Annotation of WordNet Verbs with TimeML Event Classes

Georgiana Puşcaşu[§] **and Verginica Barbu Mititelu**[†]

[§]Research Group in Computational Linguistics
University of Wolverhampton, UK
georgie@wlv.ac.uk

[†]Research Institute for Artificial Intelligence
Romanian Academy, Romania
vergi@racai.ro

Abstract

This paper reports on the annotation of all English verbs included in WordNet 2.0 with TimeML event classes. Two annotators assign each verb present in WordNet the most relevant event class capturing most of that verb's meanings. At the end of the annotation process, inter-annotator agreement is measured using kappa statistics, yielding a kappa value of 0.87. The cases of disagreement between the two independent annotations are clarified by obtaining a third, and in some cases, a fourth opinion, and finally each of the 11,306 WordNet verbs is mapped to a unique event class. The resulted annotation is then employed to automatically assign the corresponding class to each occurrence of a finite or non-finite verb in a given text. The evaluation performed on TimeBank reveals an F-measure of 86.43% achieved for the identification of verbal events, and an accuracy of 85.25% in the task of classifying them into TimeML event classes.

1. Introduction

The ability to infer the temporal structure of a text is a crucial step towards its understanding. Access to the temporal information conveyed by natural language can lead to improvement in the performance of many NLP applications, such as Question Answering (QA), Automatic Summarisation, or any other application involving information about temporally located events.

In any framework that models time and what happens or is obtained in time, the following fundamental entities are essential: events/states, temporal expressions and temporal relations. Therefore, in order to automatically gain access to the temporal structure of text, there is a need for methods capable of identifying and classifying events, of identifying and normalising temporal expressions, and of detecting temporal relations that hold between events and other events or temporal expressions. Our previous work (Puscasu, 2004; Puscasu, 2007) dealt with the tasks involving temporal expressions and temporal relations, while the present work tackles the identification and classification of events.

The main motivation for this work is the current lack of methodology to automatically identify and classify events in any natural language text. Events are most of the time expressed using verbs and nouns (see section 2 for more details), therefore our first aim is to identify a method to classify verbs into a previously established set of event classes (described in section 2), and then to establish an algorithm for porting the annotation to nominalisations. In order to be able to classify verbs as belonging to a certain event class, an annotation process is carried out on all verbs present in WordNet 2.0 (Fellbaum, 1998), assigning to each verb its most relevant event class - that is the event class that covers most of that verb's meanings (thus, we cannot make use of the synsets in which meanings are organised in WordNet). We are aware that there are verbs which, given different contexts, belong to different event classes, but our assumption is that the number of such verbs is significantly lower than the number of verbs which, irrespective of their context, trigger the same event class. This approach is similar to the one often encountered in the Word Sense Disambiguation task, where the most frequent sense is assigned to every occurrence of a word. It is obvious that this method has its drawbacks, but it can be considered as a starting point in a more detailed future investigation of event classification.

This paper will therefore report on the annotation process that will result in each English verb being assigned an event class, as well as on the evaluation of a method that identifies and classifies verbal events by employing the resulted annotation. The evaluation is performed on TimeBank (Pustejovsky et al., 2003) - a collection of newswire articles annotated with this type of temporal information. The paper is structured as follows: Section 2 defines events, presents previous approaches for their identification and classification, and distinguishes the event classes to be employed in the annotation process. Section 3 describes in detail the annotation process, featuring at the same time a detailed discussion of the issues raised throughout this process. The two independent annotations are compared to reveal cases of disagreement between annotators, at the same time allowing inter-annotator agreement figures to be established. Section 4 captures the identification and classification of verbal events on the basis of the resulted annotation, and the corresponding evaluation results on the existing annotation included in TimeBank. In the last section conclusions are drawn and future directions of research considered.

2. Events and their Classification

This research relies upon the TimeML specification language (Sauri et al., 2005b), which has been adopted as the inter-lingua for temporal markup, and on the TimeBank corpus, the proof of concept for the TimeML specifications. TimeML considers *events* as a cover term for situations that happen or occur. The TimeML specifications also consider as events those predicates describing states or circumstances in which something obtains or holds true.

Events may be expressed by means of tensed or untensed verbs, nouns, adjectives, predicative clauses,

or prepositional phrases, but, for a simplification of the annotation process, TimeML has imposed certain rules in order to select the word or group of words to be annotated as events by applying the test of headedness. By looking only at the words annotated according to these rules, statistics extracted from TimeBank reveal that the annotated extent of an event is in 64.5% of the times a verb, in 28% of the cases a noun, 3.4% of events are adjectives, 0.3% prepositions, while 3.8% are assigned a part-of-speech category called OTHER (these events are, in most cases, numeric expressions or adverbs).

