query) is recorded and taken as a measure of the frequency of the queried expression.

This methodology is not perfect, since multiple occurrences of the queried expression in the same web page are not reflected in the hit count, and in many cases the hit counts reported by search engines are just estimates and might not be very accurate. Additionally, uncarefully formulated queries can match expressions that are syntactically and semantically very different from what was intended. In any case, it has the advantages of being based on a very large amount of data and not requiring any manual annotation, which can introduce errors.

3.1 The Web as a Very Large Corpus

Hearst (1992) is one of the earliest studies where specific textual patterns are used to extract lexico-semantic information from very large corpora. The author's goal was to extract hyponymy relations. With the same goal, Kozareva et al. (2008) apply similar textual patterns to the web.

The web has been used as a corpus by many other authors with the purpose of extracting syntactic or semantic properties of words or relations between them, e.g. Ravichandran and Hovy (2002), Etzioni et al. (2004), etc. Some of this work is specially relevant to the problem of temporal information processing. VerbOcean (Chklovski and Pantel, 2004) is a database of web mined relations between verbs. Among other kinds of relations, it includes typical precedence relations, e.g. *sleeping* happens before *waking up*. This type of information has in fact been used by some of the participating systems of TempEval-2 (Ha et al., 2010), with good results.

More generally, there is a large body of work focusing on lexical acquisition from corpora. Just as an example, Mayol et al. (2005) learn subcategorization frames of verbs from large amounts of data. Relevant to our work is that of Siegel and McKeown (2000). The authors guess the aspectual type of verbs by searching for specific patterns in a one million word corpus that has been syntactically parsed. They extract several linguistic indicators and combine them with machine learning algorithms. The indicators that they extract are naturally different from ours, since they have access to syntactic structure and we do not, but our data are based on a much larger corpus.

3.2 Textual Patterns as Indicators of Aspectual Type

Because of aspectual shift phenomena (see Section 2), full syntactic parsing is necessary in order to determine the aspectual type of a natural language expression. However, this can be approximated by frequencies: it is natural to expect that e.g. stative verbs occur more frequently in stative contexts than non-stative verbs, even if there may be errors in determining these contexts if syntactic parsing is not a possibility.

If one uses Google Hits, syntactic information is not accessible. In return for its impreciseness, Google Hits have the advantage of being based on very large amounts of data.

4 Scope and Approach

In this study we focus exclusively on verbs, but events can be denoted by words belonging to other parts-of-speech. This limitation is linked to the fact that the textual patterns that are used to search for specific aspectual contexts are sensitive to part-of-speech (i.e. what may work for a verb may not work equally well for a noun).

In order to assess whether aspectual type information is relevant to the problem of temporal relation classification, our approach is to check whether incorporating that kind of information into existing solutions for this problem can improve their performance. TimeML annotated data, such as those used for TempEval, can be used to train machine learned classifiers. These can then be augmented with attributes encoding aspectual type information and their performance compared to the original classifiers.

Additionally, we work with Portuguese data. This is because our work is part of an effort to implement a temporal processing system for Portuguese. We briefly describe the data next.

<s>*Em Washington,* <TIMEX3 tid="t53" type="DATE" value="1998-01-14">*hoje*</TIMEX3>*, a Federal Aviation Administration* <EVENT eid="e1" class="OCCURRENCE" stem="publicar" aspect="NONE" tense="PPI" polarity="POS" pos="VERB">*publicou*</EVENT> *gravações do controlo de tráfego aéreo da* <TIMEX3 tid="t54" type="TIME" value="1998-XX-XXTNI">*noite*</TIMEX3> *em que o voo TWA800* <EVENT eid="e2" class="OCCURRENCE" stem="cair" aspect="NONE" tense="PPI" polarity="POS" pos="VERB">*caiu*</EVENT>.</s>
<TLINK lid="l1" relType="BEFORE" eventID="e2" relatedToTime="t53"/>
<TLINK lid="l2" relType="OVERLAP"
eventID="e2" relatedToTime="t54"/>

