Tracking of events on news pieces, in Spanish, with a temporal information approach

José Luis Vieyra Sagaón

Dissertação

Mestrado Internacional em Processamento de Linguagem Natural
e Indústrias da Língua

International Masters in Natural Language Processing
and Human Language Technologies

Trabalho efetuado sob a orientação de:
Prof. Doutor Jorge Manuel Evangelista Baptista (U.Algarve)
Prof. Constantin Orăsan, University of Wolverhampton (U. Wolverhampton)

Faro - 2014

Tracking of events on news pieces, in Spanish, with a temporal information approach

Declaração de autoria do trabalho

Declaro ser o(a) autor(a) deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

A Flor y Alejandra, los dos grandes amores de mi vida.

Abstract

Event tracking is a key task in information retrieval given the exponential growth of the information published every day. The detection and tracking of events in news pieces is of great importance for information analysis.
One key feature of event tracking is the temporal similarity that documents have between themselves, this indicates how much a document relates to each another in time, given that events also have strong temporality is natural to imply that these features improve the event tracking task.
The information conveyed in temporal expressions, the feature each has and how the similarity of temporal expressions is addressed, has not been widely studied much less for Spanish language.

For the methodology of this work a corpus was build from RSS feeds and was further annotated with temporal expressions and named entities and then classified each document in topics and then in events, annotation guidelines for the annotators, and evaluation of the annotation was done for each case. Furthermore keywords were extracted during the preprocessing to be sued as features.

Four different temporal similarity algorithms were developed and tested including a baseline, Makkonen's algorithm and two new proposals of new temporal similarity algorithms. These algorithms were used to produce the features the classifier would use in the experiments.

Furthermore a set of experiments were done using different sets of features; the goal was to classify the document in events using different temporal similarity algorithms, to try to determine, both, if temporal similarity improves event tracking, and if it is possible to implement a better temporal similarity algorithm than in the literature.

Finally evaluation of the temporal similarity algorithms trough event tracking, the corpus annotation and the manual classification were done.

The conclusion was that first temporal similarity does improve event tracking as shown in the results, second that frequencies and durations do convey important information although is not conclusive. Finally a platform for further research on temporal similarity and event tracking was achieved.

Keywords

Temporal similarity, temporal expression delimitation, temporal expressions normalization, event tracking, topic detection, machine learning, natural language processing in Spanish, Information retrieval.

Resumo

A Monitorização (ou rastreamento) de Eventos é uma tarefa fundamental na recuperação de informação dado o crescimento exponencial das informações publicadas todos os dias. A detecção e rastreamento de eventos em peças noticiosas é de grande importância para a análise da informação.
Uma característica fundamental da monitorização de eventos é a similaridade temporal que documentos têm entre si, já que isso indica o quanto um documento se relaciona com outros no tempo. Dado que os eventos também têm forte temporalidade é natural considerar que esse tipo de traços podem melhorar a tarefa de monitoramento de eventos.
A informação transmitida em expressões temporais , os traços de cada uma tem e como a similaridade de expressões temporais é abordada, são tópicos que não tem sido muito estudados e muito menos para a língua espanhola.
Para a metodologia deste trabalho, foi construído um corpus a partir de feeds RSS, que foi depois anotado, marcando-se as expressões temporais e as entidades nomeadas. Cada documento foi também classificado em tópicos e, em seguida, quanto aos eventos a que faz referência. Foram elaboradas diretrizes de anotação para os anotadores e a avaliação da anotação foi feita para cada caso.
Além disso foram extraídas palavras-chave durante o pré-processamento, a fim de serem utilizadas como traços.
Quatro diferentes algoritmos de similaridade temporal foram desenvolvidos e testados, incluindo um baseline, o algoritmo de Makkonen e duas propostas de novos algoritmos de similaridade temporal. Esses algoritmos foram usados para produzir os traços que o classificador usaria em seguida nas experiências.
Foi realizado um conjunto de experiências usando diferentes conjuntos de traços; o objetivo era classificar o documento em eventos usando diferentes algoritmos de similaridade temporal para tentar determinar, simultaneamente, se a similaridade temporal melhora a tarefa de monitorização de eventos e se é possível implementar um algoritmo de similaridade temporal melhor dos que são apresentados em trabalhos relacionados.
Finalmente, procedeu-se à avaliação da tarefa de monitorização de de eventos, dos algoritmos de similaridade, da anotação do corpus e da classificação manual.
Em conclusão, os resultados mostram que a similaridade temporal permite de facto melhorar a tarefa de monitorização de eventos; as expressões temporais de frequência e de duração parecem transmitir informações importantes, embora os resultados não sejam ainda conclusivos; finalmente, foi possível com este projeto desenvolver uma plataforma para novas pesquisas sobre similaridade temporal e monitorização de eventos.

Palavras-chave

Similaridade temporal, delimitação, classificação e normalização de expressões temporais, monitorização de eventos, detecção de tópicos, aprendizagem automática/de máquina, processamento de linguagem natural, recuperação da informação, espanhol.

Resumo Alargado

A Monitorização (ou rastreamento) de Eventos é uma tarefa fundamental na recuperação de informação dado o crescimento exponencial das informações publicadas todos os dias. A detecção e rastreamento de eventos em peças noticiosas é de grande importância para a análise da informação.

Uma característica fundamental da monitorização de eventos é a similaridade temporal que documentos têm entre si, já que isso indica o quanto um documento se relaciona com outros no tempo. Dado que os eventos também têm forte temporalidade é natural considerar que esse tipo de traços podem melhorar a tarefa de monitorização de eventos.

A informação transmitida em expressões temporais , os traços de cada uma tem e como a similaridade de expressões temporais é abordada, são tópicos que não tem sido muito estudados e muito menos para a língua espanhola.

Os objetivos deste projeto foram: melhorar a medida de similaridade temporal, proposta por Makkonen; desenvolver um sistema de classificação de eventos, usando aprendizagem automática (ou de máquina); e , finalmente, construir uma plataforma que permitisse a realização de experiências e estudos diversos na área quer da similaridade temporal quer da monitorização de eventos.

Para a metodologia deste trabalho, foi construído um corpus a partir de feeds RSS, que foi depois limpo e anotado, marcando-se as expressões temporais e as entidades nomeadas. Cada documento foi também classificado em tópicos e, em seguida, quanto aos eventos a que faz referência. Foram elaboradas diretrizes de anotação para os anotadores e a avaliação da anotação foi feita para cada caso. Além disso foram extraídas palavras-chave durante o pré-processamento, a fim de serem utilizadas como traços.

Em relação à coleta do corpus, o procedimento consistiu primeiro em obter feed de notícias RSS de jornais on-line e, em seguida, aceder automaticamente ao link que os fornece e extrair o texto dos principais jornais locais. Em seguida, foi feita uma limpeza e pré-processamento desses textos em parte para extrair características como as palavras-chave com a estatística TF- IDF e em parte para fornecer esses textos ao anotador humano num formato facilmente legível a fim de facilitar seu trabalho.

Em seguida, as expressões temporais e entidades nomeadas presentes no corpus foram anotadas e classificadas por anotadores humanos. As expressões temporais foram objeto de um processo de normalização, usando uma ferramenta especialmente desenvolvida para ajudar os anotadores nessa tarefa. Além da delimitação manual e classificação por tipos (datas, durações e frequências), as expressões temporais foram também classificados pela forma como os seus limites são definidos (completamente definido, fronteira inicial ou final difusa ou limites completamente ambíguos).

Além disso, procedeu-se a uma classificação manual das notícias em tópicos e, dentro destes, constituíram-se conjuntos de eventos, a fim de se determinar um limite máximo (ceiling) para a performance do sistema e para analisar a concordância entre os anotadores humanos contra o desempenho do sistema. Isto também foi feito para assegurar a qualidade das classificações. Três anotadores intervieram na marcação das expressões temporais e na anotação de entidades e sua classificação. Outros dois anotadores procederam à normalização das expressões temporais.

Quatro diferentes algoritmos de similaridade temporal foram desenvolvidos e testados, incluindo um baseline, o algoritmo de Makkonen e duas propostas de novos algoritmos de similaridade temporal.

A baseline consiste em usar apenas a data de publicação das peças noticiosas como característica de similaridade temporal. O algoritmo de Makkonen leva em conta as expressões temporais como intervalos de início e data de fim e este foi considerado para o desenvolvimento do resto dos algoritmos. Os outros dois algoritmos consistiram numa maior granularidade da que foi usada por Makkonen no seu algoritmo relativamente às expressões de tempo do tipo data, tendo o segundo levada também em conta as expressões de frequência e as de duração para o cálculo de similaridade temporal.

Esses algoritmos foram usados para produzir a medida de similaridade temporal que funcionará como um dos traços que o classificador usa em seguida nas experiências.

Foi realizado um conjunto de experiências usando diferentes conjuntos de traços; o objetivo era classificar o documento em eventos usando diferentes algoritmos de similaridade temporal para tentar determinar, simultaneamente, se a similaridade temporal melhora a tarefa de monitorização de eventos e se é possível implementar um algoritmo de similaridade temporal melhor dos que são apresentados em trabalhos relacionados.

Para a realização das experiências utilizou-se o algoritmo 'K-nearest neighbor' (vizinho K mais próximo) para construir o classificador, usando a implementação que é disponibilizada pelo conjunto de ferramentas de aprendizado de máquina Weka. Para economizar tempo, as experiências foram implementadas e automatizadas na linguagem de programação Java. A avaliação foi programada usando as classes de avaliação do Weka.

As experiências consistiram na combinação de diferentes conjuntos de traços, como palavras-chave e entidades nomeadas com os diferentes conjuntos que os algoritmos de similaridade temporais produziram. No total, cerca de 19 experiências foram realizadas e calcularam-se os resultados médios para obter uma medida global do desempenho dos algoritmos na tarefa de monitorização de eventos.

Finalmente, procedeu-se à avaliação da tarefa de monitorização de de eventos, dos algoritmos de similaridade, da anotação do corpus e da classificação manual.

Os resultados das experiências na tarefa de monitorização de eventos situam-se à volta de 40% de F- medida, sendo o teto (ceiling) definido com base no acordo inter-anotadores de cerca de 72%, o que dá uma ideia clara do desempenho do sistema.

Em relação à similaridade temporal, os resultados mostraram que, sem recurso a qualquer medida de similaridade temporal, é possível alcançar uma maior precisão; no entanto, se qualquer daquelas medidas for acrescentada (excepto a baseline), a precisão cai, mas a abrangência (recall) aumenta.

Isso explica-se porque a similaridade temporal traça um perfil dos documentos no tempo e em muitos eventos, especialmente se as peças noticiosas foram recolhidas durante os mesmos dias, como sucede no caso do nosso corpus, eles acabam por ter perfis temporais semelhantes.

O melhor algoritmo para similaridade temporal foi o que levou em consideração as expressões de frequência e de duração, contrariando a ideia de que este tipo de expressões só acrescentaria ruído ao sistema, o que se provou ser errado. Em segundo lugar ficou o algoritmo com a granularidade alterada, finalmente, e algoritmo de Makkonen, em terceiro lugar.

Em conclusão, os resultados mostram que a similaridade temporal permite de facto melhorar a tarefa de monitorização de eventos; as expressões temporais de frequência e de duração parecem transmitir informações importantes, embora os resultados não sejam ainda conclusivos; finalmente, foi possível com este projeto desenvolver uma plataforma para novas pesquisas sobre similaridade temporal e monitorização de eventos.

Acknowledgments

I would like to express my sincere gratitude to the **European Union** that through its **Erasmus Mundus program** granted me the scholarship to be part of this amazing Master.

Also I would like to express my gratitude to my supervisors **Prof. Jorge Bap**tista and **Prof. Constantin Oraŝan** for their strong support in the development of this work.

Also I would like to thank of the staff in the universities of the consortium that somehow helped me during these two years: **Gabriel Secondat**, **Prof. Sylviane Cardey**, **Prof. Ruslan Mitkov**, **Prof. Angels Catena**, and all the staff, teachers, assistant and all the nice people that were there when we needed them.

I also would like to thank my friends who helped me during the development of the work:

First **Eleni Tsironi**, thank you for all your support and help during endless sleepless nights of hard work, friends like you never in life!

Second **Jorge Lázaro "Banda"** for helping me in coordinating the corpus annotation as well as providing direct contact with the annotators. !*Gracias bandita!*.

Third the annotators themselves that made this work possible with their hardwork and many hours dedicated to these purpose, they are: **Erika Sanchez**, **Karina Granados**, and **Monica Gonzalez**.

I would also like to thank all my friends who were here giving their unconditional support while developing this work and during the warm Faro nights namely: **Katherin Pérez**, **Hector Peterson**, **Iacer Calixto**, **Aris Karam**, **Artur Yeghiazaryan**, **Iraklis Tsigardis**, **Nicolo Moruzzi**, **Ilia Markov**, and all the others in Wolverhampton, Barcelona and Faro.

Also to specially thank my aunt **Erika Ehnis Duhne**, my godmother **Susana Goñi** and my sister **Gabriela Vieyra** for all their support during this stage and other stages of my life. Also all other family and friends I might be forgetting.

And last but most important of all, I would like to express my full gratitude to **Blanca Flor Ramos Gamboa**, you made all this possible, so much we have had to suffer, so many things kept us apart this time, however I want you to know that this work is dedicated to you, hoping that all the things we had to sacrifice are worth it in the fruits we will harvest from all we invested in this master. Be sure to know this work was done thinking on you, as you have been in my heart all the time this master took.

Table of contents

Table Index

Figure Index

1. Introduction

This work approaches the task of event tracking and the task of temporal similarity also analyses the relation influence of temporal similarity in event tracking.

The corpus and the system were developed for Spanish news pieces taken from Mexican news papers.

1.1 Aims

This project focuses on how temporal similarity affects event tracking, how to address the problem of temporal expressions and the information they convey, and how to apply this information towards correctly identifying events in sets of news pieces. The following aims have been defined:

- Propose a new improved algorithm for temporal similarity measure of documents.

- Develop a system capable of classifying news pieces into events using the said algorithm plus conventional means like keyword vectors and named entities and perform experiments on how temporal similarity affects event tracking.

- Build a platform which could be used for temporal information analysis and further development of topic detection, event tracking and temporal information retrieval tools.

1.2 Motivation

Nowadays information retrieval is a key topic in data management, every day's global output of data increases and the need for appropriate tools to address this data is also growing.

In particular every day newspapers publish their news in their websites, on social networks and still some on them in paper. Furthermore social networks play a significant role on how this news are distributed, in particular Facebook [1]and Twitter[2].

Furthermore, newspapers, online news sites and others sources publish trough RSS feeds their content, making it available to a whole community receiving the feeds to spread it.

The analysis of said information is a key priority for governments, institutions, companies and people interested in mining this information for analysis purposes or simply to keep themselves informed in a particular topic or event.

The detection and classification of said topics, events and they information they convey motivated this work.

Furthermore, temporal information in news pieces is expected a key feature, given that the literature is not wide on this topic and that there are no reliable available tools for temporal expressions extraction and normalization for Spanish motivated this project.

First a tool to detect classify and normalize named entities and temporal expressions were desired, however the lack of appropriate tools and papers for these tasks motivated the annotation of the corpus and the manual temporal expression normalization and the tools developed for it.

This raised the question that if it is possible in a future to build a accurate event tracking system; which would have all the tools and modules (named entity recognition and classification, temporal expression extraction and normalization) which would be developed using the corpus of this project; to track events automatically starting from the RSS feed to the cluster of events.

[1] Facebook: www.facebook.com
[2] Twitter: www.twitter.com

1.3 Research questions

The research questions proposed, address the project in two stages; first the temporal similarity of documents, namely the information temporal expression convey in two documents, and second the event tracking of the corpus news pieces using amongst other things, the temporal similarity:

1. What makes two documents temporally similar?
    a. Two documents are temporally similar if they contain the same or near the same temporal intervals, durations and frequencies?
    b. Which kind of temporal expressions and which features contains more information on the document similarity?
    c. Do the types of temporal expressions for frequency and duration convey useful information?
    d. Is temporal similarity a key feature for distinguishing topics from events in news?

2. Could it be possible to identify the same event from news pieces of a given topic using temporal similarity features and traditional feature approaches? If so:
    a. Could Makkonen's temporal similarity algorithm be improved and tested in the evaluation of the performance of a machine learning algorithm that uses temporal features to cluster events?
    b. How the common high class imbalance of events in news corpus affects the tracking of events?

1.4 Overview of this work

Today's news streams and the multiple sources they come from, present a big problem for information analysis, the available tools for topic detection often do not consider temporal information, relying only in terms and simple named entities recognition [Kumaran, Allan., 2002].

In news streams, some news reflect the same event in real life i.e. *three different news from different papers reflecting a social outburst in Turkey*. The news may come before, during or after the event has taken place. The events per definition as [Yang et al., 1999] say "*An event is a dynamic topic, a subject that is discussed intensely in the news at some time*."

It's easy to suppose that the date on which the news was published is critical in this task, however the other temporal information reflected in temporal expressions either if they are explicit (*March 17th*) , implicit (*the day after tomorrow*) or ambiguous (*in a few days*) might play a big role in the news event detection.

In the literature the most common features used for topic detection and event tracking are keyword vectors, named entities of names, institutions, etc. and of course temporal similarity. These five concepts: **keyword vectors**, **named entities**, **temporal expressions**, **topic detection and event tracking**, **temporal similarity** and the **evaluation metrics** used are described on **§2 Literature review**.

A **corpus** was build from on-line resources, more in detail the RSS news feeds from on-line newspapers, and annotated with **temporal expressions**, **named entities** and **their classifications**, evaluation was done to the delimitation and classification of both, also **temporal expressions were normalized**.

Furthermore the corpus was **classified by topics** and, the in-topic documents classified in events, in order to evaluate the **event classification**; the methodology for the corpus gathering, annotation, normalization and classification is described in chapter **§3.1 Corpus**.

Some **preprocessing** was needed in the corpus, cleaning, line and word tokenization, statistical measures, etc. in order to extract the desired features; this is also described in **§3.2.1 Preprocessing**, and the feature extraction in **§3.2.3 Feature extraction**.

After with the results of the preprocessing, and the annotation, features were calculated and selected to provide a **feature space** to represent each document. Some is this features are actually temporal similarity features.

The proposal was to develop a new algorithm for temporal similarity of documents, taking in account features of them such as: vagueness, duration, ranges, most common temporal expressions, disambiguation errors, and etc.; which is described on detail on **§3.2.2 Temporal similarity.**

Already [Makkonen et al. 2002] did an event detection approach for news pieces; Her approach to the problem was to split the document in four vectors of terms, names, locations and temporal similarity information.

With this features the event tracing experiments took place which is described in **§3.2.3 Event Tracking**. The **event tracking experiments** consisted in using the k-nearest neighbors implemented in the machine learning kit Weka, using the feature spaces calculated as instances, to try to predict the event class of which each document belongs.