Since verbal events are most frequently encountered in text, we have chosen as a target for the present study the identification and classification of events expressed by means of verbs, also aiming in the future to transfer the methodology to nominalisations.

In trying to identify events, one existing approach is to consider all verbs events, with the exception of the verb "to be" and of several other forms of generics (Harabagiu and Bejan, 2006), while another approach, besides this restricted set of verbs, also considers as events certain nouns and the adjectives annotated as such in TimeBank (Sauri et al., 2005a). Other domain independent approaches consider as an event a text unit, at a coarser-grained scale the sentence (Hitzeman et al., 1995), and at a finer-grained scale the clause (Mani and Shiffman, 2004). The method employed so far by other researchers in classifying events into event classes is very preliminary, tagging events with the class that was most frequently assigned to them in TimeBank (Sauri et al., 2005a). This method obviously does not cater for verbs, nouns or adjectives not included in TimeBank and this is where our annotation and methodology brings a novel contribution by offering the research community a reliable method to identify and classify events in any natural language text.

Our idea is to annotate each English verb present in WordNet with an event class, annotation which aims not only to be useful to the research community in the assignment of an event class to a given verb, but also to be a starting point that can afterwards be refined by going at the verb sense level, or transferred to other languages using the WordNet ILI (Inter-Lingual Index) alignment.

Since our target is to obtain a tool capable of annotating any text with TimeML compliant temporal information, the event classes defined by TimeML were considered as the starting point in our investigation. These classes are:

- REPORTING: these events describe the action of a person or an organisation declaring, narrating or informing about an event, so their function is to associate the source of information with the reported event;
- PERCEPTION: this class includes events involving the physical perception of another event;
- ASPECTUAL: these events capture the aspectual predication on different facets of another event's history: initiation, reinitiation, termination, culmination, continuation;
- I_ACTION: an intensional action event introduces an event argument describing an action or situation from which we can infer something given its relation with the I_ACTION event;
- I_STATE: this class contains states that refer to alternative or possible worlds;
- STATE: states are circumstances in which something holds true;
- OCCURRENCE: an occurrence event is defined as something that happens or occurs in the world.

An analysis of these event classes and of the annotated corpus reveals many annotation inconsistencies, in the sense that the same verb in very similar contexts is annotated with different classes (for example the verb *launch* in the context *launch the offer* is in one case annotated with the class OCCURRENCE, and in the other case with I_ACTION). Even the official inter-annotator agreement figures for TimeBank reveal many inconsistencies, the inter-annotator agreement for the event class being 0.77. This figure also illustrates the fact that event annotation is not a trivial task, not even for humans. The classes OCCURRENCE and I_ACTION both include situations that happen, occur or involve change, the only difference between them being the fact that the I_ACTION event has an event argument, while the OCCURRENCE event does not. The same applies to the classes STATE and I_STATE. In order to make the human annotation process easier, and the targeted automatic annotation process doable for an automatic tool, we decided to reduce the set of event classes by merging the OCCURRENCE and I_ACTION classes into only one class (OCCURRENCE), and by also merging the STATE and I_STATE classes into one class (STATE), thus obtaining the following simplified set of event classes:

- REPORTING
- PERCEPTION
- ASPECTUAL
- OCCURRENCE
- STATE

Even if there are reasons to differentiate the OCCURRENCE and I_ACTION, as well as the STATE and I_STATE events pragmatically, we will place higher relevance on the resemblances which bring these classes together, and will neglect the differences.

Each of the five classes in the reduced set has different temporal properties. For example, a REPORTING event most commonly happens after the reported event, while perceived events happen roughly at the same time as the PERCEPTION events. The temporal consequence of ASPECTUAL events is that they indicate different stages of their argument event (beginning, end, continuation). OCCURRENCE events cover situations that involve change, processes consisting of different stages, or situations that have duration and involve an end result. STATE events cover situations that do not involve change over time. In the case of two consecutive events, typically an OCCURRENCE takes place just after a preceding OCCURRENCE, while a STATE overlaps a preceding OCCURRENCE.

3. Annotation Methodology

The annotation process takes place in two stages, at the first stage each verb is assigned one WordNet lexicographic file, while at the second stage each verb in turn is assigned one event class by two independent annotators.

3.1. Mapping verbs to WordNet Lexicographic Files

WordNet verb senses are grouped into 15 lexicographic files:

- verb.body
- verb.change
- verb.cognition
- verb.communication
- verb.competition
- verb.consumption
- verb.contact
- verb.creation
- verb.emotion
- verb.motion
- verb.perception
- verb.possession
- verb.social
- verb.stative
- verb.weather

Since one verb can have more senses, there are cases when not all verb senses are in the same lexicographic file. In fact, from a total number of 11,306 verbs in WordNet 2.0, only 7,437 verbs have all their senses in the same lexicographic file, for the remaining 3,869 verbs the senses are scattered among several lexicographic files. The first stage in the annotation process is assigning to each English verb only one lexicographic file. The assigned file is the one that maximises the score:

$$\texttt{score(file}_i\texttt{)} = \sum \texttt{(1/j)}$$

for each `j` ranging from 1 to the number of senses of the analysed verb.