Figure 2: Sample of the Portuguese data adapted from the TempEval data, corresponding to the fragment: *Em Washington, hoje, a Federal Aviation Administration publicou gravações do controlo de tráfego aéreo da noite em que o voo TWA800 caiu.*

4.1 Data

Our experiments used TimeBankPT (Costa and Branco, 2010; Costa and Branco, 2012; Costa, to appear). This corpus is an adaptation of the original TempEval data to Portuguese, obtained by translating it and then adapting the annotations. Figure 2 shows the Portuguese equivalent to the sample presented above in Figure 1. The two corpora are quite similar, but there is of course the language difference. TimeBankPT contains a few corrections to the data (mostly the temporal relations), but these corrections only changed around 1.2% of the total number of annotated temporal relations (Costa and Branco, 2012). Although we did not test our results on English data, we speculate that our results carry over to other languages.

Just like the original English corpus for TempEval, it is divided in a training part and a testing part. The numbers (sentences, words, annotated events, time expressions and temporal relations) are fairly similar for the two corpora (the English one and the Portuguese one).

4.2 Extracting the Aspectual Indicators

We extracted the 4,000 most common verbs from a 180 million word corpus of Portuguese newspaper text, CETEMPúblico. Because this corpus is not annotated, we used a part-of-speech tagger and morphological analyzer (Barreto et al., 2006; Silva, 2007) to detect verbs and to obtain their dictionary form. We then used an inflection tool (Branco et al., 2009) to generate the specific verb forms that are used in the queries. They are mostly third person singular forms of several different tenses.

The indicators that we used are ratios of Google Hits. They compare two queries.

Several indicators were tested. We provide examples with the verb *fazer* "do" for the queries being compared by each indicator. The name of each indicator reflects the aspectual type being tested, i.e. states should present high values for State Indicators 1 and 2, processes should show high values for Process Indicators 1–4, etc.

- State Indicator 1 (Indicator **S1**) is about imperfective and perfective past forms of verbs. It compares the number of hits a for an imperfective form *fazia* "did" to the number of hits b for a perfective form *fez* "did": $\frac{a}{a+b}$. Assuming the imperfective past constrains the entire clause to be a state, and the perfective past constrains it to be telic, the higher this value the more frequently the verb appears in stative clauses in a past tense.[2]

- State Indicator 2 (Indicator **S2**) is about the co-occurrence with *acaba de* "has just finished". It compares the number of hits a for *acaba de fazer* "has just finished doing" to the number of hits b for *fazer* "to do": $\frac{b}{a+b}$. In Portuguese, this construction does not seem to be felicitous with states.

- Process Indicator 1 (Indicator **P1**) is about past progressive forms and simple past forms (both imperfective). It compares the number of hits a for *fazia* "did" to the number of hits b for *estava a fazer* "was doing": $\frac{b}{a+b}$. Assuming the progressive construction is a function from processes to states (see Section 2), the higher this value, the more likely the verb can occur with the interpretation of a process.

[2]We expect this frequency to be indicative of states because states can appear in the imperfective past tense with their interpretation unchanged, whereas non-stative events have their interpretation shifted to a stative one in that context (e.g. they get a habitual reading). In order to refer to an event occurring in the past with an on-going interpretation, non-stative verbs require the progressive construction to be used in Portuguese, whereas states do not. Therefore, states should occur more freely in the simple imperfective past.

- Process Indicator 2 (Indicator **P2**) is about past progressive forms vs. simple past forms (perfective). It compares the number of hits a for *fez* "did" to the number of hits b for *esteve a fazer* "was doing": $\frac{b}{a+b}$. Similarly to the previous indicator, this one tests the frequency of a verb appearing in a context typical of processes.

- Process Indicator 3 (Indicator **P3**) is about the occurrence of *for* Adverbials. It compares the number of hits a for *fez* "did" to the number of hits b for *fez durante muito tempo* "did for a long time": $\frac{b}{a+b}$. This number is also intended to be an indication of how frequent a verb can be used with the interpretation of a process. Note that Portuguese allows modifiers to occur freely between a verb and its complements, so this test should work for transitive verbs (or any other subcategorization frame involving complements), not just intransitive ones.