The experiments consist in using different sets features to classify the events, these sets are changed, and mostly the features are: keywords, named entities, and the temporal similarity measures using the four proposed algorithms: the baseline, Makkonen's algorithm, and two improved Makkonen's algorithm proposals.

Finally, a set of **evaluation measures** were done in **§4 Evaluation**, first for the corpus annotation and manual event and topic classification; and for each topic the precision, recall and F-measure of the classifier, the overall precision, recall and F-measure of the system for each of the experiments stated above is given.

At the end discussion of the results, the implication of this work and the results and the feature work is stated in **§5 Discussion and conclusion**.

Furthermore, **§References** and relevant appendixes like annotation guidelines and a complete annotated document are included.

2. Literature review

In Natural Language Processing (NLP) and especially in Information Retrieval (IE) there are some subtasks dedicated to extract entities such as person names, organizations, places and times. Such subtasks have become critical for modern information extraction systems.

In this subtask known as Named Entity Recognition and Classification (NERC) which is treated now as a full topic worth of studying it by itself in NLP, several other subtask have been born, and even developed as a full topic of study in information retrieval, for example the Temporal Expressions Extraction and Normalization.

Furthermore, research is being carried also in the processing, comparing them, and operations between the data extracted this is the case for example of Temporal Similarity (TS) and geo-localization of documents by the places names contained on it.

There are several applications of the information gathered in IE intelligent systems, for example they can be used as features in Topic Detection (TD) and in Event Tracking (ET) technologies, also for discovering links between documents such as person relatedness, geo-temporal tracking of social networks feeds, etc.

2.1 Named entity recognition

Named entities (NE) are segments of language that might refer to a name, a location, a place, a date, currency numbers, etc.

Named entity recognition and classification are one of the main topics researched in Information Retrieval and in natural language processing, from the beginning it included the extraction of temporal expressions however given the complexity of temporal expressions and the amount of research on them, they are now considered a new topic [Nadeau, Sekine, 2009].

2.1.1 Overview on named entities recognition and classification

Name Entity Recognition (NER) was coined at the sixth Message Understanding Conference (MUC-6) by [Grishman, Sundheim. 1996], but it was born in a paper by [Rau, 1991] in which she extracted company names from text.

At the begging of the research of this field, named entity recognition consisted in extracting entities and classify them in categories, like location, person's name's organization, dates, emails, etc. Over the years the dates and other temporal information has separated in other information extraction field (the temporal expressions extraction and normalization) which will be discussed further in this work.

Named Entities recognition and classification (NERC) started its existence with the paper of Rau on 1991 to extract company names, she used handcrafted rules and heuristics to extract them, the field was not developed considerately until the 1996 with the MUC-6 task organized by Grishman and Sundheim [Grishman, Sundheim., 1996], in this task they asked to extract named entities and classify them in the following categories: organization names, person names, location (political and geographical), temporal expressions (dates and times) and numerical expressions such as monetary and percentage expressions.

From that point and on, named entity extraction has been a field in continued research, with many papers and conferences on the subject.

However there are several factors that need to be considered, the same way different research lines follow these factors. These are:

1. Language factor

Most work done in NERC is for English language [Nadeau, 2007] for obvious reasons like the language usage as *lingua franca* in the scientific community and because is the most used text on the internet. Some work in English has been done by: [Nadeau et al. 2006], [Narayanaswamy et al. 2003] and [Heng, Grishman, 2006].

Some languages independent studies exist possibly in a larger proportion than the ones that just address English [Nadeau, Sekine, 2006]. Work on this field has been done by [Steinberger, Pouliquen, 2006], [Cucerzan, Yarowsky, 1999] and in the CoNLL-2003 shared task of language independent named entity recognition [Erik, et al. 2003]

Also Spanish, Portuguese (in Harem tasks [Mota, Santos, 2008]) and German [Erik, et al. 2003] are heavily studied in this task.

In the particular case of Spanish we have the work of [Galicia-Haro, Gelbukh, 2009], [Arevalo et al. 2002] and [Toribio et al. 2010]

Also some other work has been done for other languages according to [Nadeau, Sekine, 2009] such as French, Japanese, Chinese [Wang et al, 1992], Basque, Catalan, Dutch, Danish, Russian, and Arabic [Darwish, 2013].

2. Textual domain or genre factor

This is very important factor because most of the work done in NER uses training corpus made only of one kind of genre or domain of the texts, for example work has been done for emails [Minkov et al. 2005], scientific texts (specially on the field of Biology) [Narayanaswamy et al. 2003], news corpora [Shinyama, Sekine, 2004] and even in new information sources like social networks in particular Twitter [Ritter et al., 2011].

In [Nadeau, Sekine, 2009] work they mention that most systems cannot be so easily ported to another kind of genre or domain, loosing performance of between 20% to 40% of precision and recall when tested on a different domain of texts that the one they were intended to work with.

3. Entity type factor

Other factor of consideration when working with this system is the categories in which by extent of the recognition of the named entities are classified.

As the name Named entities states the task is restricted only of those entities with one or more rigid designators. At the beginning of this field of study also temporal expressions and numerical expressions were included in the task [Nadeau, Sekine, 2009].

2.1.1.1 The delimitation of problem

Also one very relevant factor to take into account both at designing time and in evaluation is how the named entities will be delimited.

Simple decisions play a big role on named entity delimitation for example either the articles will be considered as part of a named entity or not.

Another very important thing during the evaluation of the system is if the delimitation of the entities would be evaluated, this means that if a tagged NE will be considered as correct even if their boundaries were recognized wrong, and if so, how correct is a NE not well delimited for the overall evaluation of the system.

2.1.1.2 Classification of named entities

The natural following step after extraction named entities would be classify them in categories, several categories has been used in the literature ranging from two or three to up to 200. Also sometimes the categories contain subcategories.

Even in one of the first approaches the MUC-6 task [Grishman, Sundheim., 1996], some categories for named entities were defined and the results of the classification considered on evaluation time. The categories required to be tagged were organization and persons names, locations, temporal expressions and number expressions

The term Enamex refers to the conjunction of the extraction of person and organization names, and locations, and was coined also at the MUC-6 conference [Grishman, Sundheim., 1996].

Each author according to his interest and corpus domain selects the categories in which named entities (especially Enamex ones) would be tagged, normally they define subcategories of named entities in order to be less generic or to extract only subtypes of interest. Usually temporal expressions contain its own subtypes and are now treated as a new task described further on this work.

The granularity of the categories depends on many factors and involving more and more categories provides a more complex problem of classification for the machine. Nevertheless authors have gone as far as having 200 or more categories in their approaches [Nadeau, Sekine, 2009].

2.1.2 Common approaches

In the beginning of the field, this task was mostly done by handcrafted rules encoded in regular expressions, but soon the field considered machine learning techniques using Supervised, semi supervised and unsupervised approaches.

1. Supervised Learning
   Usually the task is approached having a large annotated corpus; the system memorizes the entities and creates disambiguation rules based on discriminative features. Some machine learning algorithms using the said approach are Hidden Markov Models (HMM), Decision Trees, Maximum entropy models (ME), support vector machines (SVM) and conditional random fields. [Nadeau, Sekine, 2009].

2. Semi supervised learning approaches
   In this approach the systems are given asset of seeds of named entities for starting the learning process, the system searches for sentences with these seeds and identify contextual features common to the seed examples, then the systems looks for these contextual patterns in the corpus and obtains new named entities, then process is repeated until no more knowledge can be obtained from the corpus. [Nadeau, Sekine, 2009]

3. Unsupervised learning approaches
   These methods are normally a clustering task and rely on a big unannotated corpus, the systems perform analysis of the context of the sentences in the corpus and then cluster similar contexts and obtain the named entities. This also is usually achieved using lexical resources such as WordNet, lexical patterns and on statistics computed on the corpus. [Nadeau, Sekine, 2009].

Another very important matter is the feature spaces used for the machine learning approaches, features are information collected on different linguistic levels in the corpus, usually they lexical, positional, contextual, morphological, etc.

In the NERC task with machine learning, the set of features varies between approaches, but they can be classified in three categories:

1. Word-level features

   Several features can be used in a word level, some depends on characters, patterns, word endings and stems, etc. The following table extracted from [Nadeau, Sekine, 2009, pp. 12] resumes these kinds of features.

Feature	Examples	Comment
Case	Starts with capital letter	These features come from the letter case of the word.
	Word is uppercased	
	The word has mixed case	
Punctuation	Ends with period or has internal period	The punctuation of the named entity in the text.
	Internal apostrophe, hyphen and ampersand	
Digit	Digit Pattern	Numbers are usually found in named entities or around them.
	Cardinal and ordinal	
	Roman number	
	Word with Digits	
Character	Possessive mark, first person pronoun	The same the possessive can indicate an entity, the greek letter usually are used as or in names or entities
	Greek letters	
Morphology	Prefix, suffix, stem, singular version	The morphological features provide solid information about the context and the entity itself.
	Common ending	
Part of Speech	Proper name, verb, noun	Nouns are usually better candidate to be or be inside a named entity.
Function	Alpha, non-alpha, n-gram	The function uses features from the word structure for example a summarized pattern of uppercase, lowercase, punctuation and numbers.
	Lower, uppercase, version	
	Pattern, summarized pattern	
	Token length, phrase length	

Table 1 Word-level features for NERC

2. Dictionary features

   Dictionaries and lexical resources are widely used features; they are used for look up abbreviation and company names. In the following table also from [Nadeau, Sekine, 2009, pp. 14] the dictionary features are shown:

Features	Examples	Comments
General List	General dictionary	For disambiguating capitalized words, lists of common abbreviations, etc.
	Stop words	
	Capitalized nouns	
	Common abbreviations	
List of Entities	Organization, government, airline, educational	To look for entity already recognized or known.
	First name, Last name, celebrity	
	Astral body, countries, continent, state, cities	
List of entity cues	Typical words in organization names	To give the system clues on where a named entity could be.
	Person title, name prefix, post-nominal letters	
	Location typical word, cardinal point	

Table 2 Dictionary-level features for NERC

3. Document and corpus features
   Also, information gathered in the whole source document are used, this features rely on frequency, position in the text and meta information from the source. Again this table is taken from [Nadeau, Sekine, 2009, pp. 16]

Features	Examples	Comments
Multiple Occurrences	Other entities in context	The context, how the word is used in the text and resolution of anaphora are helpful features for this task.
	Uppercased and lowercased occurrences	
	Anaphora Coreference	
Local Syntax	Enumeration, apposition	The position of the word plays a role in the probability of it being part of a NE.
	Position in sentence, paragraph and in document	
Meta information	Uri, email header, XML section,	Some works uses meta information from the html source to aid on the NERC.
	Bulleted/numbered lists, tables, figures	
Corpus Frequency	Word and phrase frequency	The word frequency and its occurrence in the corpus are important features.
	Co-occurrences	
	Multiword unit permanency	

Table 3 Document-level features for NERC

These methods and the features used represent the great majority of the approaches used in named entity extraction and classification, features and machine learning algorithms may vary, however the guidelines are mostly the same.

2.1.3 Applications of named entities recognition and classification

Nadeau [Nadeau, 2007] describes the common applications of the named entity recognition field, some of them are:

- Personal name disambiguation.
  This field treats with trying to obtain the correct name of persons and disambiguate if a name belongs to one person or other.

- Identification of named entities definitions.
  This field this deals with extracting from the text itself not only the named entity but also the definition of it in the text.

- Named entities translation.
  They are used for multilingual systems and automatic translation systems.

- Topic based clustering.
  The NE's are used as features in the grouping of documents by their topic, also this is a very important feature in event tracking which both will be described in depth further.

- Anaphora resolution.
  Research is being done on anaphora resolution, usually named entities represent the main entity towards the articles and other referential particles will be mapped.

These are just some of the applications that rely fully or partially on NERC.

2.1.4 Evaluation of NERC and results

In NERC evaluation several considerations have to be taken into account, the first is to know what exactly will be evaluated and second how to measure it.

In the MUC tasks [Grishman, Sundheim., 1996] five type of errors were defined which are:

1. The system tagged an entity where there was none.
2. An entity was completely missed by the system.
3. The system detects an entity but categorize it wrong.
4. A system detects an entity but gets its boundaries wrong.
5. The system categorized it wrong and also got the boundaries wrong.

This being said there are three main guidelines for evaluating the NERC task. They are the MUC evaluations, the exact match evaluations and ACE evaluation.

1. In the MUC evaluation, the F measure is used as the indicator combining precision and recall of the system. Furthermore it evaluates two things, first that the entity is at least partially recognized, regardless of their boundaries, and second that the type is assigned correctly [Grishman, Sundheim., 1996].
2. The exact match evaluation considers one correct match if the system recognizes it with its boundaries correct and assigns it the correct categories, if not the match is wrong. This approach is stricter because many of the problems of NERC consist on the actual boundaries of the named entity usually only recognizing it partially. They were used mainly in the IREX [Sekine, Eriguchi, 2000] and CONLL [Erik, et al. 2003] tasks.
3. The ACE evaluations more complex assigning weights and values for all the kind of errors described above, also gives weight to the possible categories and has some adjustable error controlling values. It was used mainly in the ACE task [NIST, 2008].

Finally the results on the field had improved over the years. Given that there are several metrics to measure and different domains, it is not that easy to test one system against each other.

Several work has been done in either measuring the performance of a NER system in different domains like the one done in [Ratinov, Roth, 2009] in which they test the Stanford NER parser and the LBJ-NER against different domains getting in all of them an average of 80% F measure.

In testing different system against each other there are examples like the one of [Marrero et al. 2009] in which they test several systems known public NER systems such as Supersense[3], Afner[4], Annie[5], Freeling[6], TextPro[7], YooName[8], ClearForest[9] and

[3] http://medialab.di.unipi.it/wiki/SuperSense_Tagger
[4] http://afner.sourceforge.net/
[5] http://www.aktors.org/technologies/annie/
[6] http://nlp.lsi.upc.edu/freeling/
[7] http://textpro.fbk.eu/
[8] http://infoglutton.com/yooname-named-entity-recognition.html
[9] http://www.clearforest.com/

Lingpipe[10]. All the F measure values of both identification and classification were between 50%-50% for YoName and 85%-75% for Supersense trained with the WNSS [11]corpus.

In the particular case of Spanish work has been done like the of [Toribio et al., 2010] in which they evaluated three research systems and OpenCailas an open source NER system with Spanish as one of their languages. They evaluated the work of Ferrandez, Carreras, Florian and the OpenCalais system with the evaluation Corpus of the CONLL-2003. The results they obtained taken from [Toribio et al. 2010, pp 290] shown their best results of their research:

System	P (%)	R (%)	F
Ferrández, 2006	83.34	83.41	83.37
Carreras, 2002	81.38	81.40	81.39
Florian, 2002	78,70	79,40	79,05
OpenCalais	66.80	43.68	52.82
Baseline	26.27	56.48	35.86

Table 4 Results of NERC in Spanish with CONLL-2002

[10] http://alias-i.com/lingpipe/
[11] http://saffron.deri.ie/lrec/topic/semcor_corpus/

2.1.5 Discussion about named entities and their role on this work

The named entities contained in the documents are crucial in the development of more complex NLP systems and innovative research on the field. The NE, and their features are solid bases for topic tracking and detection.

The set of named entities contained in a text has proven to work well by themselves in document categorization and topic detection [Friburger, Maurel. 2002], temporal similarity will only improve the results as the core of the similarity would be taken by the named entities.

Different approaches has been made to solve this task, however, still there is the need to develop further this field, for example on developing new methods to treat with more topic wide texts or for general purposes.

Also even so that in Spanish wide research has been there is a lack of reliable open source tools to work with this topic, which to me is a clear indicator that NERC even so that it has been around for more than fifteen years is still subject of ongoing research.

2.2 Temporal expressions

Temporal expressions (TE) are also pieces of text which convey temporal information on them, these expression are treated now as a separate field of information extraction and from natural language processing.

The work with temporal expressions can be divided in two task, first the recognition or extraction of TIMEX, and then the normalization of them, this mains to put the expressions in a canonical form in which operations could be done on the normalized TIMEX sets.

2.2.1 Overview on temporal expressions

Temporal expressions (TE) are defined by [Ahn, Fissaha-Adafre., 2005] as language fragments that allude directly to time instants or time intervals, they not only provide temporal information by themselves, but they serve as anchor points for locating events cited in a text.

The TE's are classified in several types according to [Marsic, 2011]:

1. Fully specified
   When the value can be normalized with the information contained in the temporal expression itself, they are also known as context-independent or absolute times.
2. Underspecified
   This kind can't be normalized by themselves and need the use of another TE or context from the text itself.

Another classification also from [Marsic, 2011] is:

1. CALPOINT
   This is a point itself in the calendar in marks a time instant and temporal expressions are grouped by their granularity. The granularity would represent time sets by how long is their time frame and could be millennium, century, year, month, week, hour, second, minute and so on as required.
2. DURATION
   These are ranges that can be described both explicitly and implicitly; they refer to TE that defines a time frame in the text. In other words they express a time interval.
3. FREQUENCY
   This kind of TE's express how often events happen in time. i.e weekly, daily, monthly, etc.

The date on which the news was written is also a key aspect for normalizing this type of named entities.

There are also TE that cannot be anchored in a timeline nor normalized with the information contained directly or indirectly in the documents. For example: "*some time in the future*" or "*In the following years*"

2.2.2 Temporal expressions detection and normalization

There are two tasks in this field, the recognition and the normalization. The recognition deals with all the problems of named entities extraction, however different considerations and factors have to be kept in mind, for example the type and the extent of TIMEX that are needed to recover, the granularity of them and, of course, the purpose of the extraction.

In her work [Marsic, 2011, p 119.], defines the temporal expression annotation as:

> *The automatic TE annotation process involves two processing stages. The first stage is concerned with identifying the textual extent of the temporal expressions present in the processed text, and is normally referred to as temporal expression identification. The second stage of the annotation process is called temporal expression normalization, and its aim is to find the value that the expression designates or is intended to designate.*

Usually both stages are needed, but they have to be treated separately.

2.2.2.1 Temporal expressions extraction

The extraction of temporal expressions also started at the same time as the NERC, in the MUC conference said before, first simple expressions and basic normalization was required but with the advance of the field the complexity of the task has increased and diversified.

The extraction of temporal expressions rely on many factors of consideration, also some features are used in the machine learning algorithms as well as in the tags needed if the task include tagging in the document.

The factors of consideration are as follows:

1. Types of temporal expressions going to be extracted
2. The granularity, fuzziness, and expressions with undefined temporal boundaries
3. The method to be used in the extraction
4. The features, rules and grammars required

These factors are critical in the designing of an algorithm or system that extracts temporal expressions. Another very important matter is if the expressions are going to be tagged or not which will be treated further in this work.

Another critical factor needed to be considered is the delimitation problem of the temporal expressions, either if the articles will be included or not and to do some cleaning for the next stage of the processing which is the normalization.