In the above formula, `j` is the sense number and $\texttt{file}_i$ is the corresponding lexicographic file assigned to sense `j`. This formula chooses the lexicographic file that covers most of the important senses of a verb (as one can notice, a higher sense number corresponding to a more frequent sense gives a higher score to its lexicographic file).

3.2. Annotation Process

3.2.1. Annotation

After each verb is assigned one lexicographic file, two annotators go through each lexicographic file and assign to each verb one of the five event classes described above (REPORTING, PERCEPTION, ASPECTUAL, OCCURRENCE, STATE). Both annotators approach the annotation from two different perspectives and employing different resources.

One annotator looks only at those WordNet senses that motivated the verb's inclusion in the assigned lexicographic file and identifies the event class that offers the highest coverage of those senses.

The other annotator looks up each verb in the Oxford English Dictionary, eliminates all obsolete and rare senses and assigns the class that, according to the annotator's intuition, covers best the remaining senses.

One could argue that annotating verbs for their event type outside a context is not a proper way of doing it, as words do not have meaning in isolation, but only in the context of a sentence. However, we consider that the core meaning of a word can be captured in a lexicographic definition, and the context only favours refinements of that meaning (with some semantic traits being blocked or, on the contrary, favoured to manifest in a certain words combination).

At the end of the annotation process, the cases of agreement/disagreement were carefully analysed. This analysis revealed that, out of 11,306 verbs, the same class was assigned by both annotators in 10,945 cases, meaning an absolute agreement of 96.80%. By investigating the cases of disagreement, we discovered certain issues that were not clearly specified in the annotation guidelines. We therefore decided to clarify the guidelines and to revise the annotation accordingly.

3.2.2. Revision of Guidelines and Annotation

The cases of disagreement revealed annotation errors due to issues in the guidelines that required further clarification.

One issue refers to events that were wrongly annotated as REPORTING. Certain communicative verbs were classified as REPORTING, even if they do not have the ability to report about other events they would take as arguments (in case they could have arguments). Here are some examples of verbs wrongly annotated as REPORTING: *counsel*, *talk*, *compliment*. These verbs cannot occur with arguments denoting events they talk about. One should also be aware that the annotator's choice was influenced by the verb semantics filtered through that person's idiolect and life experience. In the case of the verb *disagree* for example, it is well known that disagreement is most frequently expressed verbally, so, as a result, this verb was initially categorised by one annotator as REPORTING. The same misinterpretation was to blame for some verbs being initially annotated as REPORTING, and only on a second thought as OCCURRENCE: *decree*, *swear*, *badmouth*, etc.

Similarly, some verbs were wrongly annotated with the class PERCEPTION, when they lacked the ability to describe the physical perception of another event, even if they referred to physical perception. Some examples are: *suffer*, *hurt*, *itch*, *miss*.

Another issue was that, in order to annotate a verb as ASPECTUAL, that verb should, in its most frequent usages, take another event as argument, whose aspectual facets it should refer to. Since this was not clearly expressed in the annotation guidelines, verbs like *break_out* or *abrogate* were wrongly annotated as ASPECTUAL, even if both *break_out* and *abrogate*, with their most frequent senses, neither take other events as arguments, nor do they refer to a certain stage in an event's evolution.

The most important problem we observed by analysing the disagreement cases was that the boundary between what we defined as STATEs and what we defined as

OCCURRENCEs was not clear-cut. In many such cases the verbs involved express inner or physiologic processes, whom one of the annotators initially considered STATEs, and the other OCCURRENCEs: *didder*, *retrospect*, *gestate*. After discussing all the above mentioned issues and clarifying the guidelines, both annotators independently adjusted their annotations accordingly for the verbs they did not agree upon, each annotator reconsidering the class they would assign to those verbs, without knowing the other annotator's decision. Finally, inter-annotator agreement was measured on the resulted annotations. Out of 11,306 verbs, the two annotators agreed on the same class being assigned to 11,087 verbs, yielding an absolute agreement of 98.06%. Cohen's kappa statistic(Cohen, 1960), taking also into consideration the proportion of chance agreement, reveals a kappa score of 0.87.

3.2.3. Final Decision

The remaining cases of disagreement (accounting for 219 verbs) were then submitted to a third annotator, in order to select only one class per verb from the two distinct annotations. A voting scheme was then applied to the three annotations, and a given verb was assigned the class two out of three annotators agreed on. Still, there were 16 verbs for which the three annotators chose three different classes. For example, in the case of the verb *give_out*, one annotator chose the class REPORTING (as it has the meaning *to announce; proclaim; report*, see [1]), another annotator chose the class STATE (as it has the meaning *to emit*, see [2]), and the third annotator chose the class OCCURRENCE (as it has the meaning *to break down, get out of order, fail*, see [3]).

