# Building a Gold Standard for Event Detection in Croatian

**Nikola Ljubešić, Tomislava Lauc, Damir Boras**

Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
{nljubesi,tlauc,dboras}@ffzg.hr

**Abstract**

This paper describes the process of building a newspaper corpus annotated with events described in specific documents. The main difference to the corpora built as part of the TDT initiative is that documents are not annotated by topics, but by specific events they describe. Additionally, documents are gathered from sixteen sources and all documents in the corpus are annotated with the corresponding event. The annotation process consists of a browsing and a searching step. Experiments are performed with a threshold that could be used in the browsing step yielding the result of having to browse through only 1% of document pairs for a 2% loss of relevant document pairs. A statistical analysis of the annotated corpus is undertaken showing that most events are described by few documents while just some events are reported by many documents. The inter-annotator agreement measures show high agreement concerning grouping documents into event clusters, but show a much lower agreement concerning the number of events the documents are organized into. An initial experiment is described giving a baseline for further research on this corpus.

## 1. Introduction

Event detection is a task of identifying all information that describes a specific event. An event is considered as something that happens at some specific time and place (Papka, 1999), or, more general, addition or loss of property (Lombard, 1986).

The most prominent initiative in this field is the TDT (Topic Detection and Tracking) initiative. It started as a 1997 pilot study (Allan et al., 1998) and developed five annotated corpora until 2004 (TDT, 2004). It is considered to be a standard in the field.

The basic notion in the TDT initiative is the topic. It is most often defined as a seminal event with all events directly related to it (TDT, 2004). The TDT initiative does not take specific events into consideration, but considers them just as parts of a specific topic. Therefore, specific events are not annotated in the corpora.

During its years different types of topics were annotated, from topics with major seminal events (such as Pinochet trial in TDT2), over very broad topics (such as the Monica Lewinski case in TDT3) to a more balanced topic selection process consisting also of singleton topics in TDT4 (TDT, 2004).

The annotation process consisted of topic selection where topics to be annotated were chosen, and the annotation process where the documents reporting about chosen topics were annotated. The collection of documents was from a specific source for a specific time span. In this large collection of documents only some documents were labeled with a topic. Additionally, it was possible for some documents to be labeled with more than one topic.

The TDT annotation method also evolved from simpler to a four-stage process in TDT4 (TDT, 2004).

## 2. Goal of the paper

The goal of this paper is to describe the process of building a corpus where each document would be annotated by the event it describes. This approach differs from the TDT approach considerably since in TDT topics are annotated, and not events.

The annotation process, consisting of a browsing and a searching step, is described. A statistical analysis of the annotated corpus is given by some distributional properties of the sample. Furthermore, two separate annotations are compared in a qualitative and quantitative manner.

The impact of using a predefined similarity threshold in the annotation process on recall, i.e. the completeness of the annotation process is also explored.

In the last section an initial event-detection experiment is described. This experiment should serve as a baseline for further research.

## 3. Building the corpus

The corpus consists of 2.398 documents published on sixteen news portals in a three days time span. The fact that the corpus consists of documents from sixteen different sources is a specificity of this corpus concerning the TDT corpora. Namely, the TDT corpora mostly consist of a single or two sources for a specific language and type. It is our belief that such sample depicts the everyday online situation more clearly, where a typical user is overwhelmed with the number of sources that offer similar content.

Another specificity of this corpus is the fact that the documents are written in Croatian. To our knowledge, this is the first event-annotated corpus for any Slavic language.