2.2.2.2 Temporal expressions normalization

Once the TE have been extracted is needed to convert them into a standard format in which the time information conveyed can be normalize to a given standard and can be used more for operations with this information.

Traditionally this work has been with the help of grammars and dictionaries, because once the temporal expressions have been extracted, is easier to apply regular expression recognition or feed them to an automata which normalize this information.

It is very important to state that the normalization using rules, usually only handles the most common temporal expressions and have some problems dealing with fuzziness and stating the boundaries of dark expressions.

2.2.2.3 Annotation schemes

The first schemes used to tag TE was designed in the MUC-7 conference [Chinchor, 1998] which was latter called the First TIMEX scheme, it was characterized for having only one attribute (type).

The later TIMEX2 had some added attributes and was annotated with the ISO-8601 [ISO8601:2004, 2004] scheme for temporal information representation. This scheme was developed on the year 2000 and was first documented by [Ferro et al.,2000] with further development.

The main advantage of TIMEX2 is that expressions can be assigned a normalized value of temporal expression in a standard format, which encourages machine computation of these values.

Also TIMEX2 provides attributes to fill extra features of the temporal expression, like the type, the value, comments and anchor values to fully normalize the temporal expression.

Another markup scheme is the TimeML developed by Pustejovsky [Pustejovsky et al. 2005] it extendes the TIMEX2 scheme but also considers a whole range of temporal information in text such as verbal tenses, verbal event, etc. However it can be used only to annotate only the temporal expressions.

2.2.2.4 Common approaches

The detection of TE as in NER is approached by two main research streams the first is the linguistic rules definition and the other is the machine learning approach.

Approaches that uses rules like TERSEO [Saquete, 2003] and Vicente-Diez's [Vicente-Díez et al. 2007] usually use a context free grammar for identify them or a finite-state automata plus some linguistic rules that are coded using mostly POS tags and another features that helps for this task. This works were made specifically for Spanish.

Other approaches use machine learning ([Ahn, Fissaha-Adafre., 2005], [Ahn, et al., 2007], [Filannino et al., 2013]), which rely on a machine learning algorithm and after

they do a post-processing identification pipeline, which is a set of rules and processes used to filter "bad" results from the ML algorithm.

In particular, the tool Tipsem [Llorens et al. 2010] and TERSEO [Saquete, 2003] are machine learning approaches for Spanish.

The previous work reflected how those approaches works, both of them has advantages and disadvantages, which will be discussed ahead with typical examples of results:

1. Rule-based

This approach is normally very rigid, and every time new patterns of TE, or new rule for classify them arise they have to be coded inside the program or defined manually in the rule set which the system uses.

On the other hand, the rules, regular expressions, grammars, etc. are easily modifiable which is an advantage at adding new rules in the systems and it is virtually impossible to do it in machine learning approaches without training the system all over again with new rules.

To show the typical structure of these systems, the structure of TERSEO is shown in the next figure:

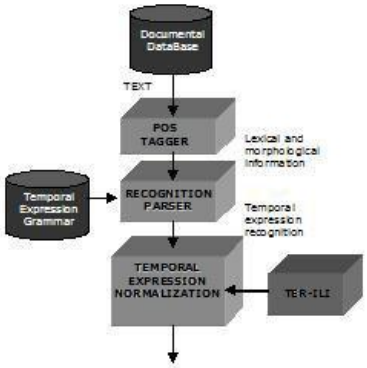


Figure 1 TERSEO system structure [Saquete, 2003]

And a typical result of these systems also from TERSEO is shown in the following table:

	SPANISH		ENGLISH
	TRAINING	TEST	TEST
No Art.	50	50	100
Real Ref	238	199	634
Treated Ref.	201	156	511
Successes	170	138	393
Precision	84%	88%	77%
Recall	71%	69%	62%
F-Measure	77%	77%	68%

Figure 2 Table of results of TERSEO [Saquete, 2003]

2. Machine learning approach

Machine learning approaches has the advantage is easier to deal with unknown cases and error reduction and noisy data control. The system has to be trained with a annotated corpus, which may be difficult to implement and is a clear disadvantage against the rule based approaches.

Machine learning algorithms can be easily tuned by changing the set and weights of features used for the extraction.

A example of this approach is ManTIME [Filannino et al., 2013] which relies on conditional random fields (CRF) a machine learning algorithm for TE extraction and normalization. The results taken from [Filannino et al., 2013] are:

#	Training data	Identification						Normalization		Overall
		Strict matching			Lenient matching			Accuracy		
run	(post-processing)	Pre.	Rec.	$F_{\beta=1}$	Pre.	Rec.	$F_{\beta=1}$	Type	Value	score
1	Human&Silver (no)	78.57	63.77	70.40	97.32	78.99	87.20	88.99	77.06	67.20
2	Human&Silver (yes)	79.82	65.94	72.22	97.37	80.43	88.10	87.38	75.68	66.67
3	Human (no)	76.07	64.49	69.80	94.87	80.43	87.06	87.39	77.48	67.45
4	Human (yes)	78.86	**70.29**	**74.33**	95.12	**84.78**	**89.66**	86.31	76.92	**68.97**
5	Silver (no)	77.68	63.04	69.60	97.32	78.99	87.20	88.99	77.06	67.20
6	Silver (yes)	**81.98**	65.94	73.09	**98.20**	78.99	87.55	**90.83**	**77.98**	68.27

Table 1: Performance on the TempEval-3 test set.

Figure 3 Performance of Filaninos's in TempEval-3 set

Another example of a machine learning approach is the one of Ahn [Ahn, Fissaha-Adafre., 2005] in which they use support vector machines for their own extraction and normalization. The results they reported and taken from [Ahn, Fissaha-Adafre., 2005] are:

System	corrVAL	corrNOVAL	actTIMEX2	P	R	F	absP	absR	absF
BL	859	32	1245	0.813	0.787	0.800	0.716	0.624	0.667
LLL	931	33	1245	0.882	0.853	0.867	0.774	0.676	0.722
LLP	938	33	1245	0.912	0.859	0.885	0.780	0.680	0.727
LPL	951	39	1245	0.916	0.871	0.893	0.795	0.694	0.741
LPP	987	39	1245	0.951	0.904	0.927	0.824	0.719	0.768
PLL	1008	63	1287	0.886	0.828	0.856	0.832	0.751	0.789
PPP	1097	70	1287	0.966	0.901	0.932	0.907	0.818	0.860
LLL	1285	33	1601	0.910	0.887	0.899	0.823	0.721	0.769
PLL	1362	63	1643	0.912	0.866	0.888	0.867	0.780	0.821
ITC-IRST	1365	35	1648	0.875	0.870	0.872	0.850	0.766	0.806

Table 6: Performance on VAL. (Top): TEXT-only. (Bottom): full document.

Figure 4 Performance of Ahn et al.

As it is shown, machine learning is on the vanguard as the approach that most likely is, and will continue to be main stream in this research, however given some needs it is possible to implement a grammar and or an automata for good results especially if we are only interested in simple temporal expressions.

2.2.3 Discussion on temporal expressions in the context of this work

First, is worth mentioning that after an analysis some types of temporal expressions won't be as decisive for similarity measures given their features. i.e. "*This has been happening since the nineteenth century*" against "*Monday, 21 June 2013 3 pm the rocket was launched...*". These cases would have to be studied to rule out the temporal expressions that won't give a good set of features for the document's temporal profile.

Considering that some TE would not be as relevant and the analysis will show that expressions which have a big granularity, or are more ambiguous, or are less clearly defined have a much lesser role on the document temporal profiling. For this reason it is necessary to provide a method to measure the similarity that takes into account all granularities.

On the other hand is important to state that for automatic annotation machine learning approaches show promising results and are more practical approach to large data sets. Nevertheless rule based approach might prove to be easy to implement for not so specialized needs. The corpus manual annotation of temporal expressions and their normalization will provide a solid starting point for development system that uses either or both of the approaches.

2.3 Topic detection and event tracking

Topic detection and tracking (TDT) according to [Allan, 2002] is a body of research and an evaluation paradigm that focuses on event-based news organization. TDT was originally funded and supported by the Defence Advanced Research Projects Agency (DARPA).

2.3.1 Overview on topic detection and event tracking

First is necessary to make a clear definition of event and topic.

These definitions taken from [Allan et al., 1998] reflect what a topic and an event is.

1. A topic is a seminal event or activity along with all directly related events and activities.

2. An event is something that happens in a specific time and place (specific elections, accidents, crimes and natural disasters are examples of events).

3. An activity is a connected set of actions that have a common focus or purpose (specific campaigns, investigations, and disaster relief efforts are examples of activities).

The mains task in topic detection is to measure the similarity of documents to see if they are similar enough to be considered of the same topic. Usually this is done with machine learning algorithms.

In the case of event tracking more features and consideration have be dealt of, in this case a simple term vector similarity is not reliable for event tracking, and more complex sets of features have to be determined.

There are certain features that are present in an event which are better described in the next questions (taken from [Masnizah, 2010]):

Features	Description
Who	The actor or person that takes part in an event.
Where	The location is where the event takes place.
When	The date or tome it took place.
What	The subject, occasion or activity involved the event.

Table 5 Main features of an event

It's clear that these questions define the events and are basically solved with named entities: person names, locations, temporal expressions, etc. Probably the hardest question to answer from a NLP task point of view is "What", because further analysis is

needed of the text for resolving it. These features are used for the machine learning algorithm in the learning process.

A point of view is that events could be identified relying mostly in the temporal information and the creation date, however, tracking events in documents in a given topic cluster, is an achievable objective with the aid of temporal expressions but other features would have to be used to achieve this goal.

According to [Masnizah, 2010]. *"Topic Detection is the task of identifying if the news are similar (on-topic) and dissimilar (off-topic) news stories in the news stream."* It is subdivided in 3 categories (also from [Masnizah, 2010]):

1. New Event Detection

   Is the task of recognizing seminal events as they arrive on the data stream.

2. Cluster Detection

   The task is to divide the processed events and cluster them in topics, It's important that, in Masnizah's work, this only alludes to events and not the documents, a similar cluster of documents would be another task but very similar.

3. Story Link

   This is the core task of topic detection, to say if a new document belongs to a topic or not.

Tracking is the task of finding new documents that belong to an event this uses events that are already known by the system and links them to the cluster.

2.3.2 Feature selection

Perhaps the most important thing in event tracking is the proper selection of features. Features have to be representative of the document, and the useless features removed.

Important features of text documents usually are keywords, name entities, and such which are strings; however usually machine learning algorithms can only work with numbers or nominal values, because of that is necessary to convert such string features to numbers or nominal values.

In the case for nominal values in words it is only if it contains the word or not (true or false), for number values is necessary to convert the string into a number usually during the preprocessing the frequency or the Tf-Idf is extracted and used as a value representative of the importance of the word in the text [Manning et al., 2008].

The Tf-Idf measure short for term frequency–inverse document frequency, is a numerical statistic that intends to reflect how important a word is to a document in a collection of documents. The following figure from [Manning et al., 2008] states the possible variants for the Tf-Idf measure:

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5+\frac{0.5\times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$ (Section 6.4.4)
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\ 0 & \text{otherwise}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

► **Figure 6.7** SMART notation for tf-idf variants. Here *CharLength* is the number of characters in the document.

Figure 5 Variants of the Tf-Idf maeasure

Other features such as the news pieces source, the published date, and another statistics might also be used.

Furthermore also features have to be condensed in a matrix to be fed to the classfier either for training or to classify.

.

2.3.3 Common approaches

Most of the work done in the field has used unsupervised learning to classify the documents.

The approach entirely depends if for example the topic categories are known or have to be discovered by the machine (classification vs clusterization) problems, in which different ML algorithms have to be used.

Practically speaking the topic and event clustering are the same mathematical problem (clusterization) the different would rely on the set of weights and features used, and the interpretation of the clusters. Also the similarity measure varies in each other.

Interesting work done in this field:

- [Song et al. 2011] propose a system that generates a tree using documents as leaves and clusters as nodes to identify closed related events and topics, it has very good performance and it is widely described in their paper. A diagram of a topic-event tree taken from [Song et al. 2011] is shown:

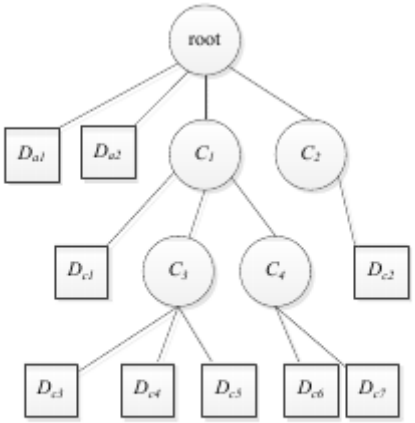


Figure 1. A sample of *MI-Tree*

Figure 6 A topic-event tree from Song's

And some results of Song:

Number of documents	accuracy	
	System-1	*System-2*
100	75%	80%
500	83.42%	87.16%
2000	89.46%	93.46%

Table II shows the CF-Feature of System-2 is large

Figure 7 Results from Song's Work

- Makkonen, in her works, uses the temporal information [Makkonen, 2004], locations and spatial information [Makkonen et al., 2003], and semantic classes of named entities. In [Makkonen et al., 2002] they provide a serious analysis of temporal expressions and their relation with topic detection, and explore the use of spatial information, which is also of considerable importance in this task. Some results from Makonnen:

Table 6. The results of detection and tracking

method	Detection			Tracking			$\frac{F1_D+F1_T}{2}$
	P	R	$F1_D$	P	R	$F1_T$	
Cosine	0.473	0.237	0.315	0.214	0.766	0.334	0.325
Cosine (SC)	0.531	0.294	0.379	0.286	0.500	0.363	0.371
Skew Divergence	0.400	0.190	0.258	0.207	0.545	0.300	0.279
Heuristic	0.551	0.905	0.685	0.688	0.450	0.544	0.620

Figure 8 Makonnen 's results

- [Kumaran, Allan, 2002] uses the baseline approach with cosine similarity of TF-IDF clustering named entities. They also made an analysis and found situations in which the NE does not perform as well as expected. They also keep in mind the "oldness" or "newness" of the articles.

- [Montalvo et al. 2007] Use fuzzy similarity, and a human knowledge ontology to improve the named entities accuracy. They worked with multilingual corpus using translated Named Entities in their algorithm.

- [Masnizah, 2010] analyze existing systems for interactive topic detection and tracking systems that retakes input from the users to help the system get feedback of its results and learn from it, besides from presenting a user interface of the found topics.

These are some examples of the approaches used by different authors on topic detection and event tracking.

2.3.4 Temporal information and its relation with event tracking

Temporal information conveyed in TE is an interesting feature for event tracking for several reasons:

- If the documents contain a date that temporal information can be anchored, it gives a good start line for temporal similarity of documents.
- Documents produced temporally closer to each other are more likely to talk about the same event that a news farther away in time.
- Temporal similarity can be applied to the documents and fix them in a time frame that might help in future researches and approaches of event tracking.

The temporal information should be approached carefully and exploiting all the advantages it gives to event tracking.

2.3.5 Discussion of topic detection and event tracking

The topic detecting and tracking tasks are widely developed, but only Makkonen addresses this problem with temporal information in dept which will be critical for this work, also the algorithm of Makkonen in all her works provides a realistic starting point for implementing a similar system.

Song's work [Song et al. 2011] is, at first glimpse, very similar of what the ideal system intends to be, the fact that a tree is formed with documents as leaves and clusters as branches, provides a very friendly user interface and, more importantly, a much quicker approach for working with new documents to be clustered.

This works (Makonnen's and Song's) appear to be the main foundations of this work, also Maznizah [Masnizah, 2010] does a research about user interaction with this kind of systems both giving a guide line for user interfaces and how the user might help training or providing knowledge to the system of its own performance.

The bottom line of topic detection in the context of this work:

- Take Makonen's works as a guideline for using temporal similarity in the topic and event detection task.
- Try to perform a tree like cluster as described in Song's work and adapt it for the features coming from the temporal similarity measure proposed and the others features taken from the documents such as term vector with TF-IDF, named entities, etc.

2.4 Temporal similarity

Temporal similarity is a field of study in which a mathematical function is defined to measure the temporal distance of two documents given their temporal information, for this reason having the temporal expressions normalized for similarity measurement is very important.

2.4.1 Overview on temporal similarity

A temporal similarity distance is no more than determining a number which relates the similarity of the temporal expressions sets in two given documents. This is analogous as measuring the similarity of two documents by its term vectors but with temporal information instead.

This field is not widely used as a main research line in NLP or IR, for two reasons, first it is relative new and unused field, and second in the cases in event tracking in which temporal similarity is needed a standard simple function with a timeline is defined or well the authors does not describe extensively how they used the temporal information on their algorithms.

The problem on temporal similarity can be primarily treated as a subset problem in which temporal information of two documents is mapped into a two dimension timeline in which each axis is the time in each document, then the overlapping area is measured and from there a similarity measure can be obtained. This with other important considerations is the core of the problem.

This being said it is to notice that it was particularly difficult to find papers referring to temporal similarity which is one main motivation for working with temporal similarity in the future.

2.4.2 Common approaches

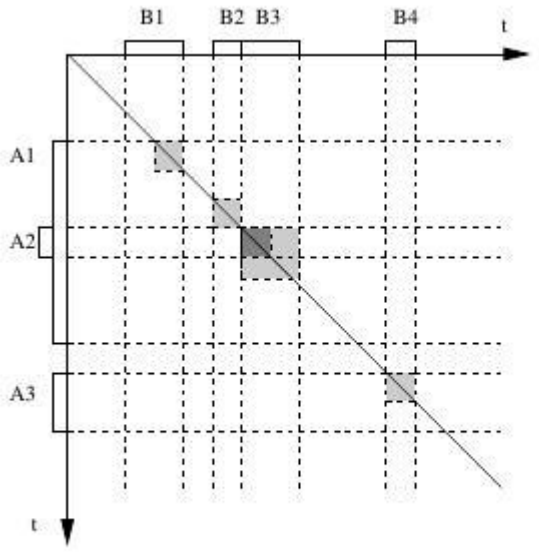
Makkonen addresses the problem in [Makkonen, 2004], taking two sets of temporal expressions which are mapped into a timeline, the vagueness is measured by relevance distributions such as uniform distribution, Gaussian distribution, Exponential distribution, Reversed Exponential distribution, etc.