[1] *He **gave out** at Macao, that he was bound to Batavia.*

[2] *The gold **gave out** its red glow.*

[3] *The Ruby's engines **gave out** for a time.*

The final classes assigned to these 16 verbs were decided by a fourth annotation.

4. Evaluating our Annotation against TimeBank

Since our initial goal was to obtain a tool capable of identifying and classifying events in natural language texts, the functionality of this tool was broken down into two parts: one which deals with the identification of events, and a second one which fills in the values of the attributes that characterise an event according to the TimeML specification (Sauri et al., 2005b). The TimeML annotation guidelines define the following attributes (and possible values) for an event:

- **eventID**: unique identification number automatically assigned to each event instance found in text;
- **class**: each event belongs to one of the following classes: REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, OCCURRENCE, I_STATE, STATE (please refer to section 2. for a detailed description of these values);
- **tense**: refers to the grammatical category of tense. This attribute can have the values: PRESENT, PAST, FUTURE, INFINITIVE, PRESPART, PASTPART, or NONE;
- **aspect**: captures the grammatical category of verbal aspect. The possible values for this attribute are: PROGRESSIVE, PERFECTIVE, PERFECTIVE_PROGRESSIVE or NONE;
- **pos**: represents the part of speech corresponding to an event. Its values can be: ADJECTIVE, NOUN, VERB, PREPOSITION, or OTHER;
- **polarity**: reveals whether the event has happened or not. The possible values for this attribute are: NEG and POS;
- **modality**: captures the modal information attached to an event (*may*, *can*, *could*, *would*, *should*, *might*).

As this paper is concerned with the annotation of verbs with event classes, it will only cover the identification of both finite and non-finite verbal events, followed by their classification according to the classes used in our annotation (i.e. filling in the value of the TimeML attribute **class** for each identified verbal event). Due to the fact that TimeBank is the reference corpus annotated according to TimeML, all our evaluations are performed on TimeBank. Since our tool is designed to work on any natural language text, the TimeBank articles are first converted to plain text by eliminating all XML tags, and then processed using Conexor's FDG Parser (Tapanainen and Jaervinen, 1997). This parser returns information on a word's part of speech, morphological lemma and it's functional dependencies on surrounding words. This information is useful for the identification of both finite and non-finite verbal events. In the following, we present individual results obtained for the identification and classification of finite and non-finite verbal events.

4.1. Finite Verb Events

4.1.1. Identification

The information provided by Conexor's FDG parser is employed in order to detect the full extent of the finite verb phrases that appear in a text. The head of each identified verb phrase, which is usually the last word in the group, is then marked as an event, except the case when the head is any form of the verb *to be*. This exclusion is due to the TimeML guidelines which clearly specify that any occurrence of the verb *to be* as finite main verb are not to be labeled as events. Therefore, all finite main verbs except the verb *to be* are considered events.

In order to compare the performance of this purely syntactic finite verb event identifier against TimeBank, we extracted only the finite verb events annotated in TimeBank, as being those events for which the attribute **pos** has the value VERB, and the attribute **tense** has any of the values PAST, PRESENT, FUTURE or NONE. Even if we have noticed non-finite verbs in infinitive that were annotated with the class NONE for the attribute **tense**, when this attribute should have received the value INFINITIVE, we considered this to be an error in the TimeBank annotation

and we did not change the way we chose finite verb events from TimeBank.
When comparing the finite verb events we identify with the ones annotated in TimeBank, the following figures are revealed:

- there are 3,845 finite verb events annotated in TimeBank.
- we identify 4,466 finite verb events in all TimeBank articles.
- in 3,602 cases the finite verbs identified by us coincide with those annotated as finite verb events in TimeBank. This leads to a precision of 80.65%, a recall of 93.68%, and an overall f-measure of 86.68% in identifying finite verb events. The low precision obtained in identifying finite verb events has as its main explanation the fact that we do no not attempt to identify verbs with generic usages or verbs present in headlines in order to avoid their annotation.
- in 3,738 cases the finite verbs identified by us are annotated as events in TimeBank. Out of these, 3602 are annotated as finite verb events, 68 as non-finite verb events, 35 have the part of speech set on NOUN, 29 on ADJECTIVE, and 4 on OTHER. A close look at those finite verb events that appear annotated in TimeBank as either non-finite verbs, nouns or adjectives revealed 46 cases wrongly annotated in TimeBank (18 finite verbs wrongly annotated as non-finite, 15 wrongly annotated as nouns, and 13 wrongly annotated as adjectives), as well as 86 parser errors.