- Process Indicator 4 (Indicator **P4**) is about the co-occurrence of a verb with *parar de* "to stop". It compares the number of hits a for *parou de fazer* "stopped doing" to the number of hits b for *fazer* "to do": $\frac{a}{a+b}$. Just like the English verbs *stop* and *finish* are sensitive to the aspectual type of their complement, so is the Portuguese verb *parar*, which selects for processes.

- Atelicity Indicator 1 (Indicator **A1**) is about comparing *in* and *for* adverbials. It compares the number of hits a for *fez num instante* "did in an instant" to the number of hits b for *fez durante muito tempo* "did for a long time": $\frac{b}{a+b}$. Processes can be modified by *for* adverbials, whereas culminated processes are modified by *in* adverbials. This indicator tests the occurrence of a verb in contexts that require these aspectual types.

- Atelicity Indicator 2 (Indicator **A2**) is about comparing *for* Adverbials with *suddenly*. It compares the number of hits a for *fez de repente* "did suddenly" to the number of hits b for *fez durante muito tempo* "did for a long time": $\frac{b}{a+b}$. *De repente* "suddenly" seems to modify culminations, so this indicator compares process readings with culmination readings.

- Culmination Indicator1 (Indicator **C1**) is about differentiating culminations and culminated processes. It compares the number of hits a for *fez de repente* "did suddenly" to the number of hits b for *fez num instante* "did in an instant": $\frac{a}{a+b}$.

For each of the 4,000 verbs, the necessary queries required by these indicators were generated and then sent to a search engine. The queries were enclosed in quotes, so as to guarantee exact matches. The number of hits was recorded for each query.

We had some problems with outliers for a few rather infrequent verbs. These could show very extreme values for some indicators. In order to minimize their impact, for each indicator we homogenized the 100 highest values that were found. More specifically, for each indicator, each one of the highest 100 values was replaced by the 100^{th} highest value. The bottom 100 values were similarly changed. This way the top 99 values and the bottom 99 values are replaced by the 100^{th} highest value and the 100^{th} lowest value respectively.

Each indicator ranges between 0 and 1 in theory. In practice, we seldom find values close to the extremes, as this would imply that some queries would have close to 0 hits, which does not occur very often (after all, we intentionally used queries for which we would expect large hit counts, as these are more likely to be representative of true language use). For this reason, each indicator is scaled so that its minimum (actual) value is 0 and its maximum (actual) value is 1.

5 Evaluation

As mentioned before, in order to assess the usefulness of these aspectual indicators for the tasks of temporal relation classification, we checked whether they can improve machine learned classifiers trained for this problem. We next describe the classifiers that were used as the bases for comparison.

5.1 Experimental Setup

In order to obtain bases for comparison, we trained machine learned classifiers on the Portuguese corpus TimeBankPT, that is adapted from the TempEval data (see Section 4.1). We took inspiration in the work of Hepple et al. (2007).

This was one of the participating systems of TempEval. It used machine learning algorithms implemented in Weka (Witten and Frank, 1999). For our experiments, we used Weka's implementation of the C4.5 algorithm, `trees.J48` (Quinlan, 1993), the RIPPER algorithm as implemented by Weka's `rules.JRip` (Cohen, 1995), a nearest neighbors classifier, `lazy.KStar` (Cleary and Trigg, 1995), a Naïve Bayes classifier, namely Weka's `bayes.NaiveBayes` (John and Langley, 1995), and a support vector classifier, Weka's `functions.SMO` (Platt, 1998) . We chose these algorithms as they are representative of a wide range of machine learning approaches.

Recall that the tasks of TempEval are to guess the type of temporal relations. Each train or test instance thus corresponds to a temporal relation, i.e. a `TLINK` element in the TimeML annotations (see Figures 1 and 2). The classification problem is to determine the value of the attribute `relType` of TimeML `TLINK` elements. These temporal relations relate an event (referred by the `eventID` attribute of `TLINK` elements) to another temporal entity, that can be a time (pointed to by the `relatedToTime` attribute), in the case of tasks A and B, or, in the case of task C, another event (given by the `relatedToEvent` attribute).