### 3.1. Annotation of the corpus

Our approach to the annotation process differs considerably from the TDT annotation approach in two ways:

1. documents are annotated with event identifiers, not topic identifiers
2. all documents in a time span are annotated

In our opinion, annotating specific events (which can in a later phase be connected into topics) is a much more informative annotation method enabling much broader experiments. Furthermore, by annotating all documents in a document collection, the problem of event selection is omitted. Since every document is annotated, such a corpus enables a clear insight into the relationship of events in a document sample.
In the annotation process, every document is annotated with just one event. Defining an event list that could lead the annotators in the annotation process was not possible because of the overwhelming number of implicit events. Therefore the list of events was built up indirectly during the annotation process.
The corpus is annotated with a software developed for this purpose (Bakarić, 2009). Since the number of documents pairs is overwhelming (n*(n-1)/2, i.e. 2.874.003), for every document the software offers a list of documents sorted by their similarity to the main document. The similarity is calculated as the harmonic mean of following normalized similarity measures:

- cosine similarity
- Jaccard coefficient as defined in (Grefenstette, 1994)
- Dice coefficient as defined in (Curran, 2004)
- Jensen-Shannon divergence as defined in (Curran, 2004)

Thereby, the bias towards a specific similarity measure is avoided. The decision on when to quit browsing the document list is left to the annotator and his intuition that no additional documents will be found easily by browsing the list further.
After the browsing step, a search step is undertaken for all documents that were previously not browsed.
The corpus is annotated twice by separate annotators given same instructions.

### 3.2. Distributional properties of events

The 2.398 documents are annotated with 1.214 different event identifiers, i.e. in the sample, 1.214 different events are reported on.
The basic distributional property of the sample, concerning described events, is that some events are described by many documents while most events are described by just a few or a single document. Out of 1.214 event clusters, 817 of them, i.e. 67%, contain just a single document. The largest cluster contains 51 documents. The distribution of event clusters concerning the number of documents they contain is given in figure 1.
The distribution shows that more than 30% of documents describe a singleton event. These singleton events present 67% of all events. Event clusters up to the size of three documents contain 61% of documents. These documents describe 88% of the events.
Since the document sample is annotated by two independent annotators given the same instructions, an inter-annotator agreement can be measured. Two different inter-annotator agreements are calculated. The first one is the

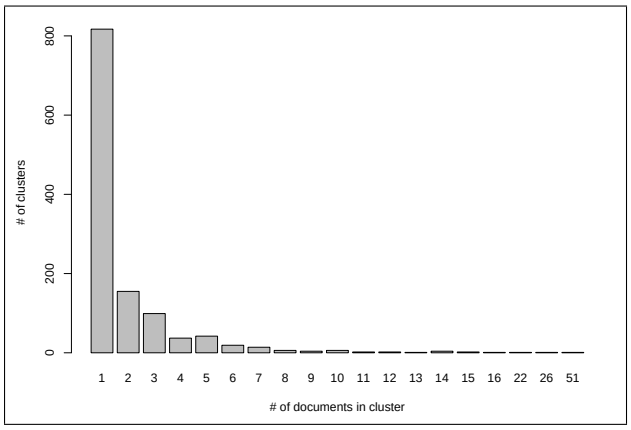


Figure 1: Distribution of event clusters concerning the number of documents they contain

classic $\kappa$ coefficient (Cohen, 1960) - $\kappa = 0.684$. The second one is a modified $\kappa$ - $\kappa_{mod} = 0.91$. The $\kappa_{mod}$ does not take into account different number of event clusters, i.e. if $A_1$ and $A_2$ are sets of document pairs from same clusters, $\kappa$ is calculated as $2 * |A_1 \cap A_2|/(|A_1| + |A_2|)$ whereas $\kappa_{mod}$ is calculated as $|A_1 \cap A_2|/min(|A_1|, |A_2|)$. Namely, in the first annotation, 1.214 events are annotated, while in the second one, only 955 are annotated.
An example of the differences in the annotation is the report about the Cyclone Nargis. Namely, one annotator distinguished the first reports about the disaster and the first appeals for help to the international community as two different events, whereas the other one considered them as a single event. Thereby, the first annotator formed the a large cluster, while the second annotator formed two smaller ones. The instructions given to the annotators were the same - not separating a complex event into subevents only in case it is considered to be impossible. The difference between the $\kappa$ coefficient and the modified $\kappa$ shows clearly that annotators were confronted with such problems quite often.