Afterwards, we eliminate the verbs annotated as events in TimeBank from the set of finite verb events we identify (4,466 - 3,738 => 728), and check how many should have been annotated in TimeBank. The answer is that 284 occurrences of verbs should have been annotated in TimeBank and were not. The remaining 444 finite verb events identified, that should not have been annotated in TimeBank, are due to different reasons.
There are 318 cases that should not receive an annotation according to certain specifications in the guidelines. Generic usages of verbs are not supposed to be annotated, and, since we do not attempt to identify generic finite verbs, we annotate them in 140 cases (e.g. [4]). Events occurring in the headlines of the articles should not receive an annotation (there are 88 cases of finite verb events we identify in headlines, e.g. [5]). Modal verbs and auxiliary verbs not followed by a main verb are also excluded from annotation, and we annotate such verbs in 83 cases (e.g. [6] and [7], respectively). There are also finite verbs appearing in fixed phrased that do not contribute to the meaning of the sentences and they should not be annotated (we annotate 7 such finite verbs, e.g. [8]).

[4] *Ethnic Albanians **comprise** 90 percent of the population in Kosovo, but Serbs maintain control through a large military and police presence.*

[5] *Saddam **Seeks** End To War With Iran.*

[6] *We will continue to do everything we **can** to establish what has happened.*

[7] *Service industries also showed solid job gains, as **did** manufacturers, two areas expected to be hardest hit when the effects of the Asian crisis hit the American economy.*

[8] *You **know**, since he's been here the stock skyrocketed so, Yeah I think he's doing the right thing.*

There are 126 errors of identification produced by the syntactic parser. These comprise all those cases in which nouns (e.g. [9]), adjectives (e.g. [10]), adverbs (e.g. [11]), prepositions (e.g. [12]) or conjunctions (e.g. [13]) were annotated as finite verbs, and also cases of ungrammatical sentences (e.g. [14]), and non-finite verbs (e.g. [15]) that are tagged as finite ones.

[9] *The Pentagon said that Defense Secretary Dick Cheney is considering urging Bush to order a national callup of armed forces **reserves** for active duty because of the drain on units sending soldiers abroad.*

[10] *Last year, Russian officials assailed Ukraine for holding **joint** naval exercises with NATO in the Black Sea an area Moscow considers its own turf.*

[11] ***Live** from Atalanta, good evening Lynne Russell, CNN headline news.*

[12] *His advisers said the results reflected not just from balancing the budget, but also initiatives **like** improved access to education and training and the opening of foreign markets to trade.*

[13] *Prime Minister Benjamin Netanyahu told his Cabinet on Sunday that Israel was willing to withdraw from southern Lebanon **provided** Israel's northern frontier could be secured.*

[14] *In Hong Kong, is always **belongs** to the seller's market.*

[15] *In a long verbal attack **read** on Iraqi television Thursday, Saddam repeatedly called Bush "a liar" and said a shooting war could produce body bags courtesy of Baghdad.*

4.1.2. Classification

The aim of annotating the WordNet verbs with TimeML classes was to automatically apply it to any natural language text and assign a class to each identified verbal event. Therefore, in order to evaluate and demonstrate the usefulness of our annotation, we apply it to the finite verb events we correctly identify in the TimeBank articles (in terms of text span and according to the existing TimeBank annotation), i.e. 3,602 cases (see 4.1.1.). This is done by looking up the class assigned to each finite verb identified and comparing it against the one annotated in TimeBank.
From the 3,602 finite verb occurrences, 3,526 are found in WordNet and therefore a corresponding class exists in our annotation. The remaining 76 do not appear in WordNet (73 are phrasal verbs, like *succeed in*, one is a parser error - *placed*, one is a verb whose absence from WordNet is quite strange - *pend*, and the last one is *nose-dive* which appears in WordNet as *nosedive*). In the case of phrasal verbs, we automatically assign them the class corresponding to the original verb obtained by deleting the particle, even if we are aware that the meaning, and consequently the attached class, may be different.
When comparing the class that is assigned in our annotation to a certain verb, with the class annotated in TimeBank, we correctly classify 3,053 cases out of 3,602 (i.e. 84.75%).