As for the features that were employed, we also took inspiration in the approach of Hepple et al. (2007). These authors used as classifier attributes two types of features. The first group of features corresponds to TimeML attributes: for instance the value of the `aspect` attribute of `EVENT` elements, for the events involved in the temporal relation to be classified. The second group of features corresponds to simple features that can be computed with string manipulation and do not require any kind of natural language processing.

Table 2 shows the features that were tried and employed.

The *event* features correspond to attributes of `EVENT` elements, with the exception of the `event-string` feature, which takes as value the character data inside the corresponding TimeML `EVENT` element. In a similar spirit, the *timex3* features are taken from the attributes of `TIMEX3` elements with the same name. The `tlink-relType` feature is the class attribute and corresponds to the `relType` attribute of the TimeML `TLINK` element that represents the temporal relation to be classified. The *order* features are the attributes computed from the document's textual content. The feature `order-event-first` encodes whether the event terms precedes in the text the time expression it is related to by the temporal relation to classify. The classifier attribute `order-event-between` describes whether any other event is mentioned in the text between the two expressions for the entities that are in the temporal relation, and similarly `order-timex3-between` is about whether there is an intervening temporal expression. Finally, `order-adjacent` is true iff both `order-timex3-between` and `order-event-between` are false (even if other linguistic material occurs between the expressions denoting the two entities in the temporal relation).

	Task		
Attribute	A	B	C
event-aspect	×	✓	✓
event-polarity	✓	✓	✓
event-POS	×	×	✓
event-stem	×	✓	×
event-string	✓	×	×
event-class	✓	×	✓
event-tense	✓	✓	✓
order-event-first	✓	N/A	N/A
order-event-between	✓	N/A	N/A
order-timex3-between	×	N/A	N/A
order-adjacent	✓	N/A	N/A
timex3-mod	✓	×	N/A
timex3-type	×	×	N/A
tlink-relType	✓	✓	✓

Table 2: Feature combinations used in the classifiers used as comparison bases. Features inspired by the ones used by Hepple et al. (2007) in TempEval.

In order to arrive at the final set of features (marked with a check mark in Table 2), we performed exhaustive search on all possible combinations of these features for each task, using the Naïve Bayes algorithm. They were compared using 10-fold cross-validation on the training data. The feature combinations shown in Table 2 are the optimal combinations arrived at in this way.

These are the classifiers that we used for the

comparison with the aspectual type indicators. We chose this straightforward approach because it forms a basis for comparison that is easily reproducible: the algorithm implementations that were used are part of freely available software, and the features that were employed are easily computed from the annotated data, with no need to run any natural language processing tools whatsoever.

As mentioned before in Section 4.1, the data used are organized in a training set and an evaluation set. The training part is around 60K words long, the test data containing around 9K words. When tested on held-out data, these classifiers present the scores shown in italics in Table 3.

These results are fairly similar to the scores that the system of Hepple et al. (2007) obtained in TempEval with English data: 0.59 for task A, 0.73 for task B, and 0.54 for task C. They are also not very far from the best results of TempEval. As such they represent interesting bases for comparison, as improving their performance is likely to be relevant to the best systems that have been developed for temporal information processing.

	Task		
Classifier	A	B	C
`trees.J48`	*0.57*	*0.77*	*0.53*
With best indicator			**0.55**
`rules.JRip`	*0.60*	*0.76*	*0.51*
With best indicator	**0.61**		**0.54**
`lazy.KStar`	*0.54*	*0.70*	*0.52*
With best indicator		**0.73**	**0.53**
`bayes.NaiveBayes`	*0.50*	*0.76*	*0.53*
With best indicator	**0.53**		**0.54**
`functions.SMO`	*0.55*	*0.79*	*0.54*
With best indicator	**0.56**		**0.55**

Table 3: Evaluation on held-out test data of classifiers trained on full train data. Values for the classifiers used as comparison bases are in italics. Boldface highlights improvements resulting from incorporating aspectual indicators as classifier features, and missing values represent no improvement.