In [Makkonen et al., 2003], they also use temporal similarity for topic detection; they proposed a cross-tabulation of the temporal intervals of the documents to see the overlapping area of the ranges. Also they work considering that there are seven possible relations between two ranges (equals, meets, overlaps, begins, falls within, etc.). Represented in the following figure taken from [Makkonen et al., 2003, pp. 7] :

$[t_i, t_j]$ **is before** $[t_k, t_l]$	if $t_j < t_k$
$[t_i, t_j]$ **meets** $[t_k, t_l]$	if $t_j = t_k$
$[t_i, t_j]$ **overlaps** $[t_k, t_l]$	if $t_i < t_k < t_j < t_l$
$[t_i, t_j]$ **begins** $[t_k, t_l]$	if $t_i = t_k \wedge t_j < t_l$
$[t_i, t_j]$ **falls within** $[t_k, t_l]$	if $t_i < t_k \wedge t_j < t_l$
$[t_i, t_j]$ **finishes** $[t_k, t_l]$	if $t_i < t_k \wedge t_j = t_l$
$[t_i, t_j]$ **equals** $[t_k, t_l]$	if $t_i = t_k \wedge t_j = t_l$

Figure 9 Posible relations between time ranges

And a figure which represents the graphical timeline mapping, also taken from [Makkonen et al., 2003, pp. 5] is:



A cross-tabulation of two sets of intervals A and B.

Figure 10 Tabulation of two sets of time intervals

After that, they confront the results with a weight function that considers all the seven relations between ranges, and get a value which is the sum of all areas after being weighted returning a similarity measure between zero and one.

They also work with locations using a spatial taxonomy and trying to get the distance between the places if they are not the same.

2.4.3 Discussion about temporal similarity applied in this work

Makkonen gives an approximation of what this work is intended to be. She did not use other relevancy measures that are not directly related to the timeline itself like durations or frequencies.

Another line of thought is if a temporal expression is closer to a more relevant Named Entity, or how far in time are the temporal expressions are from the document published date, etc.

Something to be considered is that in topic tracking, maybe very large topic clusters might have widely distributed temporal expressions, and measuring the temporal profile of each document against a new document is costly from the computer point of view. This should be controlled and expected at testing time.

2.5 Lazy machine learning

Lazy learning is a machine learning approach that does not stores training data or simple does some minor processing and waits until data is received to classify in the other its counterpart the eager learning given a training sets constructs a classification model before receiving data to classify [Bontempi et al., 1999].

Generally the lazy learning algorithms take less time in training but more time in predicting than their counterparts the eager learning algorithms. Also the lazy learning algorithms use a richer hypothesis space that covers the entire instance space and their counterparts must commit to a single hypothesis.

As it was said the lazy learning algorithms stores the training data until a new instance is going to be classified [Atkenson et al., 1997].

Two typical approaches of lazy learning are:

- K-nearest neighbors
  Instances are represented as points in a Euclidean feature space.

- Locally weighted regression
  A local approximation is constructed.

More over k-nearest neighbor is an algorithm that can be configured with different sets of options.

2.5.1 K-nearest neighbors algorithm

In this lazy learning algorithm the instances to be classified are mapped trough their features into a Euclidean vector space with *n-dimensions* in which *n* is the number of features and each feature is one dimension in the space.

For classifying instances, the algorithm maps it into a vector in the feature spaces and simply calculates its Euclidean distance against all the training instances points and assigns the class label to the k-nearest neighbors of that point thus the name of the algorithm.

For a value of $K=1$ the class is assigned to the closest neighbor, and rules have to be done to decide how to select the label for the instances if the value of K is bigger than one. This can be done with the mean distance or a weighting function. The following figure[12] exemplifies it:

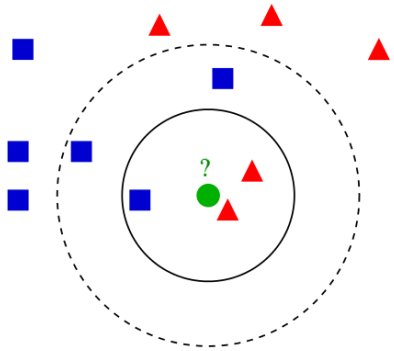


Figure 11 K-nearest neighbors class label decision

The system is robust to noise data averaging the k-nearest neighbors; nevertheless a big problem is the multidimensionality of the space commonly seen in information retrieval problems and meaningless features or less relevant attributes should be ruled out as features.

[12] Taken from Wikipedia: http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

2.5.2 Discussion on machine learning algorithms

In a real case scenario where oncoming news pieces are coming everyday and the training has to be updated regularly to considered new cases a eagerly learning algorithm seems like the most reliable tool, as in it can be trained once every new events are created and each oncoming news in between would be easily classified.

However in this case the intention of the work is to test the temporal similarity and its availability as features to topic tracking; thus the implementation and time consuming training of this kind f algorithms makes them less reliable on perform experiments (given that they take too long to perform many experiments).

For that reason it was decided that even though multidimensionality is a big problem in k-nearest neighbors it may proof be a good platform to perform the experiments given that its training and testing time and its performance are acceptable.

Furthermore there is more documentation on k-nearest neighbors algorithm online using the selected machine learning tool used Weka[13]; mainly because is far more used for these kind of problems therefore easier to implement the experiments,

[13] Weka3: Data Mining Software. http://www.cs.waikato.ac.nz/ml/weka/

2.6 Evaluation metrics

In this project two kind of evaluation metrics were considered, first the precision, recall and F-Measure and secondly the Cohen's kappa statistical measure for inter annotator agreement.

It was decided for this work to rate the event tracking with precision, recall and F-Measure, also the quality of annotation of named entities and temporal expressions was measured with this metrics.

In the other hand the topic classification and the event classification was measured using the inter annotator agreement to measure how well humans perform in this task to be able to compare against the machine learning algorithm results.

2.6.1 Precision, recall and F-measure

In information retrieval evaluation usually a statistic used to measure the results is the precision and the recall. The precision is the fraction of retrieved instances that are relevant and the recall is the fraction of relevant instances that are retrieved.

This can be graphically explained on the next figure[14]:

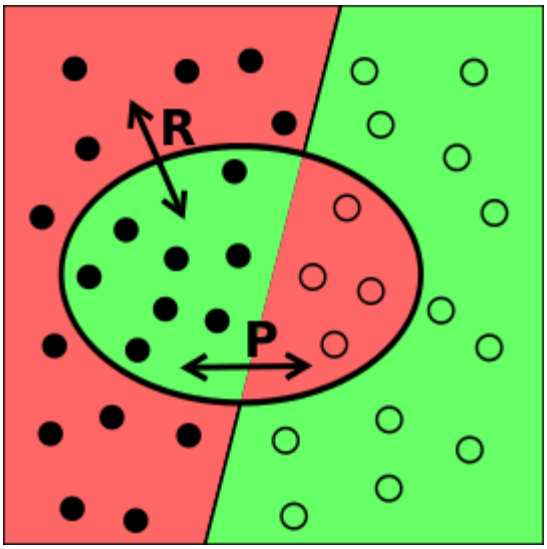


Figure 12 Graphical representation of precision and recall.

[14] Image taken from Wikipedia. The image is public domain and was taken from: http://en.wikipedia.org/wiki/Precision_and_recall

The next formulas expres these concepts:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Given the figure there are four possible outcomes from a classifying problem in a given class [Powers, 2007]; which are represented in the table below:

	Condition positive	**Condition negative**
Test is positive	Correct classification True positive	Type I error False positive
Test is negative	Type II error False negative	Correct classification True negative

Table 6 Possible outcomes of a classifying problem

As can been seen from the table there are two types of correct answers and two types of errors. Type I the rejection of a null hypothesis that is actually true. Type II errors refer when a null hypothesis that is false is not rejected.

Given the table above the Precision and recall can be calculated with the next formulas:

$$\text{Recall} = \frac{tp}{tp + fn} \qquad \text{Precision} = \frac{tp}{tp + fp}$$

The F-measure is the harmonic average of both precision and recall. The F-measure can be taken as how good perform a system in a test. The formula for calculating the F-measure is:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

2.6.2 Inter-annotator agreement

The inter-annotator agreement is the degree of agreement between annotator in a given test. In information retrieval in particular in annotation, classification and other common information retrieval tasks, the inter-annotator agreement is used to qualify the input data and also to provide a ceiling of the proficiency of the system because it is implied that no system can perform better than human counterparts.

One statistical measure to rate the inter-annotator agreement is Cohen's kappa. It is a statistical measure for qualitative categorical items, this is classification problems, it is a robust statistic because it takes into account the agreement occurring by chance [Smeeton, 1985].

The formula for the Cohen's kappa is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

The kappa statistic is calculated given the relative agreement between annotators Pr(a) and calculating the probability of chance selecting that class Pr(e). Kappa is a number between cero and one, one means the annotators are in perfect agreement and cero in total disagreement.

The following figure explains the interpretation of the inter-annotator agreement [Viera, 2005]:

Interpretation of Kappa

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Figure 13 Interpretation of the Kappa measure

2.6.3 Discussion of the evaluation metrics

Looking into the need of evaluation for this project, the metrics selected namely, precision, recall and F-measure, and Cohen's kappa, looks representative in the evaluation of the system.

There is to mention that there is the need to measure with kappa the ceiling proficiency of the system which would represent the best the system might do. And a bottom line which in this case is the baseline proposed.

Furthermore kappa will give us an idea of the complexity of classifying the news pieces into events.

It is common in literature as seen above to use precision, recall and F-measure to rate how good a system performs, especially in event tracking or topic detection.

For those reasons these evaluation metrics were selected.

2.7 Overall discussion of the literature review

For this work, Song's work seems to be the most oriented on what it is ideally intended, following his methodology and giving special attention to temporal similarity; however the implementation of such complex system requires a great amount of time therefore a more feasible approach was selected.

The chosen approach is k-nearest neighbor classifier to track the topics and train it with temporal features; this would provide a starting platform for temporal similarity analysis and event tracking experiments.

Named entity extraction is already provided by some tools, dictionaries or ontologies and their normalization and filtering should be made or at least considered, however the problem that there are none tools available makes a fact that named entities are required

Temporal expressions extraction is a bigger problem; the only available tools found are from Alicante University. Contact has been made with Dr. Saquete and access to one (or both) tools was requested however no answer was received. Therefore the corpus was manually annotated looking for better results and to have material to work in the experiments.

Temporal similarity measure will be heavily based on Makkonen's works as her works were the only found. She takes mathematical properties of ranges and considers them for time range similarity measuring also she uses several distributions for ambiguity treatment which is really interesting because this could be applied in others dimensions until a better model is found.

Empirical research needs to be done for improving Makkonen's temporal similarity Also testing and weighting on the diverse features (temporal or not) should be considered.

In conclusion research has been done in the area but temporal similarity lacks works about it, and it is considered to be an important feature of documents especially of news. Improving Makkonen's algorithm should be a priority and analyzing how she works and implementing it to the corpus temporal expressions.

3. Methodology

The method proposed for this works includes the gathering of corpus, the development of the temporal similarity algorithms and the construction of the clustering system and the evaluation of the results with a set of experiments.

First the following a corpus was gatherer from RSS feeds of online newspapers and their correspondent online news pieces, this was achieved with RSS recovering tools and text cleaning tools.

The corpus was then cleaned and distribute along with a set of guidelines to annotators for its named entity and temporal expressions annotation and classification. Then the temporal expressions were normalized furthermore a classification of the news pieces in topics and then in events was carried out.

A set of temporal similarity algorithms were programmed including a baseline, Makkonen's algorithm, and two new proposals in which one frequencies and durations types of named entities are considered.

Then a set of features were extracted and implemented a k-nearest neighbor Weka classifier in Java, using the sets of features per topic. Finally a set of experiments considering the different temporal similarity algorithms and the other non temporal features were done.

Finally evaluation of the annotation, classification in topics and events and the performance of the classifier were done, using Cohen's kappa or, precision, recall and F-measure.

The following figure explains the overall process of the methodology:

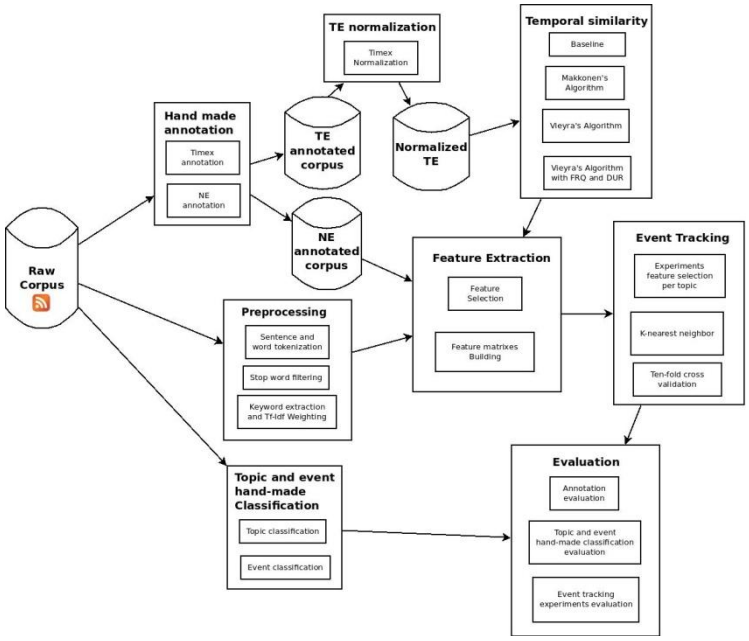


Figure 14 Methodology followed during this project

3.1. Corpus

The corpus was made of news pieces that were fed by a Java program. The texts consist mainly in two parts; first metadata recollected in the process and second the text itself of the news.

The corpus was gathered from news of different newspaper from 13th of October 2013 to 28 of October 2013.

The news pieces were selected mostly from "Politics" section nevertheless some news pieces of totally distinct topics were left for evaluation purposes.

It is needed to tag the corpus with some basic information that will help during the preprocessing and evaluation of reliable normalized temporal expression, the work proposed over the corpus is:

- Annotation of Named Entities (locations and names)
- Annotation of temporal expressions.
- Metadata collection.
- Loose topic manual classification.

The news then would be preprocessed this is tokenized, stemmed, normalized using dictionaries of common synonyms, etc.

3.1.1 Corpus gathering

The process of obtaining a corpus which is suitable for the needs of this work is described ahead and contains two basic steps, first the corpus gathering itself and then a shallow filtering of event related news.

The gathering uses a set of tools for the Java programming language. The tools used are:

1. ROME:[15] A Java library used to extract RSS Feeds and the metadata related to them.
2. Tika: [16] Another Java tool which amongst other things is able to extract webpages, their metadata and the text related to them.
3. Boilerpipe[17]: A tool that works with heuristics and recognize the content filled text between all the text that comes from a html text.

[15] ROME: https://rometools.jira.com/wiki/display/ROME/Home
[16] Tika: http://tika.apache.org/

[17] Boilerpipe: http://code.google.com/p/boilerpipe/

The gathering of the corpus used RSS feeds from 4 distinct newspapers online and their sections where the event news might be. The sources are:

- La Jornada[18]: A well known left wing paper.
- Excelsior[19]: A well known right sided paper.
- El Universal[20]. One of the most read papers in Mexico.
- Proceso[21]. A news magazine specialized in politics.

And the sections used and the URL's:

Newspaper	Section	URL
La Jornada	Politics	http://www.jornada.unam.mx/2013/10/15/politica
La Jornada	Society and Justice	http://www.jornada.unam.mx/2013/10/15/sociedad
La Jornada	States	http://www.jornada.unam.mx/2013/10/15/estados
La Jornada	Capital	http://www.jornada.unam.mx/2013/10/15/capital
Excelsior	All (only one RSS feed)	http://www.excelsior.com.mx/
El Universal	All (only one RSS feed)	http://www.eluniversal.com.mx/noticias.html
Proceso	All (only one RSS feed)	http://www.proceso.com.mx/

Table 7 Newspaper sections used for corpus gathering and their URL

The method consisted on some straight forward steps to collect all the news and metadata from the desired news papers:

1. ROME was fed with the RSS URL's, ROME extract each individual RSS feed and extract all the metadata. In this metadata there is the actual URL of the news piece in the newspaper website.
2. Then this second URL is fed to TIKA, which crawls the URL extracting all the metadata plus the complete HTML code of the news piece.
3. After collecting the relevant metadata the brute HTML text is consumed by Boilerpipe, this cleans all non-relevant text from the HTML (comments, indexes, navigation text, etc.) returning on the main text block of the HTML code.
4. Automatic cleaning is done to each of the news papers, this is done to remove text that was in the main text content of the webpage, mainly the text removed is non-related with the news piece for example the name of the newspaper, the reporters names, the pictures tittles, the texts *"Lo más leido"* (the most read), *"comentar"* (comment), and lists of links to related news.
5. All this information (metadata and text) is joined into a java object and printed into files, this files can be parsed to recover the java objects for the actual processing of the corpus.
6. Manual filtering was done to discard pieces of news that were comments, overviews, reviews and non-topic related pieces of news.
7. A new Corpus file was done with the clean news pieces in a readable format for annotation purposes.

[18] Periódico La Jornada: http://www.jornada.unam.mx/
[19] Periódico El Excelsior: http://www.excelsior.com.mx/
[20] Periódico El Universal: http://www.eluniversal.com.mx/noticias.html
[21] Semanario Proceso: http://www.proceso.com.mx/

The process of how corpus is collected is described here:

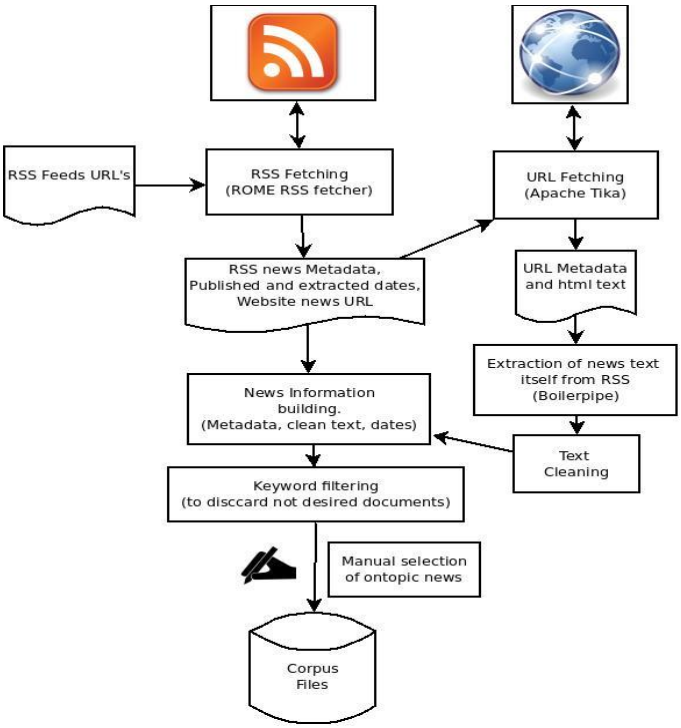


Figure 15 Corpus gathering and filtering process

A daily harvest of RSS feeds was done during the given period of time (13-10-2013 to 28-10-2013) with exceptions of some days that internet was not available. A total of 552 news pieces were recovered as shown in the table:

Date	Number of news	Date	Number of news
13-10-2013	7	21-10-2013	34
14-10-2013	19	22-10-2013	53
15-10-2013	96	23-10-2013	27
16-10-2013	55	24-10-2013	61
17-10-2013	40	25-10-2013	34
18-10-2013	26	26-10-2013	2
19-10-2013	11	27-10-2013	8
20-10-2013	48	28-10-2013	31
Total 552			

Table 8 Number of news pieces per date

For the days 26th and 27th the RSS harvest was not done because of internet failure however some of the news collected on the 28th belonged to the 26th and 27th and that's why these days have news pieces.