One baseline we could consider would be assigning to all finite verb events the most frequent class encountered in TimeBank (i.e. OCCURRENCE), this resulting in 1,982 correctly classified cases, yielding an accuracy of 55.02%.
Apart from this baseline, a classifier is trained by ten fold cross validation on TimeBank to assign to each verb the most frequent class assigned to it by manual annotation in TimeBank, this resulting in 3,116 verb occurrences being correctly classified. This yields an accuracy of 86.50%, only 1.75% higher than the precision and recall obtained by applying our annotation. Therefore, we can conclude that our annotation applied to the TimeBank articles provides the correct class for a number of cases that is very close to the maximum number of cases that can be correctly identified by adhering to the "one class per verb" paradigm.
In section 4.1.1. we discussed the fact that we identified 284 finite verb occurrences that should have been annotated in TimeBank and were not. These cases were annotated manually with the corresponding event classes, and then we automatically checked how many of these cases we would have correctly assigned a class to by using our annotation, and the result was 245 (i.e. 86.26%).
If, instead of looking at each finite verb occurrence in TimeBank, we examine individual verbs (lemmas), we notice that there are 769 unique finite verbs appearing in TimeBank. In 649 (i.e. 84.39%) of the cases the class we assigned to a particular verb is equal to the most frequent class assigned to it in TimeBank.
The finite verbs having the class assigned by our annotation different than the most frequent class encountered in TimeBank, accounting for 120 cases, were analysed in detail, in order to identify what caused this disagreement.
In most of the cases, the verb senses used in TimeBank are different to the most frequent senses a verb is normally used with. For example, the verb *abandon* appears twice in TimeBank (e.g. [16]), and both times it is annotated as ASPECTUAL. But its usage with the sense of putting an end to an event is encountered more seldom than the senses of leaving behind, of emptying, of deserting. This verb has received the class OCCURRENCE in our annotation, but its most frequent class found in TimeBank is ASPECTUAL.

[16] *However, StatesWest isn't* ***abandoning*** *its pursuit of the much-larger Mesa.*
(ASPECTUAL in TimeBank)

Also, there are 31 verbs for which the most frequent class assigned in TimeBank should have been the one we annotated. This is due to errors of annotation in TimeBank. One example would be the verb *split*, which appears once in TimeBank annotated as ASPECTUAL (see [17]), while in our annotation it is assigned the class OCCURRENCE. Another example would be the verb *state*, which appears twice as finite verb in TimeBank and is once annotated as OCCURRENCE (see [18]), and once as REPORTING (see [19]), the most frequent class selected being OCCURRENCE. In our annotation the verb *state* is annotated as REPORTING.

[17] *No successor was named, and Mr. Reupke's duties will be* ***split*** *among three other senior Reuters executives, the company said.*
(ASPECTUAL in TimeBank)

[18] *I was pleased that Ms. Currie's lawyers* ***stated*** *unambiguously this morning... that she's not aware of any unethical conduct.*
(OCCURRENCE in TimeBank)

[19] *Organizers* ***state*** *the two days of music, dancing, and speeches is expected to draw some two million people.*
(REPORTING in TimeBank)

We have also encountered errors in our annotation by checking these cases of disagreement with the most frequent class annotated in TimeBank. There are 6 verbs for which we concluded that we have assigned the wrong class. One example would be the verb *plan*, which we have seen as an on-going process of devising a plan, and therefore we assigned it the class OCCURRENCE. In TimeBank it appears 17 times denoting STATEs (e.g. [20]), being probably understood with the sense of having a certain intention.

[20] *Kuchma also* ***planned*** *to visit Russian gas giant Gazprom , most likely to discuss Ukraine's dlrs 1.2 billion debt to the company.*
(I_STATE in TimeBank)

Even if there are cases in which our annotation fails to provide the most appropriate class for a certain verb occurrence, the results obtained so far prove that our methodology and verb annotation can be useful not only in detecting the event classes for already annotated TimeBank events, but also in detecting and classifying new events missed by the TimeBank annotators.

4.2. Non-finite Verb Events

4.2.1. Identification

Based on the output of Conexor's FDG parser, the full extent of all non-finite verb phrases is identified. As in the case of finite verbs, only the head of each non-finite verb phrase is automatically annotated as an event. The only exception to this process is any non-finite form of the verb *to be*.
The non-finite events annotated in the gold standard corpus (TimeBank) are extracted by selecting only those events for which the attribute **pos** has the value VERB, and the attribute **tense** ranges over the values INFINITIVE, PRESPART and PASTPART.
A comparison of the non-finite verbal events annotated in TimeBank with the ones automatically identified by the parser revealed the following:

- there are 1,274 non-finite verb events annotated in TimeBank.
- we identify 1,819 non-finite verb events in all TimeBank articles.
- in 1,136 of the cases the non-finite verb events we identify are also annotated in TimeBank. This leads to a precision of 62.45%, a recall of 89.16%, and an overall f-measure of 73.45% in identifying non-finite verb events.

- in 1,356 cases the non-finite verbs identified by our parser are annotated as events in TimeBank. Out of these, 1,136 are annotated as non-finite verb events, 123 as finite verb events, 84 have the part of speech set on NOUN, 12 on ADJECTIVE, and 1 on OTHER. A careful examination of those non-finite verb events that appear annotated in TimeBank as either finite verbs, nouns or adjectives reveals 125 cases wrongly annotated in TimeBank (70 non-finite verbs wrongly annotated as finite, 48 wrongly annotated as nouns, and 7 wrongly annotated as adjectives), as well as 94 parser errors.