5.2 Results and Discussion

After obtaining the bases for comparison described above, we proceeded to check whether the aspectual type indicators described in Section 4.2 can improve these results.

For each aspectual indicator, we implemented a classifier feature that encodes its value for the event term in the temporal relation (if it is not a verb, this value is missing). In the case of task C, two features are added for each indicator, one for each event term.

We extended each of these classifiers with one of these features at a time (two in the case of task C), and checked whether it improved the results on the test data. So for instance, in order to test Indicator S1, we extended each of these classifiers with a feature that encodes the value that this indicator presents for the term that denotes the event present in the temporal relation to be classified. In the case of task C, two classifier features are added, one for each event term, and both for the same Indicator S1. For instance, for the (training) instance corresponding to the `TLINK` in Figure 2 with the `lid` attribute that has the value `l1`, the classifier feature for Indicator S1 has the value that was computed for the verb *cair* "go down", since this is the `stem` of the word that denotes the event that is the first argument of this temporal relation. After adding each of these features, we retrained the classifiers on the training data and tested them on the held-out test data. In order to keep the evaluation manageable, we did not test combinations of multiple indicators.

Table 3 shows the overall results. For task A, the best indicators were **P4** (with `JRip`), **A1** (`NaiveBayes`) and **S1** (`SMO`). For task B the best one was **P4** (`KStar`). For task C, the best indicators were **P3** (`J48`), **A1** and **P3** (`JRip`), **C1** (`KStar`), **A1** (`NaiveBayes`) and **P2** (`SMO`). Each of the indicators **S2**, **P1** and **A2** either does not improve the results or does so but not as much as another, better indicator for the same task and algorithm.

It seems clear from Table 3 that some tasks benefit from these indicators more than others. In particular, task C shows consistent improvements whereas task B is hardly affected. Since task C is about relations involving two events, the classifiers may be picking up the sort of linguistic generalizations mentioned in Section 2 about *when* clauses.

`J48` and `JRip` produce human-readable models. We checked how these classifiers are taking advantage of the aspectual indicators. For task C, the induced models are generally associating high

values of the indicators **A1** and **P3** with overlap relations and low values of these indicators with other types of relations. This is expected. On the one end, high values for these indicators are associated with atelicity (i.e. the endpoint of the corresponding event is not presented). On the other hand, both indicators are based on queries containing the phrase *durante muito tempo* "for a long time", which, in addition to picking up events that can be modified by *for* adverbials, more specifically pick up events that happen *for a long time* and are thus likely to overlap other events.

For task A, `JRip` also associates high values of the indicator **P4**—which constitute evidence that the corresponding events are processes (which are atelic)—with overlap relations. This is a specially interesting result, considering that the queries on which this indicator is based reflect a purely aspectual constraint.

6 Concluding Remarks

In this paper, we evaluated the relevance of information about aspectual type for temporal processing tasks.

Temporal information processing has received substantial attention recently with the two TempEval challenges in 2007 and 2010. The most interesting problem of temporal information processing, that of temporal relation classification, is still affected by high error rates.

Even though a very substantial part of the semantics literature on tense and aspect focuses on aspectual type, solutions to the problem of automatic temporal relation classification have not incorporated this sort of semantic information. In part this is expected, as aspectual type is very interconnected with syntax (cf. the discussion about aspectual coercion in Section 2), and the phenomenon of aspect shift can make it hard to compute even when syntactic information is available.

Our contribution with this paper is to incorporate this sort of information in existing machine learned classifiers that tackle this problem. Even though these classifiers do not have access to syntactic information, aspectual type information seemed to be useful in improving the performance of these models. We hypothesize that combining aspectual type information with information about syntactic structure can further improve the problems of temporal information processing, but we leave that research to future work.