### 3.3. Setting a threshold for the annotation process

In the end of this section the possibility of setting a similarity threshold in the annotation process is explored. The idea is to limit the documents showed in the first step of annotation - the browsing - to a specific similarity value. Three variables are observed:

1. the percentage of document pairs left for inspection,
2. the percentage of relevant document pairs filtered out by the threshold
3. and the threshold itself

The purpose of the threshold is to minimize the first variable while maximizing the second one, i.e. shortening the browsing list as much as possible while losing as few as possible relevant document pairs. Figure 2 shows the relation between the first two variables. The graph shows that the list of possible document pairs can be reduced down to 4% with almost no loss of relevant document pairs (0.3%).

If the document pair list is reduced to 1%, a loss of 2% of relevant document pairs occurs. These two moments are marked in figure 2 with dots.
It is our opinion that setting a threshold for the browsing step of the annotation process is justified. In order to find the document pairs omitted because of the threshold, this step has to be followed by the search step. The decision on where to set the threshold value depends on the specific implementation.

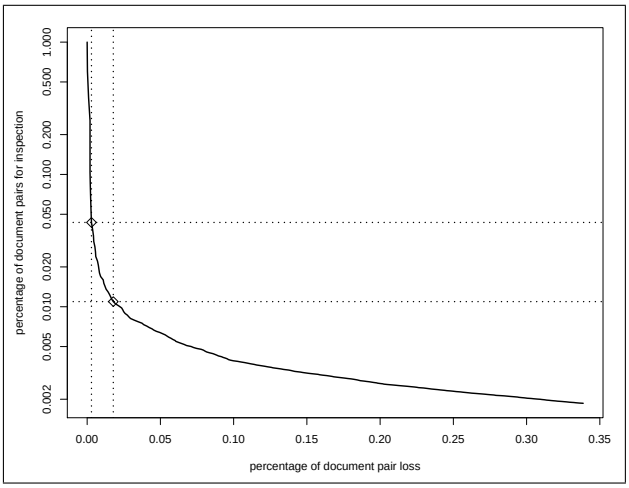


Figure 2: The relationship between the number of document pairs for review and the number of relevant document pairs contained in the list

## 4. The baseline experiment

The baseline experiment consists of experimenting with different clustering algorithms. The experiment is performed on the data published on May 4th and May 5th. The May 4th data sample is used as a development set for threshold parameter estimation, while the May 5th data sample is used as a test set.
The following clustering algorithms are evaluated:

- hierarchical complete-link
- hierarchical average-link
- single-pass (single-link)

The hierarchical single-link algorithm is omitted since its output is identical to the single-pass algorithm.
In the document formalization step lowercased tokens are used as features. The TF-IDF measure is used as the feature weight measure. As the similarity measure, the cosine similarity is applied. All these decisions rely on the results achieved in previous research (Ljubešić, 2009). In addition, no linguistic feature selection or extraction is performed since previous research shows that there is no positive impact of such methods (at least of these that were at disposal at that time) on the task of event detection (Ljubešić, 2009).
Evaluation is performed in a rather novel way for the field of event and topic detection since the whole data set is annotated with respective events. That is the reason why not only a few events are given as evaluation tasks, but the whole sample subset with all events that are described in it. Thereby much more realistic evaluation conditions are provided.

Since evaluation is done on the whole event list, the clustering evaluation method from information retrieval is used (Manning et al., 2008a). Namely, all documents that are grouped into one cluster are transformed into a set of document pairs. Calculating the contingency table produces acceptable values for all parameters but the true negatives since that number is regularly extremely high (there is a large number of documents, i.e. document pairs, that are not grouped into the same cluster). A typical example of the results in the contingency table would be $a = 901, b = 241, c = 432, d = 462.592$. That is why the false negative and false positive rates, i.e. miss and false alarm rates, which are regularly computed in TDT tasks, are not informative. Namely, they produce undistinguishable values for false alarm rates.

For that reason the F-measure is used. That measure relies on precision and recall and therefore it does not use the true negatives value. As the maximizing argument, $F_{0.5}$ is used since precision is considered more important than recall in a multi-source environment where information is regularly repeated. As stated before, the argument maximization, i.e. parameter estimation is performed on the May 4th data set.