An analysis of the non-finite verbs identified by our parser, but not annotated as events in TimeBank (463 cases) reveals the fact that 252 of them should have been annotated.
For the remaining 211 cases, their presence in our list of non-finite verbs not labeled as events is fully justified, as they were not supposed to be annotated according to certain specifications in the guidelines.
As in the case of the finite verbs, generic usages should not be annotated, but since we do not attempt to identify generic verbs, we annotate them in 64 cases (e.g. [21]). Among these, 19 instances account for certain verbs that form generic expressions used to elaborate in more detail on something previously mentioned (e.g. *related to* [22]). There are also 15 cases we consider generic because the non-finite verbs are employed in noun phrases to qualify certain characteristics of the noun they syntactically depend on (e.g. *civil rights* ***monitoring*** *group*, ***detonating*** *cord*).

[21] *So for Hong Kong, it's time, as investment bankers like to* ***say****, to reposition.*

[22] *In addition, Hadson said it will write off about $3.5 million in costs* ***related*** *to international exploration leases where exploration efforts have been unsuccessful.*

Apart from generic usages, we also encountered 64 non-finite verbs occurring in article headlines, which should not receive an annotation (e.g. [23]). Modal and auxiliary verbs, also excluded from annotation, were identified 4 times (e.g. [24]). Three non-finite verbs appear in fixed phrases that do not contribute to the sentence meaning and they should not be annotated (e.g. [25]).

[23] *Qantas to* ***run*** *daily flights between Australia and India*

[24] *"Those fumes will exhaust themselves, and the manufacturing sector is going to start* ***getting*** *beat up in the spring."*

[25] *He added, "This has nothing to* ***do*** *with Marty Ackerman and it is not designed, particularly, to take the company private."*

There are 76 errors of identification produced by the syntactic parser. These include the cases in which nouns (e.g. [26]), finite verbs (e.g. [27]), but mostly prepositions (e.g. [28]) are annotated as non-finite verbs.

[26] *And nails found in the Atlanta abortion clinic bombing are identical to those discovered at Rudolph's storage* ***shed*** *in north Carolina.*

[27] *Geraldine Brooks in Amman, Jordan, and Craig Forman in Cairo, Egypt,* ***contributed*** *to this article.*

[28] *Ranariddh's loyalists,* ***including*** *Nhek Bunchhay, his top military commander, went into hiding or fled the capital.*

It is a well known fact that annotating events is by far a very difficult and tedious task, even for human annotators. It is normal for annotators either to annotate extra events that should not have been annotated, or to miss out events that they probably did not consider relevant or that they simply did not notice because of being tired or bored. Therefore, we assume it is only normal to find events that should have been annotated and were not, even if there were a human annotator and not an automatic tool performing the annotation. One thing we have noticed is that the percentage of finite verbs that should have been annotated (284 out of 728 analysed cases => 39.01%) is much lower than the percentage of non-finite verbs that should have been considered events (252 out of 463 analysed cases => 54.42%). This reflects the assumption that finite verbs capture the most important information in a sentence, and therefore the information expressed by non-finite verbs is more often not considered relevant for annotation purposes.

4.2.2. Classification

At this stage our annotation is applied to the non-finite verbal events identified by our parser that are also annotated in TimeBank (1,136 occurrences). The class assigned to each verb in our annotation matches the class annotated in TimeBank in 991 cases, thus the precision and recall we obtain in assigning the correct class to non-finite verbal events is 87.23%. Only two verbs do not appear in WordNet (*dole* and *downsize*).
A baseline scenario could correspond to all non-finite verbal events receiving the most frequent class annotated in the corpus (i.e. OCCURRENCE), this being successful in 966 cases (i.e. 85.03%).
By applying ten fold cross validation on TimeBank (i.e. splitting all occurrences of non-finite verbal events into 10 files, then choosing for each verb its most frequent class annotated in nine files, and finally assigning the chosen classes to all verbs in the remaining file) 995 instances are annotated correctly, yielding an accuracy of 87.58%.
Our annotation is also applied to those 252 instances of non-finite verbs that should have received an annotation in TimeBank (please refer to section 4.2.1. for more details). The result of this automatic classification process is compared against the manually annotated data, revealing 213 non-finite verb instances correctly classified (84.52%).
By examining, instead of verb occurrences, individual verbs (lemmas), 470 unique non-finite verbs are found annotated as events in TimeBank. In 416 cases the class we assign to a verb coincides to the one most frequently annotated in the corpus, therefore our decision on an event class coincides with the class revealed by existing annotated data for 88.51% of the verbs.
For 54 verbs, the most frequent class annotated in TimeBank is different to the one we considered appropriate. In most cases, it is just a matter of a particular sense or usage that appears more frequently in the TimeBank articles. For example, the verb *include* appears only once in TimeBank (see [29]), that instance being annotated as OCCURRENCE, as the verb is used in the sense of adding

as part of something else or putting in as part of a set, group, or category (third sense in WordNet). Still, the verb *include* is assigned the class STATE in our annotation, as it is more frequently used with the sense of having as a part or being made up out of (first sense in WordNet, see [30]).