Also the news pieces recovered by source are shown as follows:

Source Newspaper	News pieces
Proceso	176
Excelsior	59
Jornada	273
Universal	44
Total	**552**

Table 9 Number of news pieces by source

And the table containing the news pieces by source and by date is:

Source	Date	News pieces	Source	Date	News pieces
Proceso	13-10-2013	7	Excelsior	13-10-2013	0
	14-10-2013	15		14-10-2013	0
	15-10-2013	21		15-10-2013	7
	16-10-2013	8		16-10-2013	8
	17-10-2013	9		17-10-2013	0
	18-10-2013	16		18-10-2013	5
	19-10-2013	11		19-10-2013	0
	20-10-2013	8		20-10-2013	3
	21-10-2013	9		21-10-2013	3
	22-10-2013	12		22-10-2013	6
	23-10-2013	18		23-10-2013	8
	24-10-2013	21		24-10-2013	8
	25-10-2013	5		25-10-2013	6
	26-10-2013	2		26-10-2013	0
	27-10-2013	8		27-10-2013	0
	28-10-2013	6		28-10-2013	5
Total Proceso	**176**		Total Excelsior	**59**	
Jornada	13-10-2013	0	Universal	13-10-2013	0
	14-10-2013	0		14-10-2013	4
	15-10-2013	66		15-10-2013	2
	16-10-2013	35		16-10-2013	4
	17-10-2013	28		17-10-2013	3
	18-10-2013	0		18-10-2013	5
	19-10-2013	0		19-10-2013	0
	20-10-2013	32		20-10-2013	5
	21-10-2013	18		21-10-2013	4
	22-10-2013	31		22-10-2013	4
	23-10-2013	0		23-10-2013	1
	24-10-2013	26		24-10-2013	6
	25-10-2013	21		25-10-2013	2
	26-10-2013	0		26-10-2013	0
	27-10-2013	0		27-10-2013	0
	28-10-2013	16		28-10-2013	4
Total Jornada	**273**		Total Universal	**44**	
Total			**552**		

Table 10 News pieces by source and by date

The token count and relevant statistics on the full Corpus are shown further in this work.

3.1.2 Corpus annotation

The corpus needed to be annotated with several entities required by the system. It was decided to do some basic preprocessing and make a readable format text easy to tag by human taggers which only contains the id, the title, the source file and the text itself.

The entities needed to be tagged were basically named entities, and temporal expressions.

There were three major procedures done for tagging and normalizing the corpus, these are described ahead:

1. Named entities tagging
2. Temporal expressions tagging
3. Temporal expressions normalization

3.1.2.1 Quality control

For controlling the quality of the human taggers, fifty texts were tagged carefully and revised by two different annotators, to be used as the golden standard both for named entities and temporal expressions.

Packets of fifty non tagged texts were done, in each of the packets five texts of the golden standard were inserted without the human annotators knowing it.

After when the annotator delivered the results, an automatic evaluation of the annotator's proficiency is done against the golden standard.

The golden standard consisted on 50 random texts taken out from the corpus, and only could be identified by the ids in the file the annotators got.

The results of the quality control are shown separately for named entities and temporal expressions further in this work.

The annotation was done by three students which are studding Language and Literature Course in México's National University (UNAM), a questionnaire was given to them for a profile of the annotators and the answer were:

Annotator	Is Spanish her Mother Language	Age	Level of Studies	Is she Mexican?	Is she familiarized with the topic of the news?	Does she know the newspapers of the news?
A	Yes	26	Language Bachelor	Yes	Yes	Yes
B	Yes	26	Language Bachelor	Yes	No	Yes
C	Yes	27	Language Bachelor	Yes	Yes	All but Excelsior

Table 11 Annotator's profile data

As is shown above all the annotator are Mexican and are somehow familiarized with the topics of the news, also they all know basic annotation theory and have experience on it. They did all the temporal expressions and named entities annotation.

In the case of the temporal expression normalization, two annotators used the normalization tool in the golden standard's temporal expressions and their results compared. The annotators were the author of this project and the other Prof. Baptista one of the supervisors.

Furthermore the classification of the events and topics was done by two different annotators and was measured with inter-annotator agreement kappa.

For the evaluation of the delimitation of named entities and temporal expression precision, recall and F-measure was used, also for their classification. In the case of the measure of the temporal expression normalization the Cohen's kappa was used for the delimitation and precision, recall and F-measure for the open-boundary classification.

The results of the annotation quality evaluation are shown in **§4.1 Annotation evaluation**.

3.1.2.2 Named entities annotation

The tagging consisted on human taggers wrapping between brackets the named entities and putting a single letter after which describes the type of named entity i.e.:

El ministro en retiro {Genaro Góngora Pimentel}P fue exonerado por sus pares del {DF}L en el caso contra la madre de sus hijos.

> (The minister in retirement {Genaro Gongora Pimentel}P was exonerated by his collegues en the {DF}L (Federal district) in the case against the mother of his children).

In this case there are two named entities with different types tagged here first is a person name (*Genaro Góngora Pimentel*) and a Place name (*DF*).

Name entities were tagged by six kinds of entities, this were selected thinking first in the usual categories used in literature and also taking into account the events reflected on the corpus. The categories of named entities and the rules to annotate them are:

1. Person names (*P*)
   The names of persons were tagged. A set of rules were defined only for proper names and examples were given to the taggers, the letter to identificate the names in the corpus is *P*. Some examples of tagged names are:
   The rules for the taggers were:
   i. Full names, short names (not nicknames), last names and first names of persons only were considered. I.E. these names are from the same person and these cases are common on the text:
      - *José Luis Perez Rodriguez*
      - *Perez Rodriguez*
      - *Luis Perez*
      - *Perez*

   ii. Common abbreviations or acronyms of persons names were also tagged I.E.
      - *EPN* as in *Enrique Peña Nieto*
      - *AMLO* as in *Andres Manuel Lopez Obrador*

   iii. Names of persons given to others that are not proper a person were not tagged I.E.
      - *Escuela Justo Sierra* (Justo Sierra School)

   iv. If the person name was given full and next to it its acronym were tagged as one entity. I.E.
      - *Andres Manuel Lopez Obrador (AMLO)*

   v. If a person is referred by another means or by anaphora it was not annotated. I.E.
      - *{El presidente}P de la Republica* ({The president} of the Republic) is **Incorrect**
      - *{Él}P dijo que no lo haría* ({He}P said he would not do it) is **Incorrect**

   vi. If the person was accompanied also by his title, only the person name is marked. I.E.
      - *El presidente de la Republica, {Enrique Peña Nieto}P* (The president of the Republic, {Enrique Peña Nieto{P})

Some correctly tagged examples of valid person names are:

- *{Édgar Elías Azar}P* , *{Elías Azar}P* , *{José Antonio Meade}P* , *{Meade}P* , *{EPN}P. {AMLO}P*

2. Natural disasters names (*D*)
   Most if not all of these cases are specific to hurricanes, tropical storms and natural phenomena names, mostly because of two reasons, first during the gathering of the corpus there were several hurricanes and tropical storms of importance over México and second amongst natural disasters only hurricanes, typhoons, tropical and polar storms, and long term natural phonomena are named. These names were tagged also wrapped in brackets and identified with the letter *D*.

   In this case no rules for the taggers were needed but only to pay attention to the text because the names of hurricanes are usually first names of persons; and that if the name is accompanied by "hurricane" or "typhoon" to tag it to. Some correctly tagged disaster names are:

   - *{Ingrid}D impacto hoy a {Guerrero}L* ({Ingrid}D hit {Guerrero}L today)
   - *El {huracán Manuel}D* (The {hurricane Manuel}P)

3. Institutions names (I)
   Also names of non-human entities were tagged, because to say Institution is a very vague term, a set of rules had to be defined to guide the taggers, the Institutions names were tagged with the letter *I*.

   The rules are:
   i. Full names, acronyms, synonyms (if they are names) were tagged, in many cases institutions are mentioned by their full name only once in the text and then the acronyms or abbreviation are used. I.E
      - *{Camara de Senadores}I* (Senate)
      - *Segob* short for *Secretaria de Gobernación* (govern secretariat)
      - *FIL* as in *Feria internacional del libro* (International book fair)

   ii. Political Entities which are also names of places I.E Mexico when referred as the government of Mexico and not the country itself were tagged as places nevertheless.
   iii. Entities which are referred by its full name and its acronym were marked as only one institution.
   iv. Common names that refer to institutions should be tagged only if they play a significant role in the sentence and in the news piece were tagged. I.E.
      - *El {gobierno federal}I y los {gobiernos estatales}I deberán demostrarle a la sociedad[…]* (The {federal government}I and the {States governments}I will have to show society […]) is **Correct**

   v. Anaphoras should not be tagged. I.E.
      - *La {empresa}I elaboró[…]* (The {company}I maid) is **Incorrect**

Some correctly tagged institutions names are:

- *{Instituto Nacional de Bellas Artes}I* ({National Institute of Fine Arts}I), *{INBA}I*, *{Secretaría de Gobernación (Segob)}I* ({Secretariat of govern, Segob}I).

4. Places Names (*L*)
   The names of places also presents a challenge for the taggers, they were tagged with the letter *L*.

   The rules defined are:
   i. Only names of political geographic entities were tagged, this includes counties, neighborhoods, states, countries geopolitical regions of the world, geographic zones, etc. I.E.
      - *{CIUDAD DE MÉXICO}L* ({Mexico city}L)
      - *{México}L*
      - *{Latinoamérica}L* ({Latin-America}L)
   ii. Names of places describing a geographic landmark were not tagged. I.E.
      - *{Río Bravo}L* ({Bravo River}L) is **Incorrect**
   iii. Also abbreviations, acronyms and synonyms were tagged. I.E.
      - *{S.L.P}L* as in San Luis Potosí
   iv. If the place name is given by stating the name of a place and followed by another name in a bigger political division, it should be tagged as one:
      - *{Tixtla, Guerrero}L*
   v. The same as before if one entity is accompanied by its acronym or abbreviation it was tagged as just one entity.
      - *{Chiapas (Chis.)}L*
   vi. In the case of lists, usually the fist entity contains the word that identifies the type of entity, this should be marked along with all the elements of the list. I.E
      - *{municipios de Escárcega}L, {Carmen}L, {Champoton}L, {Calkiní}L y {Hopelche}L del estado de {Campeche}L*; ({counties of Escárcega}I, {Carmen}L, {Champoton}L, {Calkiní}L and {Hopelche}L of the state of {Campeche}L)

5. Other names (*O*)
   Some other names were identified that did not fit in any of the previous categories, these are rare in the text and only a few of them were found in the quality control texts. These were tagged with the letter *O*.

   No actual rules were defined for this kind except that the name is consistent enough to be considered within the text. Also if the name is accompanied by its acronym it is tagged as just one entity. Some examples of this kind correctly tagged are:

   - *{IVA (Impuesto al Valor Agregado)}O* ({added value tax (IVA)}O)

6. Unidentifiable
   There are very few cases that are impossible to classify them in these categories, however they were considered to be important. In the golden standard piece of corpus (fifty texts) only one occurrence of this case was found.
   In the only example it was impossible to identify of which kind the entity was so it was marked as unknown, the example found is:

   - *En una canoa, desalojan una población de {Tixtla, Guerrero}L ({EFE}?);* (In a small boat, is evacuated one population of {Tixtla, Guerrero}L ({EFE}?)

   These cases were not be tagged.

3.1.2.2 Named entities annotation statistics

The global statistics of named entities that were annotated in the corpus (552 documents) following the rules above are:

Named Entities Type	Number of named entities	Percentage %
Place (L)	4,243	27.08 %
Disaster (D)	234	1.49 %
Institutions (I)	5,397	34.45 %
Persons (P)	4,018	25.65 %
Other (O)	1,772	11.25 %
Total	**15,664**	**100 %**

Table 12 Named entities statistics in the corpus

As can be seen the disasters is the category with less entities, this is because in the corpus only a few hurricanes and storms are explicitly mentioned. By fat eh institutions are the more numerous with 34% followed close by place and person names, which clearly indicates that these 3 named entitieare majority in the corpus.

3.1.2.3 Temporal Expressions annotation

A key feature needed for the corpus was the tagging of temporal expressions, for this key process a set of guidelines were done and the same files used for named entity tagging were used, choosing not to overlap the tags from both stages of tagging (named entities and temporal expressions tagging).

The rules were defined as follows, taking into account basic features of temporal expressions such as frequency or duration and choosing not to tag certain types of temporal expressions which are hard or impossible to normalize.

The tagging procedure was similar to the name entity tagging, but the difference is that the temporal expressions were tagged as they came on the text only wrapping them in brackets without choosing a single letter to differentiate the type of temporal expressions instead the tags *DAT*, *DUR*, *FRQ* were used according to the typeas shown below; I.E:

{Viernes 21 de Septiembre de 2011}DAT (Friday, 21st of September 2011)

There were three kinds of temporal expressions needed to be tagged each with their own rules and exceptions which are as follows:

1. Dates (*DAT*)
   This kind of temporal expressions refers to a single time point, they might be explicit or implicit; and are annotated with the *DAT* tag after the brackets.
   - Explicit dates
     Some examples of this kind of temporal expressions are:
     - *{15/05/2013}DAT*
     - *{Martes 15 de Mayo de 2013}DAT* (Tuesday, 15th of May 2013)
     - *{El lunes}DAT* (Monday)
     - *{a las tres de la tarde con veintitres minutes}dat* (At three on the afternoon with twenty three minutes).
     - *{Cuarto para las cuatro}DAT* (quarter to four).
   - Implicit Dates
     This kind of temporal expressions conveys the date in word that express time:
     - *{el més pasado}DAT* (last month)
     - *{ayer}DAT, {mañana}DAT* (yesterday, tomorrow)
     - *{mañana por la mañana}DAT* (tomorrow morning)
     - *{antiguamente}DAT* (in ancient times)
     - *{En un futuro}DAT* (in the future)

2. Expressions of duration (*DUR*)
   Also time expressions that express a duration were tagged, the *DUR* tag was used on this kind of temporal expressions I.E.:
   - *{Toda la semana}DUR* ([during] All week)
   - *{Durante tres años}DUR* (During three years)
   - *{por muchos años}DUR* (during many years)

3. Frequency expressions (*FRQ*)
   Also expressions that convey frequencies were tagged, the *FRQ* tag was used on these temporal expressions I.E.:
   - *{Todos los dias}FRQ* (every days)
   - *{Diariamente}FRQ, {diario}FRQ* (daily)
   - *{Mensualmente}FRQ* (monthly)
   - *{cada tercer dia}FRQ* (every third day)
   - *{cada quince dias}FRQ* (every fifteen days)

4. Exceptions
   Also there were cases that seemed to be temporal expressions (or actually were) but was decided to not consider them for this work, this is because either, the expressions are part of a name, or the disambiguation and normalization is impossible or completely fuzzy.
   The exceptions were:
   - Age expressions:
     - *Tirso Cruz Yuca, de 46 años* (Tirso Cruz Yuca, 46 years old)
   - Expressions that determine a name:
     - *el ciclo escolar 2013-2014* (The scholar cycle of 2013-2014)

These were the guidelines given to the human annotators, further in this work the results of the annotation are presented.

3.1.2.4 Temporal Expressions normalization

First the rules defined to normalize the temporal expressions are shown ahead, these were done according to the experiment needs and do not follow a proper normalization standard.

Dates, frequencies and durations are addressed for the normalization purposes into two sets of time points, beginning and ending.

All the normalizing dates were chosen to be in the format: *YYMMdd HH:mm* I.E *131022 00:00.*

1. Dates (DAT)
    a. Dates that cannot be normalized at all I.E. *The day the people rises.* Should be marked with both dates as the first minute of 1950. I.E.
        - *The day the people rises:*
            - Start: *500101 00:00*
            - End: *500101 00:00*

    b. Regardless of granularity the dates were normalized in a minute granularity which is the minimum considered in this work. I.E.
        - *22th of October (2013) :*
            - Start: *131022 00:00*
            - End: *131022 23:59*
        - *2013:*
            - Start: *130101 00:00*
            - End: *131231 23:59*
        - *At 3 pm (22th of October 2013)*
            - Start: *131022 15:00*
            - End: *131022 15:59*

2. Frequencies (FRQ)
    a. For this type it was only needed to state the duration of the frequency and state it with a starting and ending date that reflected this duration.
    b. Similarly frequencies that convey two temporal expressions inside like *The 19th of every month.* Should be normalized the same way as the un-normalizable dates are. I.E.:
        - *19th of every month:*
            - Start: *500101 00:00*
            - End: *500101 00:00*

    c. Frequencies starting and ending dates should be tough towards 1950. Not that it matters, but it helps readability. I.E.
        - *Every week*
            - Start: *500101 00:00*
            - End: *500106 23:59*

3. Durations (DUR)
    a. Similarly durations are marked with a starting and ending dates which reflects the duration.
    b. Again dates should be marked in the year 1950. I.E.

The prisoner got {25 years}DUR of prison.

- Start: *500101 00:00*
- End: 741231 *23:59*

Furthermore another classification was done in the case that the temporal expressions did not have both boundaries defined. The table below states this boundaries classification:

Open Boundary	Value	Example	Notes
Direct Date (no open boundary)	**DD**	*October the 22nd*	This is when the date is complete and explicit.
Indirect Date (no open boundary)	**ID**	*Yesterday*	When is a complete date but it's expressed in an indirect way.
Open Left boundary	**ODL**	*Some days ago*	When the start point can't be defined.
Open Right boundary	**ODR**	*In the next hours*	When the end point can't be defined.
Open both sides	**ODA**	*Sooner or later*	When the date is completely ambiguous.

Table 13 Boundaries classification of dates

A tool was developed to aid in the normalization of the temporal expressions simplifying the work. this tool is a simple command line program that reads the Timex annotated corpus and loops trough each document while presents the user the **raw temporal expression**, **the reference date** and **the full line in the text** where the Timex occurred.

Then the results are saved in a file which is later read to add the normalized Timex to the corpus on memory.

This tool has greatly reduced the time of normalization considering common granularity approaches so the normalizing task is easier. Furthermore the tool automatically saves the progress after the temporal expressions of one document has been finished.

3.1.2.5 Temporal expressions annotation statistics

In the same way the global statistics of temporal expressions that were annotated in the corpus (552 documents) following the guideline are:

Timex Type	Number of Timex	Percentage %
Dates (DAT)	2,061	80.32 %
Frequency (FRQ)	68	2.65%
Duration (DUR)	437	17.03 %
Total	**2,566**	**100 %**

Table 14 Temporal expressions statistics in the corpus

It is to notice that most information is conveyed in the date expressions which means that in the corpus we have very good number of these expressions.