The time complexity of the hierarchical algorithms is polynomial ($O(n^2)$, i.e. $O(n^2 \log n)$) while of the single-pass algorithm it is linear ($O(n)$) (Manning et al., 2008b).

The results of the experiment on the May 5th data set is shown in figure 3 in the form of a precision-recall graph.

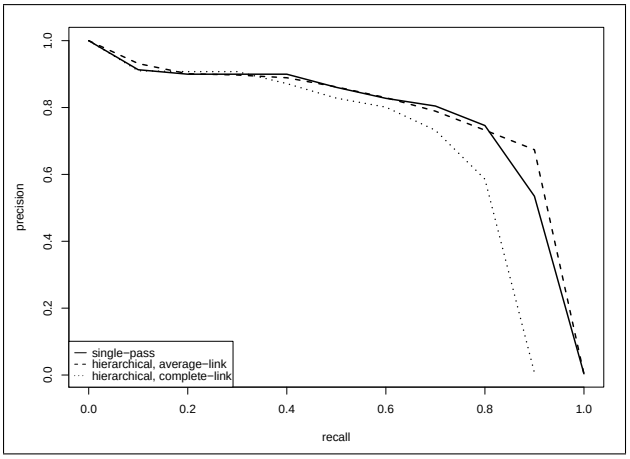


Figure 3: The precision-recall graph for the three clustering algorithms

The results show a rather similar performance of all three clustering algorithms. As expected, the performance differs on high recall, i.e. low threshold levels. The complete-link algorithms underperforms both the average-link and single-pass algorithm. The average-link algorithm outperfoms the single-pass algorithm slightly. For the reason of significant simplicity of the single-link algorithm towards hierarchical algorithms, the single-link algorithm is chosen as the baseline for further research.

## 5. Conclusion and further work

In this paper we described the process of building a gold standard for event detection in the Croatian language. Our approach differs from the approaches in the TDT initiative in more aspects. The corpus is built from documents from sixteen sources simulating the information online environment more accurately. Furthermore, documents are annotated by events they describe, not by topics, and every document in the collection is tagged by a specific event identifier. The statistical analysis of the annotated corpus shows that most documents describe singleton events, i.e. events described by just one document. Therefore most event clusters contain just a few documents, while just a few of them contain a large amount of documents. The two calculated $\kappa$ coefficients show a very high annotator agreement concerning the cluster content, but different cluster organization concerning their size. Additionally, by introducing a threshold in the browsing step of the annotation, the browsing list can be reduced drastically by loosing just a small portion of the relevant document pairs. An initial experiment, combined with results of previous research, has set the baseline for further research. In that experiment, a novel evaluation method is used where not a selection of events is given, but a complete document section with a much higher event number.

Since a very limited number of documents is included in the sample, future plans are to combine this sample with a much broader one where only a selection of events found in this sample would be annotated. Thereby the time limit which is present at this point would be eliminated. Furthermore, at this point a single document is annotated with just one event identifier. Future plans include also experimenting with annotating one document with more than one event identifier. Finally, linking related events concerning their topic could be undertaken, enabling all TDT tasks to be applied on this sample.

## 6. References


James Allan, Jaime Carbonell, G. Doddington, J. Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study: Final report. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

Nikola Bakarić. 2009. Aplikacija za ručno stvaranje klastera dokumenata prirodnog jezika. Master's thesis, Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

J. R. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA.

Nikola Ljubešić. 2009. *Pronalaženje događaja u višestrukim izvorima informacija [Event Detection in Parallel Information Sources]*. Ph.D. thesis, University of Zagreb.

Lawrence B. Lombard. 1986. *Events: A Metaphysical Study*. Routledge & Kegan Paul, Boston, MA.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, 2008a. *Information retrieval*, chapter Evaluation of Clustering. Cambridge University Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008b. *Information retrieval*. Cambridge University Press.

Ron Papka. 1999. *On-line New Event Detection, Clustering and Tracking*. Ph.D. thesis, University of Massachusetts.

Linguistic Data Consortium, 2004. *TDT 2004 - Annotation Manual*.