[29] *The Internet, the global network of computers, is now far reaching into the country - extending its embrace to* ***include*** *every nook and cranny of the nation.*

[30] *The list* ***includes*** *the names of many famous writers.*

There are also a number of cases corresponding to errors in TimeBank, where the class should have been the one present in our annotation. One example would be the verb *quit* appearing once as a non-finite verb and wrongly annotated in TimeBank as ASPECTUAL (see [31]), when the class should have been OCCURRENCE.

[31] *If the government succeeds in seizing Mr. Antar's assets, he could be left without top-flight legal representation, because his attorneys are likely to* ***quit****, according to individuals familiar with the case.*
(ASPECTUAL in TimeBank)

In certain cases there are errors in our annotation - the class most frequently annotated in TimeBank being more suitable to characterise a verb than the one present in our annotation. One example is the verb *lead*, which we consider a stative verb, but it is probably used more frequently as an OCCURRENCE.

5. Conclusions

This paper covers efforts towards the development of a methodology to automatically identify and classify events in any natural language text.

It mainly addresses the process of annotation, targeting to assign to each WordNet verb one TimeML event class. Each WordNet verb was assigned an event class by two independent annotators who chose, according to their intuition, the class that best covered most of that verb's important senses. The inter-annotator agreement was in terms of absolute agreement 96.80%, and in terms of kappa statistics 0.87. The cases of disagreement were clarified with a third, and in some cases, a fourth annotation, and finally each verb mapped to exactly one event class.

An automatic method employing the resulted language resource was then evaluated on TimeBank to measure its performance in identifying and classifying events expressed through verbs. The evaluation was done separately for finite and non-finite verbs.

At the level of event identification, we considered as events all finite and non-finite verbs, except any form of the verb *to be*. The result of the current identification method across all verbs, without making the finite/non-finite distinction, was then compared with all verbal events annotated in TimeBank, i.e. those events having the **pos** attribute set on VERB. This comparison revealed a precision of 78.42%, a recall of 96.28%, and an F-measure of 86.43% for the identification of events expressed using verbs. The relatively low precision is due to over-annotation, therefore, in the future, we aim to refine this method to be able to identify generic verb usages, verbs in headlines and modals/auxiliaries not followed by a main verb, in order to avoid their annotation.

The correctly identified verbal events are then automatically assigned the corresponding class found in our annotation. The evaluation against the event classes manually annotated in TimeBank reveals an accuracy of 85.25% in the task of classifying verbal events into TimeML event classes. A baseline system that always assigns the class OCCURRENCE to each identified event would have an accuracy of 62.54%.

Unlike most previous work on event classification, the present effort provides a reliable way to classify every verbal event into one TimeML class, irrespective of the presence or absence of that verb in TimeBank. Our approach intended to be as domain independent as possible and catering for most of the verbs in the English language, WordNet offering us an almost complete coverage. In terms of unique verbs, TimeBank can provide the most frequent class for 926 verbs, while our annotation also caters for 10,380 extra verbs.

In the future, we will try to identify an automatic method to port this verb annotation to nominalisations and adjectivisations that might also express events.

6. References

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. In *Education and Psychological Measurement*.

C. Fellbaum, editor. 1998. *WordNet: An Eletronic Lexical Database*. The MIT Press.

S. Harabagiu and C. Bejan. 2006. An Answer Bank for Temporal Inference. In *Proceedings of LREC 2006*.

J. Hitzeman, M. Moens, and C. Grover. 1995. Algorithms for Analyzing the Temporal Structure of Discourse. In *Proceedings of the Annual Meeting of the European Chapter of the Association of Computational Linguistics (EACL'95)*.

I. Mani and B. Shiffman. 2004. Temporally Anchoring and Ordering Events in News. In J. Pustejovsky and R. Gaizauskas, editors, *Time and Event Recognition in Natural Language*. John Benjamins.

G. Puscasu. 2004. A Framework for Temporal Resolution. In *Proceedings of the LREC 2004*.

G. Puscasu. 2007. Discovering Temporal Relations with TICTAC. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizuaskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*.

R. Sauri, R. Knippen, M. Verhagen, and J. Pustejovsky. 2005a. Evita: A Robust Event Recognizer For QA Systems. In *Proceedings of HLT/EMNLP 2005*.

R. Sauri, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. 2005b. TimeML Annotation Guidelines Version 1.2.1. http://www.timeml.org.

P. Tapanainen and T. Jaervinen. 1997. A non–projective dependency parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing, ACL.*