An interesting question that we hope will be addressed by future work is how these results extend to other languages. We cannot provide an answer to this question, as we do not have the data. However, this experiment can be replicated for any language that has (i) TimeML annotated data, (ii) a reasonable size of documents on the Web and a search engine capable of separating them from the documents in other languages and (iii) an aspectual system similar enough that the question being addressed in this paper makes sense (and useful patterns for queries can be constructed, even if not entirely identical to the ones that we used). The second criterion is met by many, many languages. The third one also seems to affect many languages, as the existing literature on aspectual phenomena indicates that these phenomena are quite widespread. The second criterion is, at the moment, the hardest to fulfill as not many languages have data with rich annotations about time (i.e. including events and temporal relations). We speculate that our results can extend to English, although a different set of query patterns may have to be used in order to extract the aspectual indicators that are employed. We believe this because the two languages largely overlap when it comes to aspectual phenomena.

References

Florbela Barreto, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes, and João Silva. 2006. Open resources and tools for the shallow processing of Portuguese: the TagShare project. In *Proceedings of LREC 2006*.

António Branco, Francisco Costa, Eduardo Ferreira, Pedro Martins, Filipe Nunes, João Silva, and Sara Silveira. 2009. LX-Center: a center of online linguistic services. In *Proceedings of the Demo Session, ACL-IJCNLP2009*, Singapore.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for fine-grained semantic verb relations. In *In Proceedings of EMNLP-2004*, Barcelona, Spain.

John G. Cleary and Leonard E. Trigg. 1995. K*: An instance-based learner using an entropic distance measure. In *12^{th} 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 *Proceedings of ACL 2010*.

Francisco Costa and António Branco. 2012. TimeBankPT: A TimeML annotated corpus of Portuguese. In *Proceedings of LREC2012*.

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

Henriëtte de Swart. 1998. Aspect shift and coercion. *Natural Language and Linguistic Theory*, 16:347–385.

Henriëtte de Swart. 2000. Tense, aspect and coercion in a cross-linguistic perspective. In *Proceedings of the Berkeley Formal Grammar conference*, Stanford. CSLI Publications.

David R. Dowty. 1979. *Word Meaning and Montague Grammar: the Semantics of Verbs and Times in Generative Semantics and Montague's PTQ*. Reidel, Dordrecht.

Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, , Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in KnowItAll. In *Proceedings of the 13th International Conference on World Wide Web*.

Eun Young Ha, Alok Baikadi, Carlyle Licata, and James C. Lester. 2010. NCSU: Modeling temporal relations with Markov logic and lexical ontology. In *Proceedings of SemEval 2010*.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, volume 2, pages 539–545, Nantes, France.

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.

George H. John and Pat Langley. 1995. Estimating continuous distributions in Bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo.

Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of ACL-08: HLT*, pages 1048–1056, Columbus, Ohio. Association for Computational Linguistics.

Laia Mayol, Gemma Boleda, and Toni Badia. 2005. Automatic acquisition of syntactic verb classes with basic resources. *Language Resources and Evaluation*, 39(4):295–312.

Congmin Min, Munirathnam Srikanth, and Abraham Fowler. 2007. LCC-TE: A hybrid approach to temporal relation identification in news text. pages 219–222.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

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*.

Georgiana Puşcaşu. 2007. WVALI: Temporal relation identification by syntactico-semantic analysis. In *Proceedings of SemEval-2007*, pages 484–487, Prague, Czech Republic. Association for Computational Linguistics.

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

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

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of ACL 2002*.

Graeme D. Ritchie. 1979. Temporal clauses in English. *Theoretical Linguistics*, 6:87–115.

Eric V. Siegel and Kathleen McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 24(4):595–627.

João Ricardo Silva. 2007. Shallow processing of Portuguese: From sentence chunking to nominal lemmatization. Master's thesis, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal.

Zeno Vendler. 1967. Verbs and times. *Linguistics in Philosophy*, pages 97–121.

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, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The TempEval challenge: identifying temporal relations in text. Language Resources and Evaluation.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of SemEval-2010*.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.