3.1.3 Topic classification of the news pieces

The experiment considers that the news pieces will already have an assigned topic on which to extract the events, so a manual classification was done on the corpus to group the documents by topics.

These topics roughly follow the topics used for the gathering of the corpus; however more refinement was done in this classification. The topics and the documents in each are shown ahead:

Topic Id	Topic	Number of News pieces
1	*Maestros, Reforma Educativa* Teachers, Education Reform	96
2	*Partidos Politicos* Political Parties	24
3	*Justicia, Policiacos* Justice, Police related	129
4	*Desastres naturales, enfermedades, accidentes* Natural disasters, diseases and accidents	72
6	*Reformas (menos la educativa)* Government reforms (not the education one)	69
7	*Instituto Federal Electoral (IFE)* Federal Electoral Institute	24
9	*Impuestos, Ley Aduanera* Taxes, Customs duty law	36
8	*Otros Politica* Others in Politics	53
5	*Otros* Others	49
Total	**All**	**552**

Table 15 Topics in which the news pieces were classified

For this purpose a command line tool was developed showing the topics and the titles of the news giving the possibility to see the full text of the news piece if the title was not enough to assign topic.

From the set of news pieces of each topic is where the Event detection annotation and the experiments described further were applied.

This classification was evaluated using the inter annotator agreement, the results of such evaluation are presented on **§4.2.1**.

3.1.4 Event classification of the news pieces

To train and evaluate the system, a manual annotation of the events in the topics sets of news pieces was done.

Id's were given to each event and then assigned to the news pieces, in case the news piece was no part of any particular event the event Id was set to "*0*", which it does not belong to any event or is an event reflected only in one piece of news.

Similarly as in the topic classification a tool was developed, so that for each topic events could be made an assigned while annotating, facilitating the task.

The events assigned are shown below by topic:

- **Topic 1: Teachers, Education Reform**

Topic 1: Teachers, Education Reform *(Maestros, Reforma Educativa)*		
Event Id	**Event**	**Number of News pieces**
0	Non Topics or just one topic per news piece	24
1	Writ of Amparo[22] against the education reform *Amparo contra la reforma educativa*	6
2	Teacher's protest in Michoacán *Protesta de maestros en Michoacán*	9
3	Attack on journalist from teachers in Veracruz *Agresión de maestros a periodistas en Veracruz*	3
4	Confrontation between police and teachers in Cancún *Enfrentamiento entre maestros y policías en Cancún*	5
5	Teacher's protest in Oaxaca *Protesta de maestros en Oaxaca*	5
6	Teacher's protest in DF *Protesta de maestros en DF*	18
7	General teacher's strike *Paro de labores general de maestros*	4
8	Teacher's protest in Veracruz *Protesta de maestros en Veracruz*	7
9	Teacher's protest in Quintana Roo *Protesta de maestros en Quintana Roo*	8
10	Teacher's protest in Chiapas *Protesta de maestros en Chiapas*	7
Total Documents	**All**	**96**

Table 16 Events assigned to the news pieces for topic 1

[22] Writ of Amparo, A Mexican legal resourcehttp://en.wikipedia.org/wiki/Recurso_de_amparo

- **Topic 2: Political Parties**

Topic 2: Political Parties (*Partidos Politicos*)		
Event Id	**Event**	**Number of News pieces**
0	Non Topics or just one topic per news piece	1
1	National Regeneration Movement (Morena) becoming a political party *Morena se convierte en partido político*	3
2	Democratic Revolutionary party (PRD) scandals *Escandalos PRD*	9
3	Impeachment of the leader of the Revolutionary Institutional Party PRI in Morelos *Destitución del lider del PRI en Morelos*	2
4	Candidacy of Josefina Vazquez Mota to the National Action Party leadership *Candidatura de JVM a presidencia del PAN*	4
5	Accusations against Governor Graco Ramirez *Declaración y acusaciones a Graco Ramirez*	3
6	Leadership of PRD *Dirigencia del PRD*	2
Total Documents	**All**	**24**

Table 17 Events assigned to the news pieces for topic 2

- **Topic 3: Justice, Police related**

Topic 3: Justice, Police related (*Justicia, Policiacos*)		
Event Id	**Event**	**Number of News pieces**
0	Non Topics or just one topic per news piece	64
1	Alberto Pashistán's Pardon *Indulto de Alberto Pashistan*	8
2	Federal Electricity Commission (CFE) excessive charges *Cobros excesivos de CFE*	7
3	Governor Andrés Granier Fraud *Desfalco de Andrés Granier*	5
4	Cibernetic threats *Amenazas Cibernéticas*	2
5	Televisa and Azteca Tv Amparos *Amparos de Tv Azteca y Televisa*	2
6	Ex president Fox ex-wife involved in money laundering *Ex esposa de Fox lavado de dinero*	3
7	Televisa's military fake scene *Montaje militar de Televisa*	4
8	Ombudsman election *Elección del Ombudsman*	2
9	Peasants strike *Protesta campesina*	4
10	Narco-violence and formation of auto-defense groups *Narco-violencia y surgimiento de auto-defensas*	9
11	Chief of Government Gabriel Mancera criminalizes citizen protests *Mancera criminaliza la protesta ciudadana*	3
12	Assault on indigenous leader *Atentado a dirigente indígena*	4
13	Elba Esther Gordillo case *Caso Elba Esther Gordillo*	3
15	Results on security *Resultados en seguridad*	9
Total Documents	**All**	**129**

Table 18 Events assigned to the news pieces for topic 3

- **Topic 4: Natural disasters, diseases and accidents**

Topic 4: Natural disasters, diseases and accidents (*Desastres naturales, enfermedades, accidents*)		
Event Id	**Event**	**Number of News pieces**
0	Non Topics or just one topic per news piece	17
1	Hurricane Manuel *Huracán Manuel*	5
2	Hurricane Raymon *Huracán raymond*	23
3	Government response to disasters *Respuesta del gobierno a desastres*	12
4	Cholera outbreak *Brote de cólera*	3
5	Monster truck accident *Accidente de Monster truck*	4
6	Tropical storm Octave *Tormenta tropical Octave*	4
8	Evacuation of Victims *Evacuación de víctimas*	4
Total Documents	**All**	**72**

Table 19 Events assigned to the news pieces for topic 4

- **Topic 5: Others**

Topic 5: Others (*Otros*)		
Event Id	**Event**	**Number of News pieces**
0	Non Topics or just one topic per news piece	39
1	Juan Manuel Márquez box fight *Pelea de Juan Manuel Márquez*	2
2	International book fair on DF *Feria internacional del libro en el DF*	2
3	President Peña Nieto on the Iberoamerican summit *Peña Nieto en la cumbre Iberoamericana*	4
4	Lou Reed's death *Muerte de Lou Reed*	2
Total Documents	**All**	**49**

Table 20 Events assigned to the news pieces for topic 5

- **Topic 6: Government reforms but the education one**

Topic 6: Government reforms but the education one (*Reformas menos la educativa*)		
Event Id	**Event**	**Number of News pieces**
0	Non Topics or just one topic per news piece	10
1	Energetic Reform *Reforma energética*	23
2	Tax reform *Reforma hacendaria*	15
3	Political-electoral reform *Reforma político-electoral*	6
4	Goverment expenses reform *Reforma fiscal*	13
Total Documents	**All**	**69**

Table 21 Events assigned to the news pieces for topic 6

- **Topic 7: Federal Electoral Institute (*Instituto Federal Electoral IFE*)**

Topic 7: Federal Electoral Institute (*Instituto Federal Electoral IFE*)		
Event Id	**Event**	**Number of News pieces**
0	Non Topics or just one topic per news piece	1
1	Disappearance of the Federal Electoral Institute IFE *Desaparición del IFE*	2
2	IFE counselors election *Consejeros del IFE*	6
3	Removal of the autonomy of the state electoral institutes *Remoción de autonomía de institutos electorales estatales*	4
4	Gender equality in the IFE *Equidad de género en el IFE*	2
5	Leonardo Valdés Zurita finishes his period a general counselor of the IFE *Termina periodo de Valdés Zurita*	2
6	Reform of the PAN statutes *Reforma a los estatutos del PAN*	5
7	Electoral re-distritation *Redistritación electoral*	2
Total Documents	**All**	**24**

Table 22 Events assigned to the news pieces for topic 7

- **Topic 8: Others from Politics**

Topic 8: Others from Politics (*Otros de Política*)		
Event Id	**Event**	**Number of News pieces**
0	Non Topics or just one topic per news piece	25
1	Assignation of the capital city fund *Asignación del fondo de capitalidad*	13
2	Expense of the state on bureaucracy *Gasto del estado en burocracia*	2
3	Deal for México *Pacto por México*	3
4	Return of Fausto Vallejo as governor of Michoacán *Fausto Vallejo regresa a la gobernatura de Michoacán*	3
5	Bank of Mexico deficit *Déficit del Banco de México*	2
6	Privatization of Puebla's public water *Privatización del agua potable en Puebla*	2
Total Documents	**All**	**53**

Table 23 Events assigned to the news pieces for topic 8

- **Topic 9: Taxes, Customs duty law**

Topic 9: Taxes, Customs duty law (*Impuestos, Ley Aduanera*)		
Event Id	**Event**	**Number of News pieces**
0	Non Topics or just one topic per news piece	10
1	Raise on the value add tax *Aumento del IVA*	9
2	Customs law approbal *Ley de aduanas*	4
3	Tax on refreshments and junk food *Impuesto a los refrescos y la comida chatarra*	9
4	Unemployment ensurance *Seguro de desempleo*	4
Total Documents	**All**	**36**

Table 24 Events assigned to the news pieces for topic 9

Such were all the events manually classified from the corpus, and against these values the prediction of the machine learning algorithm was evaluated.

The classification was evaluated using the inter annotator agreement to measure how good the human classifier did against another human, these results are presented on **§4.2.2**.

3.2. Methodology on event tracking and temporal similarity

The experiment consisted in two tasks, first do a event detection system from the documents of each topic, for this several features were used including temporal similarity features which consist on the second task of this experiment.

Besides the main tasks a preprocessing and feature extraction had to be done in the corpus, this was done to collect features to represent documents in such a way that they are used by the machine learning algorithm to cluster events.

The methodology is described ahead:

3.2.1 Preprocessing

First as the corpus was gathered and annotated, the preprocessing and feature extraction was built.

The main effort was to find and implement a NER in Spanish and a Timex tagger and normalizer also for Spanish.

3.2.1.1 Text cleaning

As explained when the corpus was gathered a java tool was used to extract the text, BoilerPipe. This tool automatically extracts the main text content of a website; however it does not care about more extensive cleaning which had to be programmed.

The first consideration was to remove all lines that were written in all news pieces of a given newspaper, in other words, removal of the news paper name, the picture descriptions, comments and the leftover navigational links were removed.

Furthermore, the news pieces were explored to refine this process even more, for example patterns depicting the reporters' names, and links and references to related news.

3.2.1.2 Sentence and word level preprocessing

After the first cleaning process was done, another tool was used to perform the task of sentence and word tokenization, surprisingly enough there are not so many Java tools designed for Spanish.

After testing it was decided that best suited library for sentence tokenization would be a modification of the open source library Cue.Language23 which is a Java library for text analysis including Spanish sentence and word tokenization amongst other things.

Cue.Language, by itself does a fine job in sentence tokenization, however some extra rules had to be added, including Spanish abbreviations like *Dip.* (*Diputado*) (Congressman), *Mtro.* (*Maestro*) (Teacher); and others, which were common in the corpus but were not implemented in the library.

After this process the corpus only had one line per sentence.

After the Sentence tokenization was done, testing was carried for word tokenization, Cue.Language was tested and considered but t the end the library TagHelper Tools24 had better results and also had stemming implemented, which proved fitted for the task.

In this case the work consisted in first tokenize the words of each sentence, applying a stop word list, then perform stemming in the tokens and then building of n-grams, only bi-grams and tri-grams were considered.

The stop word list consisted in the TagHelper Tools had around 470 most common words in Spanish and several common words in the corpus were implemented with approximately fifteen stop words.

The retrieval of the Stem is already implemented in the library however it was chosen to keep both the original token and the stem, mainly to build more accurate stop word list and analyze the behavior of the library.

Afterwards the tokens in the line are joined in bi-grams and tri-grams, this is not automatically done by the library, for this task it was considered that no more than two stopped tokens could be between possible n-grams; otherwise it was not made an n-gram.

The information about the tokens is saved in a Java object, from this object a Type dictionary is done with the stems. This builds a high performance tree which contains the types, the token from where they come from, and other basic information on the types and tokens.

[23] Cue.language: A small Java library for simple text analysis - counting strings, identifying languages, and removing stop words. https://github.com/vcl/cue.language

[24] TagHelper Tools: Facilitating Reliable Content Analysis of Corpus Data. http://www.cs.cmu.edu/~cprose/TagHelper.html

The statistics results from this process are shown below:

Feature	Value
Total Tokens	127,997
Total Types (Unigrams)	11,286
Total Types (Bigrams)	81,031
Total Types (Trigrams)	88,733
Total Types (all n-grams 1-3)	181,050

Table 25 Tokens and Types in the corpus

All this tokens and types are available when the corpus is loaded into memory inside a java object and were used later as features in the feature selection stage.

3.2.1.3 Named entities preprocessing

After the named entity recognition and classification task some processed were done to clean and normalize the result of the annotation.

A parser was implemented to upload the NE's and add them to a Java object that contains all the information of a document's named entities.

First given that the text of a named entity might have stop words, they were removed from text itself of the named entities.

Then all symbols that were not numbers or letters were removed and the letters with accents were transformed to their ASCII equivalent.

Furthermore it was considered to make a map or a dictionary in which the different variants of named entities that represent a single entity would be joined. This was programmed in the preprocessing and some of the most common named entities were added to it. However the extreme length of the named entities corpus made this task highly time consuming and so it was not completed.

Finally such map is made available to the annotation evaluation task and the feature selection task.

3.2.2 Temporal similarity

For calculating the temporal similarity, as input the output file of the normalizing tools was used, this contained information of the temporal expressions including the reference date, the normalized dates start and end, the type of temporal expression and the classification according their boundaries.

Makkonen's [Makkonen, 2003], gives a guideline of how she obtained the temporal similarity of the documents, this will be described in detail ahead. All the four methods used for temporal similarity were done following Makkonen's general procedure, modifying accordingly to the needs of every algorithm.

3.2.2.1 Makkonen's temporal similarity algorithms

In Makkonen's work, she only addresses the problem using *date* type temporal expressions, in particular it can be said that only the temporal expressions of the type date, and the boundary classification direct date (DD) and indirect date (ID) are selectable to be used in Makkonen's algorithm, for that only these were selected to perform the temporal similarity, the rest were ignored.

Makkonen's proposal is to use the dates and divide them in a beginning and an end, which was already done in the normalization stage, then calculate the overlapping are of the temporal intervals on a jointed timeline as expressed in the figure:

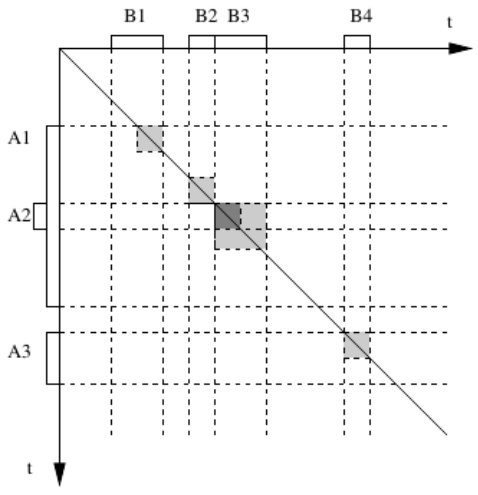


Figure 16 A cross tabulation of two sets of intervals

To calculate said overlapping a set of possible relations of intervals were considered between temporal intervals, these are shown in the figure:

$[t_i, t_j]$ **is before** $[t_k, t_l]$	if $t_j < t_k$
$[t_i, t_j]$ **meets** $[t_k, t_l]$	if $t_j = t_k$
$[t_i, t_j]$ **overlaps** $[t_k, t_l]$	if $t_i < t_k < t_j < t_l$
$[t_i, t_j]$ **begins** $[t_k, t_l]$	if $t_i = t_k \wedge t_j < t_l$
$[t_i, t_j]$ **falls within** $[t_k, t_l]$	if $t_i < t_k \wedge t_j < t_l$
$[t_i, t_j]$ **finishes** $[t_k, t_l]$	if $t_i < t_k \wedge t_j = t_l$
$[t_i, t_j]$ **equals** $[t_k, t_l]$	if $t_i = t_k \wedge t_j = t_l$

Figure 17 Possible relations of intervals

Besides from these possible relations the relation *After* was also calculated nevertheless in not used along with *Before* and *Meets* because the rend no overlapping area.

Afterwards the overlapping are was calculated a statistical measure was done to obtain a number between 0 and 1 the represents the overlapping area between two intervals (temporal expressions), this equation is expressed ahead:

$$\mu_t([t_i, t_j], [t_k, t_l]) = \frac{2\,\Delta([t_i, t_j] \cap [t_k, t_l])}{\Delta(t_i, t_j) + \Delta(t_k, t_l)},$$

Figure 18 Statistical μt to compare two temporal intervals

Where $\Delta : T \times T \rightarrow IR$, $\Delta_{(ti, ti)} = 1$ is the duration (in days). The result is 1 where they are a exact match and 0 if they do not overlap at all.

It is important to notice that Makkonen explicitly only uses the granularity of the Δ's in days.

Furthermore when all the μt are calculated a similarity matrix called the cover matrix is done comparing all the temporal expressions against each other as shown below:

	$T_{2,1}$	...	$T_{2,m}$	max
$T_{1,1}$	$\mu_t(T_{1,1}, T_{2,1})$	...	$\mu_t(T_{1,1}, T_{2,m})$	$v_{1,1}$
⋮	⋮		⋮	
$T_{1,n}$	$\mu_t(T_{1,n}, T_{2,1})$	...	$\mu_t(T_{1,n}, T_{2,m})$	$v_{1,n}$
max	$v_{2,1}$		$v_{2,m}$	

Figure 19 Cover Matrix

Where $v_{1,1} = \max_{j \leq m}(\mu t(T_{1,1}, T_{2,j}))$.

Finally a single number between 0 and 1 is obtained comparing the cover matrix using the following formula:

$$cover_t(T_1, T_2) = \frac{\sum_{i=1}^{n} v_{1,i} + \sum_{j=1}^{m} v_{2,j}}{n+m}.$$

Figure 20 Coverage between two sets of temporal expressions

After this a similarity matrix is done between all documents in the corpus with this number.

This is Makkonen's algorithm and is the base for the other proposed three methods.

3.2.2.2 Proposed temporal similarity algorithms

Three Algorithms were proposed, first the baseline, second the same μt function but with the granularity stated for minutes, which will take into account all selectable temporal expression in the corpus, and then this second μt along with a new defined frequency and duration temporal expressions similarity.

3.2.2.2.1 Baseline

First as baseline only the reference date was used, that is, the date when the news was published as a temporal expression, was used to test document similarity.

For this Makkonen's μt function was used, along with the cover matrix to fill the temporal similarity o the documents, it is obvious to state that the similarity will only be 1 if the published dates are the same or 0 if different.

3.2.2.2.2 Modified μt (Vie)

During the normalization time of the temporal expressions it was noticed that many of them had the minute granularity, this is that represented time intervals that could only be expressed with a minute granularity i.e. at 15:30, during 6 minutes, a minute of silence, half an hour later, etc.

While reproducing Makkonen's experiment, it was noticed that the fact she calculates the difference (Δ) in days, of the μt function, and that was going to be an error source for more granulated expressions like the ones stated above.

That was the reason that it was decided to use a modified difference (Δ) with a minute granularity instead of days, this did not affected the other calculations and enabled the algorithm to handle the more granulated expressions.

So basically it follows all the procedure of Makkonen but changes the Δ to $\Delta : T \times T \rightarrow IR$, $\Delta(ti, ti) = 1$ is the duration minutes.

This algorithm was named **Vie**.

3.2.2.2.3 Modified μt plus frequency and duration similarity (VieAll)

The other new method proposed to use the modified μt function plus two new measures of similarity one for frequency (*FRQ*) expressions and one for duration expressions (*DUR*).

First it is to state that both of these measures are only perform with those types of temporal expressions and is not done for documents without them. Furthermore these are averaged with the μt function only in the case both documents contains expressions

of their kind. In other words the frequency comparison is only done and averaged if both documents contain frequency expressions, and the duration comparison is only done if both documents contain duration expressions; otherwise only the μt is used for the similarity matrix.

Frequency expressions similarity

The similarity between two frequency expressions was defined as a boolean operation, either is the same, or not.

The reason for these is that frequency expressions reflect the duration of time of a repetitive event; i.e. *every day, weekly, every 5 minutes*, etc.

In the case of more complex frequency expressions i.e. *The 19th of every month, 5 minutes every hour*; were excluded at normalization the reason of these is that they convey two expressions as one which are dependant of each other, as no normalizing consideration was done for this expressions, they were ignored at measuring the similarity.

Duration expressions similarity

In the case of the duration expressions, the proposed algorithm consists on comparing the exact duration of the normalized expressions.

This was done by matching the overlapping of durations, obtaining a number between 0 and 1, where 1 is assigned if the durations are the same or cero if they do not overlap, as all durations overlap, cero represents un-normalizable durations such as *For some years*

Average and weighting of duration and frequency similarity

After obtaining the frequency and or duration similarity (if any) a weighting function was done assigning three constant for averaging the following table shows the procedure:

Case of TE	α	β	γ	Averaging Function
DAT	-	-	-	$Sim = Cover_{DAT}$
DUR	-	-	-	$Sim = Cover_{DUR}$
FRQ	-	-	-	$Sim = Cover_{FRQ}$
DAT+ DUR	0.9	0.1	-	$Sim = \alpha\ Cover_{DAT} + \beta\ Cover_{DUR}$
DAT+ FRQ	0.95	0.05	-	$Sim = \alpha\ Cover_{DAT} + \gamma\ Cover_{FRQ}$
DAT+DUR+FRQ	0.85	0.1	0.05	$Sim = \alpha\ Cover_{DAT} + \beta\ Cover_{DUR} + \gamma\ Cover_{FRQ}$

Table 26 Average Functions and their constants

These constants are configurable for experimentation.

After obtaining the similarity with the given functions the full similarity matrix in all the documents is done. This algorithm was named **Vie-All**.

3.2.3 Feature extraction

Features obtained from preprocessing and the named entities and temporal similarity algorithms were fitted in a data model suited for Weka algorithms to process.

The features were selected only with documents on the same topic effectively treating each topic as a separate problem from each other. The sets of features selected are three:

1. Keywords

   In this case the selection of keywords was as follows:

   - The top ten Tf-Idf keywords of each document were selected.
   - If in those ten keywords are unigrams, bigrams and or trigrams that correspond to the same "keywords" they are joined and more keywords are fed until having again ten.
   - The keywords are mapped into a tree map in which keys are the Tf-Idf values and the values are a list of keywords in the top ten that has that Tf-Idf value; then in case there is a key in the map with more than one keyword as value, they are shuffled.
   - Finally the top five keywords of each document using after the above process are selected.
   - A big map is done to contain all the keywords avoiding having repeated values.

2. Named entities

   In the case of named entities the process is really similar to keywords:

   - The top five Tf-Idf named entities of each type in the annotation document were selected.
   - The NE's are mapped into a tree map in which keys are the Tf-Idf values and the values are a list of NE's in the top five of each type that has that Tf-Idf value; then in case there is a key in the map with more than one named entity as value, they are shuffled for that given value.
   - Finally top ten named entities of each document using after the above process are selected assuring that at least one named entity for each type (if the document has this type of NE).
   - A big map is done to contain all the named entities avoiding having repeated values.

3. Temporal similarity

   In the case of temporal similarity, the results of the algorithms are received as an array of id-number pair in which id is another document in the corpus and the number is the calculated temporal similarity between them. There are four temporal similarity arrays each containing the results of each temporal similarity algorithm. These were the temporal similarity features included in the corpus.

The following table contains the sets of features fed to the machine learning algorithm:

Feature(s)	Number of Features	Description
Id	One	The document Id
Event Id	One	The event in which the document was manually classified
Keywords	Depends on the number of documents on topic	The top five Tf-Idf ranked keywords of each document on topic
Named Entities	Depends on the number of documents on topic	The top ten Tf-Idf ranked NE'd of each document on topic
Temporal similarity	4 * Number of documents on topic	Temporal similarity of each document against each other for all four algorithms

Table 27 Feature sets description

After the feature matrix was done on memory, a comma separated value file (CSV) was produced for the Weka machine learning algorithm to read. One feature CSV file were done per topic.

3.2.4 Event tracking

The event tracking experiment took part in two stages first a system set up, with the used Weka k-nearest neighbor classifier, and second the parsing and measuring of the results with a ten-fold cross validation.

After the CSV matrixes are loaded and the feature prepared to be used in the experiments.

Then the second stage was when the sets of experiments with the temporal similarity features took place, changing the temporal and non temporal features.

3.2.4.1 Experiments set up

For the set up a program was developed which read the CSV files containing the feature matrixes of all topics.

The machine learning package Weka was selected for this task for three main reasons:

1. The existence of an explorer in which to try the first experiments. And select and set up the classifier module.
2. The same classifier can be implemented programmatically and evaluated for all experiments at once without having to repeat the experiments over and over again.
3. It is written in the Java programming language giving the whole project continuity and integration against using a different language package.

Then a feature set selection module was implemented in which the different features were mapped out of the feature map and converted into Weka Instances Objects.

The classifier selected was the IBK classifier which is an implementation of the K-nearest neighbor classifying algorithm. With the parameters set to -K 2 -W 0 -X which means that the classifier assigns the class label of only the two closest neighbors.

These parameters were tested on the Weka explorer, at this end this configuration was selected because it rendered the best results on the test topics.

After the classifier parameters were selected, it was proceeded to implement it using Weka classes within the system.

A feature filter was done both to remove the useless features such as the document id, and to select only the temporal similarity features of the algorithm to be tested which is described further.

3.2.4.2 Feature sets selection on the experiments

When the experiment was setup the feature selected for every experiment where fed to the classifier. Given that the temporal similarity features of the four different algorithms the experiments develop around these features contrasting them against the others sets namely named entities and keywords and the combination of them.

The experiments performed for each of the topics are summarized in the next table:

Feature combination in the experiments

Non Temporal \ Temporal	Nothing	Base Line	Makonnen's	Vieyra	Vieyra with Frq and Dur
None	----------	BL	Mak	Vie	VieAll
Only Keywords	Only Kw	BL + Kw	Mak + Kw	Vie + Kw	VieAll + Kw
Only Named Entities	Only Ne	BL + NE	Mak + NE	Vie + NE	VieAll +NE
Keywords and Named Entities	Kw + NE	BL + Kw + NE	Mak + Kw + NE	Vie + Kw + NE	VieAll + Kw + NE

Table 28 Feature combinations used in the experiments

In the table, the four temporal similarity algorithms are shown against the non temporal features namely keywords and named entities. In total there were 19 experiments done.

It is to mention that the first three experiments shown in the table in the first row, which are using only the temporal similarity were carried out but they results were not averaged and were used as a guidance of how the classifier varies.

3.2.4.3 Automatic evaluation

Weka's cross validation module was used for the evaluation; the parameters for the evaluation were a ten-fold validation, selecting the event feature (manually classified event) as the class to be classified on.

The results of each topic cross evaluation, namely the precision, recall and f-measure were done with Weka's implemented methods however it is to notice that this methods give more weight in classes with more documents.

After all this an overall evaluation of each temporal similarity experiment were done summarizing the results with the different non temporal feature sets, leaving out the ones that only use temporal features as stated before.

4. Results of the evaluation and discussion

The evaluation is divided in two parts, first the evaluation of the corpus annotation and hand-made topic and event classification, and second the event tracking experiments results.

4.1 Named entities and temporal expressions annotation evaluation

First the results of the named entity delimitation and classification quality control are given. Second, the same evaluations but for temporal expressions plus the evaluation of the temporal expressions normalization.

4.1.1 Named entities annotation evaluation

The quality control results of the proficiency of the annotators in named entities delimitation, they were measured by precision, recall and F measure and are shown ahead:

Feature	Value
Total NE in Golden Standard	1,450
Total NE in Sub-Corpus50	1,457
True Positives	1,402
False Positives	55
False Negatives	48
Precision	**96.225 %**
Recall	**96.690 %**
F Measure	**96.457 %**

Table 29 Quality measure of named entities detection

Of these results, the true positives (1402) are matched pairs of correspondent named entities both in the golden standard (GS) and in the Sub-corpus given for quality control purposes (Sub-Corpus50).

After the matched pairs were done, a quality measure was done for the classification of the named entities, which was also a task for the annotators; the results by type are given ahead:

For the Place named entity type (L):

Place (L)	Value
Total	338
True Positives	324
False Positives	14
False Negatives	14
Precision	**95.858 %**
Recall	**95.858 %**
F Measure	**95.858 %**

Table 30 Quality results for the Place tag (L)

For the Disaster named entity type (D):

Disaster (D)	Value
Total	20
True Positives	19
False Positives	0
False Negatives	1
Precision	**100.00 %**
Recall	**95.000 %**
F Measure	**97.436 %**

Table 31 Quality results for the Disaster tag (D)

For the Institution named entity type (I):

Institution (I)	Value
Total	582
True Positives	516
False Positives	24
False Negatives	66
Precision	**95.556 %**
Recall	**88.660 %**
F Measure	**91.979 %**

Table 32 Quality results for the Institution tag (I)

For the Person name named entity type (P):

Person (P)	Value
Total	382
True Positives	369
False Positives	6
False Negatives	13
Precision	**98.400 %**
Recall	**96.597 %**
F Measure	**97.490 %**

Table 33 Quality results for the Person name tag (P)

And finally for the Other named entity type (O):

Other (O)	Value
Total	80
True Positives	67
False Positives	63
False Negatives	13
Precision	**51.538 %**
Recall	**83.750 %**
F Measure	**63.810 %**

Table 34 Quality results for the Other tag (O)

From the results, it can be observed that the tag for Other (O) has a very bad precision; this is due that the type itself it's difficult to detect and correctly identify the named entity. Usually Others (O) is used for Taxes names, or an entity that can't be unambiguously classified.

All the others result are very good and assure that the corpus quality is sufficient for this work purposes and any others that may come out from this corpus.

4.1.2 Discussion on the named entities annotation and evaluation

The annotation of named entities provides a series of challenges to overcome, given that the quantity of named entities is very large, some categories ambiguous, human annotation errors, and the repetitive work and considerable amount of time required to annotate and classify them.

For the annotation and classification of named entities there is to notice some problems and facts that need to be discussed:

1. The large amount of named entities

   One of the things to notice is that in the corpus a large number of named entities were found, this implies two things first that named entities are representative of each of the types and also that having so many to deal with is inevitable to come across unclassifiable cases and weird exceptions.

2. Ambiguous, hard to delimitate or incomplete named entities

   Other important thing to notice is that given the large quantity and the diversity of named entities, there are some cases that are hard to delimitate or the entities are incomplete this can be a consequence of cleaning errors during preprocessing or parsing errors of the corpus or human errors at annotation time.

3. Ambiguity in some class of named entities

   For some types of named entities it is complicate to differentiate them of the other for the classification, for example some uncommon tax name annotated as an institution, or a hurricane annotated as a person. Sometimes this can be because the annotator did not read the entire context around the named entities or simply because it is not specified in the text.

4. Markup mistakes in some types

   In some types, especially in the places type, there is complicated to distinguish different places or the markup is complicated due that the entities are in a complicated list.

5. Annotation mistakes and amount of work

   Given the large numbers and the attention span needed to annotate the entities it is impossible to not mention that several human mistakes are to be done, in the best case scenario the delimitation is wrong and can be fixed with temporal expressions, however sometimes the errors are more complicated or simply the annotator did not put enough attention making rendering the correct parsing of said expressions complicated if not impossible.

Finally these details sum up in the final quality of the annotated corpus, even though the evaluation gives an idea of how good this was done there is to notice that in particular complicated are expected to have some level of inaccuracy given the features of such named entities.

Now, regarding the results of the evaluation of the annotation and classification first the delimitation results are to be considered very good given the big number of expressions.

It is to notice that an F-measure of 96% indicates an excellent precision and recall and indicates that for the human annotators the delimitation of named entities is not a problem for them.

In the case of the results of the classes, is evident that the classification provides a harder problem for the human annotators, even so the results for all the categories are very good which also proves the efficiency of the annotators at this task

It is only to notice the low performance of the annotators in the Other (O) type, the type itself is ambiguous and highly subjective and depends on what each annotator considers to be an entity of a given class. The precision is fairly low however the recall is very high indicating that most of these expressions were annotated but many were wrongly annotated.

Finally it is worth mentioning that both the annotation and classification of named entities was a task very well done for the annotators, two reasons for this can be first that the guidelines were highly specific for what it was required and, second that annotators were highly familiarized with the entities in the corpus. Also there is to consider that many of the entities were highly repetitive which facilitated in some way the task.

4.1.3 Temporal expressions annotation evaluation

In this part first the results of the temporal expression delimitation and classification are given and second the results of the evaluation of temporal expression normalization are shown.

4.1.3.1 Temporal expressions annotation and classification evaluation

Quality control results of the annotator's proficiency in temporal expressions extraction, measured by precision, recall and F measure are given ahead:

Feature	Value
Total Timex in Golden Standard	255
Total Timex in Sub-Corpus50	188
True Positives	172
False Positives	16
False Negatives	83
Precision	**91.489 %**
Recall	**67.451 %**
F Measure	**77.652 %**

Table 35 Quality measure of temporal expressions detection

Of these results, the true positives (172) are matched pairs of correspondent temporal expressions both in the golden standard (GS) and in the Sub-corpus given for quality control purposes (Sub-Corpus50).

After the matched pairs were done, a quality measure was done for the classification of the temporal expressions, which was also a task for the annotators; the results by type are given ahead:

For the Date temporal expression type (DAT):

Date (DAT)	Value
Total	136
True Positives	131
False Positives	10
False Negatives	5
Precision	**92.908 %**
Recall	**96.324 %**
F Measure	**94.585 %**

Table 36 Quality results for Dates (DAT)

For the Frequency temporal expression type (FRQ):

Frequency (FRQ)	Value
Total	2
True Positives	1
False Positives	1
False Negatives	1
Precision	**50.00 %**
Recall	**50.00 %**
F Measure	**50.00 %**

Table 37 Quality results for Frequency (FRQ)

And finally for the Duration temporal expression type (DUR):

Duration (DUR)	Value
Total	34
True Positives	24
False Positives	5
False Negatives	10
Precision	**82.759 %**
Recall	**70.588 %**
F Measure	**76.190 %**

Table 38 Quality results for Duration (DUR)

It is to notice that in the text selected as a golden standard, the number of named entities of the type frequency were very low (only two occurrences) thus is hard to measure the evaluation with those low numbers.

4.1.3.2 Temporal expressions normalization evaluation

For the evaluation of the temporal expressions normalization the same metric as before was used, first precision recall and F-measure are given, and then a metric of the precision recall and F-measure of the boundary classification was are presented; there is to notice that the boundary classification is only done for date types of temporal expressions.

For this the temporal expressions of the golden standard were selected and normalized by two persons using the normalizer helper tools, these two results of the golden standard's temporal expressions were the ones used for the evaluation.

The way to measure the normalization of the expressions was simply to compare the normalized temporal expression starting and ending date string as a single string, this implies that both dates have to be correct to count as correctly normalized. Furthermore the probability of error was calculated to be very low given that the strings have to be perfect so it was set at a very low value[25].

The results of the normalization are shown ahead:

Feature	Value
Total Timex	255
Correctly normalized	201
Agreement Pr(a)	0.7882
Chance Probability Pr(e)	0.0500
Inter-annotator agreement	**0.7771**

Table 39 Inter-annotator agreement for temporal expression normalization

In the case of the open-boundary classification of the normalized named entities, precision recall and F-measure was done only in the date temporal expressions in the golden standard.

[25] Given that the string is really long there is really low probability of chance a random low value was chosen to represent the hypothetical chance agreement scenario.

The overall results of the open-boundary classification are shown on the table below:

Open-boundary classification	**Value**
Total Date Temporal expressions	199
True Positives	152
False Positives	18
False Negatives	47
Precision	**89.411 %**
Recall	**76.381%**
F Measure	**82.383%**

Table 40 Results of the open-boundary classification of Date temporal expressions

The discussion of these and the previous results is given ahead.

4.1.4 Discussion on Temporal expressions annotation and normalization evaluation

The annotation, classification and normalization of named entities provides even more challenges to overcome, given that the quantity of temporal expression is very large, some categories ambiguous, human annotation errors, and the repetitive work and considerable amount of time required to annotate and classify them not mention the even more time consuming normalization.

For the annotation, classification and normalization of named entities there is to notice some problems, facts and remarks that need to be discussed:

1. The large amount of temporal expressions

   One of the things to notice is that in the corpus a large number of temporal expressions were found, this implies the same things that in named entities, better representativity but more complicated cases and ambiguous to normalize and classify.

2. Ambiguous, hard to delimitate or incomplete temporal expressions

   The same problem as in named entities arises, delimitation might be even more complicated and many are well hidden in words and can be hard to spot or properly delimit.

3. Ambiguity in some class of temporal expressions

   This problem aggravates with temporal expressions, the corpus shows that the annotators often confuse durations with dates and vice versa.

4. Markup mistakes in some types

   In some types, especially in the frequencies and some dates, there is complicated to distinguish different entities and correctly classify them.

5. Big set of rules during normalization

   Even as it does not seem to be that many rules in the normalization guidelines, quickly when normalizing conventions, durations and frequencies rules and human ways to simplify the job, all of which provide error sources but prove to be inevitable given the large quantity of work.

6. Complicated cases

   Some named entities prove very difficult if not impossible to normalize for example *The 19th of every month*, or, *two hours every two days*; these were left behind however it's worth mentioning the noise and confusion they provided at normalization time.

7. Annotation mistakes and amount of work

The same problem exacerbated for the hard, time-consuming task of normalization, inevitably several human mistakes were done, finding the normalization one particularly hard to find and repair.

Now regarding the results of the evaluation of the delimitation, classification and normalization:

The results of the delimitation task has very good precision 91% meaning that the annotator were familiarized with temporal expressions and that they were able to identify them, however the recall is too low only 67% meaning that there are a lot of expressions not seen or not annotated, this is especially true with expression embedded within words like *today*.

Now for the type classification the best results were for the date type this is mainly because it is the easiest type to spot, the expressions are more explicit than others. The annotators got very good proficiency this type.

In the case of the duration type, the precision is good 82% but the recall really low the reason for this is that many duration expressions were incorrectly annotated as dates and vice versa, however the F-measure is acceptable.

Finally for the frequencies; it is obvious to notice the lack of this kind of expressions, there were only two in the golden standard making it effectively impossible to measure accurately. Nevertheless still even with two samples of frequencies, we can see there is a mistake.

In the case of normalization, the agreement is 0.78 and the inter-annotator agreement 0.77 which means that the annotators are in substantial agreement but still there is almost 20% of incorrect normalized expressions.

These errors are more evident in the full normalization of the corpus because it took many hours of repetitive work and the attention span cannot be maintained for long.

These numbers also gives a good idea of how humans perform in normalizing temporal expressions.

Finally the open boundary classification results proves to also be somewhat problematic with high precision 89% but fairly low recall 76% which means that the annotators had trouble classified given the ambiguity or human mistakes.

To conclude the discussion it is to say that temporal expressions normalization is a far more complicated task than expected with several cases and rules to be considered, furthermore human annotators even if trained cannot achieve a near perfect agreement in all the stages from the annotation, classifications and normalization.

4.2 Topic and event hand-made classification

The topic and event handmade classification evaluation was done using the Cohen's kappa for inter-annotator agreement. For the topics, only one measure is given as an overall agreement, but for the events the results are given by topic.

4.2.1 Topic classification

The results of the handmade topic classification are shown ahead:

Feature	Value
Total Documents	552
Total topics	9
Correctly classified	417
Agreement Pr(a)	0.7554
Chance Probability Pr(e)	0.1111
Inter-annotator agreement	**0.7249**

Table 41 Inter annotator agreement for topic classification

As we can see the probability of mere chance is rather high given that the topics are diffuse and news pieces might be in several topics.

4.2.2 Event classification

The results of the event handmade classification by topic are shown in the table below:

Topic	#Documents	Number of events	Correctly Classified	Pr(A)	Pr(e)	Kappa
Topic 1	96	11	72	0.7500	0.0909	0.7250
Topic 2	24	7	16	0.6667	0.1429	0.6111
Topic 3	129	15	92	0.7132	0.0667	0.6927
Topic 4	72	9	55	0.7639	0.1111	0.7344
Topic 5	49	5	44	0.8980	0.2000	0.8724
Topic 6	69	5	48	0.6957	0.2000	0.6196
Topic 7	24	8	20	0.8333	0.1250	0.8095
Topic 8	53	7	46	0.8679	0.1429	0.8459
Topic 9	36	5	29	0.8056	0.2000	0.7569
			Average	**0.7771**	**0.1422**	**0.7402**

Table 42 Inter-annotator agreement of the event classification per each topic.

In these cases the events are obviously topic-related and it might be confusing to classify the news pieces correctly as can be seen in the results, nevertheless we can see the average kappa to be 0.74 which means that there is substantial agreement between the annotators.

This also means that the system in those cases cannot perform better effectively making this the ceiling proficiency of the event tracking experiments.

4.2.3 Discussion on the handmade topic and event classification

The handmade classification for topics and events makes a crucial part of this work, first for defining the actual topics and events and then in the evaluation to provide a clear picture of how human annotators perform in the same task as the classifier.

A clear point to discuss is that the events and topics were selected in a subjective way that could not be avoided, this is given the total news pieces first a single list of topics and then of events were done by a single person, and after other human classified the news pieces in those events and topics.

This means that the second human classifier, had to o try to fit the news pieces into topics already given to him and given that some news pieces may belong to more than one topic and the same for events, the annotator performed with a reasonable agreement but not as good as it was expected.

Regarding the topics, the selection of the topics was done looking to spread the events the most balanced way however for example the topic *Political parties* is far more semantically distanced from *Disasters* than from *Federal Electoral Institute*. This gives the impression that certain granularity on topics is present and that is violated by selecting these topics.

However selecting topics by their so call granularity or better said specificity going to more general to more particular would mean more time in annotation and less examples in each event of news pieces to cluster thus it was left the way it is.

The *Other Political* and *Others* topics are of particular interest and sources of confusion, given that these topics actually contains all documents that could not be fitted in different topic, so the events of them are fewer and the news pieces per event even less.

Other important point to remark is that the number of documents is not in any way proportional to the number of documents per topic, meaning that the topics are highly imbalanced; this was noticed from the beginning but was chosen to left it as it is because this would provide a more realistic scenario of this kind of work.

Finally about the topics classification, given the said reasons the kappa statistic gave only 72% which still means considerable agreement.

Then regarding the event handmade classification the same concerns about the topics arises, the other event makes noise in the classification confusing the annotators regarding in which event to classify the ambiguous news pieces.

This is especially true in the topic *Others* the event *Others* contains most of the news pieces (around 90%) making the topic extremely imbalanced in its events and most of the texts were classified in the *Others* event.

Regarding the results of the quality control we can see that in the *Others* and the *Others Politics* got the highest agreement, around 80%, this is representative of what was said, then the topic Federal Electoral Institute got 80% which is an almost perfect agreement.

The lowest was 61% in *Political Parties* and the reason is that many news pieces mention other events and link them in the text therefore was hard for the annotator to correctly classify.

Finally the average of the inter-annotator agreement for the event handmade classification was 74% which means the annotators are in considerable agreement, providing a ceiling for the event tracking experiments to expect.

4.3 Event tracking evaluation

In the case of the evaluation of the event tacking first the results of the experiments per topics are given and then an overall evaluation and ranking of the temporal similarity algorithms performance.

4.3.1 Event tracking evaluation per topic

The results of the event tracking evaluation presented per topic are shown below, it is to notice that in the tables, *Nothing* means no temporal similarity used, *BL* is the baseline, *Mak* is Makkonen's algorithm and *Vie* and *VieAll* are the proposed algorithms, furthermore *Kw* means keywords and *NE* means named entities, and top TS is the best temporal similarity algorithm in that case:

For topic 1:

Topic 1 Teachers, Education Reform

Precision

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.16432	0.18190	0.17492	0.21633	**VieAll**
Only Kw	0.20551	0.16432	0.18190	0.17492	0.21633	**VieAll**
Only NE	0.29142	0.25666	0.20089	0.34243	0.34243	**Vie - VieAll**
Kw and NE	0.29142	0.25666	0.20089	0.34243	0.34243	**Vie - VieAll**
Average	0.26278	0.22588	0.19456	0.28659	0.30040	**VieAll**

Recall

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.20833	0.20833	0.19792	0.20833	**BL - Mak - VieAll**
Only Kw	0.25000	0.20833	0.20833	0.19792	0.20833	**BL - Mak - VieAll**
Only NE	0.25000	0.28125	0.20833	0.25000	0.25000	**BL**
Kw and NE	0.25000	0.28125	0.20833	0.25000	0.25000	**BL**
Average	0.25000	0.25694	0.20833	0.23264	0.23611	**BL**

F-Measure

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.18121	0.18140	0.17202	0.18155	**BL - Mak - VieAll**
Only Kw	0.20925	0.18121	0.18140	0.17202	0.18155	**BL - Mak - VieAll**
Only NE	0.17348	0.24779	0.14833	0.17996	0.17996	**BL**
Kw and NE	0.17348	0.24779	0.14833	0.17996	0.17996	**BL**
Average	0.18540	0.22560	0.15935	0.17731	0.18049	**BL**

Table 43 Event Tracking results for topic 1

For topic 2:

Topic 2 Political Partiesit is

Precision

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.21930	0.51202	0.42521	0.40923	**Mak**
Only Kw	0.56667	0.21930	0.51202	0.42521	0.40923	**Mak**
Only NE	0.60328	0.26649	0.40923	0.50577	0.50577	**Vie-VieAll**
Kw and NE	0.60328	0.26649	0.40923	0.50577	0.50577	**Vie-VieAll**
Average	0.59108	0.25076	0.44349	0.47892	0.47359	**Vie**

Recall

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.41667	0.50000	0.45833	0.41667	**Mak**
Only Kw	0.58333	0.41667	0.50000	0.45833	0.41667	**Mak**
Only NE	0.58333	0.41667	0.41667	0.50000	0.50000	**Vie-VieAll**
Kw and NE	0.58333	0.41667	0.41667	0.50000	0.50000	**Vie-VieAll**
Average	0.58333	0.41667	0.44444	0.48611	0.47222	**Vie**

F-Measure

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.28274	0.46816	0.41023	0.38447	**Mak**
Only Kw	0.55701	0.28274	0.46816	0.41023	0.38447	**Mak**
Only NE	0.56637	0.31762	0.38447	0.46370	0.46370	**Vie-VieAll**
Kw and NE	0.56637	0.31762	0.38447	0.46370	0.46370	**Vie-VieAll**
Average	0.56325	0.30599	0.41237	0.44588	0.43729	**Vie**

Table 44 Event Tracking results for topic 2

For topic 3:

Topic 3 Justice, Police related

Precision

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.39004	0.33522	0.32600	0.32600	**BL**
Only Kw	0.34514	0.39004	0.33522	0.32600	0.32600	**BL**
Only NE	0.50738	0.38332	0.43658	0.39442	0.39442	**Mak**
Kw and NE	0.50738	0.38332	0.43658	0.39442	0.39442	**Mak**
Average	0.45330	0.38556	0.40279	0.37161	0.37161	**Mak**

Recall

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.38760	0.39535	0.38760	0.38760	**Mak**
Only Kw	0.48837	0.38760	0.39535	0.38760	0.38760	**Mak**
Only NE	0.53488	0.44961	0.49612	0.48062	0.48062	**Mak**
Kw and NE	0.53488	0.44961	0.49612	0.48062	0.48062	**Mak**
Average	0.51938	0.42894	0.46253	0.44961	0.44961	**Mak**

F-Measure

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.37414	0.35858	0.34939	0.34939	**BL**
Only Kw	0.39673	0.37414	0.35858	0.34939	0.34939	**BL**
Only NE	0.44737	0.40296	0.40365	0.39108	0.39108	**BL-Mak**
Kw and NE	0.44737	0.40296	0.40365	0.39108	0.39108	**BL-Mak**
Average	0.43049	0.39335	0.38863	0.37718	0.37718	**BL**

Table 45 Event Tracking results for topic 3

For topic 4:

Topic 4 Natural disasters, diseases and accidents

Precision

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.22385	0.30958	0.27093	0.27093	**Mak**
Only Kw	0.33491	0.22385	0.30958	0.27093	0.27093	**Mak**
Only NE	0.20465	0.33650	0.45258	0.43087	0.43087	**Mak**
Kw and NE	0.20465	0.33650	0.45258	0.43087	0.43087	**Mak**
Average	0.24807	0.29895	0.40491	0.37756	0.37756	**Mak**

Recall

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.30556	0.33333	0.40278	0.40278	**Vie-VieAll**
Only Kw	0.37500	0.30556	0.33333	0.40278	0.40278	**Vie-VieAll**
Only NE	0.31944	0.41667	0.47222	0.47222	0.47222	**Mak-Vie-VieAll**
Kw and NE	0.31944	0.41667	0.47222	0.47222	0.47222	**Mak-Vie-VieAll**
Average	0.33796	0.37963	0.42593	0.44907	0.44907	**Vie-VieAll**

F-Measure

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.25771	0.29739	0.31031	0.31031	**Vie-VieAll**
Only Kw	0.28599	0.25771	0.29739	0.31031	0.31031	**Vie-VieAll**
Only NE	0.22235	0.35296	0.41373	0.39458	0.39458	**Mak**
Kw and NE	0.22235	0.35296	0.41373	0.39458	0.39458	**Mak**
Average	0.24356	0.32121	0.37495	0.36649	0.36649	**Mak**

Table 46 Event Tracking results for topic 4

For topic 5:

Topic 5 Others

Precision

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.66639	0.66327	0.66327	0.66327	**BL**
Only Kw	0.83230	0.66639	0.66327	0.66327	0.66327	**BL**
Only NE	0.72109	0.72109	0.72109	0.72109	0.72109	**Same**
Kw and NE	0.72109	0.72109	0.72109	0.72109	0.72109	**Same**
Average	0.75816	0.70286	0.70181	0.70181	0.70181	**BL**

Recall

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.81633	0.79592	0.79592	0.79592	**BL**
Only Kw	0.87755	0.81633	0.79592	0.79592	0.79592	**BL**
Only NE	0.83673	0.83673	0.83673	0.83673	0.83673	**Same**
Kw and NE	0.83673	0.83673	0.83673	0.83673	0.83673	**Same**
Average	0.85034	0.82993	0.82313	0.82313	0.82313	**BL**

F-Measure

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.73378	0.72356	0.72356	0.72356	**BL**
Only Kw	0.83284	0.73378	0.72356	0.72356	0.72356	**BL**
Only NE	0.76933	0.76933	0.76933	0.76933	0.76933	**Same**
Kw and NE	0.76933	0.76933	0.76933	0.76933	0.76933	**Same**
Average	0.79050	0.75748	0.75407	0.75407	0.75407	**BL**

Table 47 Event Tracking results for topic 5

For topic 6:

Topic 6 Natural disasters, diseases and accidents

Precision

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.16831	0.30179	0.30595	0.19430	**Vie**
Only Kw	0.26059	0.16831	0.30593	0.30595	0.19430	**Mak-Vie**
Only NE	0.26781	0.24187	0.30593	0.29836	0.26731	**Mak**
Kw and NE	0.26781	0.24187	0.30179	0.29836	0.26731	**Mak**
Average	0.26540	0.21735	0.30455	0.30089	0.24298	**Mak**

Recall

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.28986	0.37681	0.37681	0.36232	**Mak-Vie**
Only Kw	0.37681	0.28986	0.39130	0.37681	0.36232	**Mak**
Only NE	0.36232	0.33333	0.39130	0.39130	0.37681	**Mak-Vie**
Kw and NE	0.36232	0.33333	0.37681	0.39130	0.37681	**Vie**
Average	0.36715	0.31884	0.38647	0.38647	0.37198	**Mak-Vie**

F-Measure

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.21118	0.29214	0.27885	0.25060	**Mak**
Only Kw	0.26493	0.21118	0.32826	0.27885	0.25060	**Mak**
Only NE	0.29200	0.26202	0.32826	0.32420	0.30389	**Mak-Vie**
Kw and NE	0.29200	0.26202	0.29214	0.32420	0.30389	**Vie**
Average	0.28298	0.24507	0.31622	0.30909	0.28613	**Mak**

Table 48 Event Tracking results for topic 6

For topic 7:

Topic 7 Federal Electoral Institute

Precision

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.13194	0.28526	0.28526	0.28526	**Same**
Only Kw	0.43056	0.13194	0.28526	0.28526	0.28526	**Same**
Only NE	0.43830	0.32143	0.42361	0.42361	0.42361	**Same**
Kw and NE	0.43830	0.32143	0.42361	0.42361	0.42361	**Same**
Average	0.43572	0.25827	0.37749	0.37749	0.37749	**Same**

Recall

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.16667	0.33333	0.33333	0.33333	**Same**
Only Kw	0.45833	0.16667	0.33333	0.33333	0.33333	**Same**
Only NE	0.41667	0.33333	0.54167	0.54167	0.54167	**Same**
Kw and NE	0.41667	0.33333	0.54167	0.54167	0.54167	**Same**
Average	0.43056	0.27778	0.47222	0.47222	0.47222	**Same**

F-Measure

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.14719	0.28582	0.28582	0.28582	**Same**
Only Kw	0.41888	0.14719	0.28582	0.28582	0.28582	**Same**
Only NE	0.35419	0.29444	0.45253	0.45253	0.45253	**Same**
Kw and NE	0.35419	0.29444	0.45253	0.45253	0.45253	**Same**
Average	0.37575	0.24536	0.39696	0.39696	0.39696	**Same**

Table 49 Event Tracking results for topic 7

For topic 8:

Topic 8 Others in Politics

Precision

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.43105	0.48473	0.48473	0.48473	**Same**
Only Kw	0.36522	0.43105	0.48473	0.48473	0.48473	**Same**
Only NE	0.60175	0.49790	0.52683	0.52683	0.52683	**Same**
Kw and NE	0.60175	0.49790	0.52683	0.52683	0.52683	**Same**
Average	0.52291	0.47562	0.51280	0.51280	0.51280	**Same**

Recall

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.35849	0.58491	0.58491	0.58491	**Same**
Only Kw	0.30189	0.35849	0.58491	0.58491	0.58491	**Same**
Only NE	0.62264	0.45283	0.56604	0.56604	0.56604	**Same**
Kw and NE	0.62264	0.45283	0.56604	0.56604	0.56604	**Same**
Average	0.51572	0.42138	0.57233	0.57233	0.57233	**Same**

F-Measure

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.36824	0.52197	0.52197	0.52197	**Same**
Only Kw	0.24721	0.36824	0.52197	0.52197	0.52197	**Same**
Only NE	0.58787	0.45655	0.52673	0.52673	0.52673	**Same**
Kw and NE	0.58787	0.45655	0.52673	0.52673	0.52673	**Same**
Average	0.47432	0.42711	0.52514	0.52514	0.52514	**Same**

Table 50 Event Tracking results for topic 8

For topic 9:

Topic 9 Taxes, Customs duty law

Precision

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.28205	0.23214	0.25289	0.26687	**BL**
Only Kw	0.17097	0.28205	0.23214	0.25289	0.26687	**BL**
Only NE	0.39028	0.55556	0.28411	0.24815	0.29398	**BL**
Kw and NE	0.39028	0.55556	0.28411	0.24815	0.29398	**BL**
Average	0.31717	0.46439	0.26679	0.24973	0.28494	**BL**

Recall

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.30556	0.27778	0.30556	0.33333	**VieAll**
Only Kw	0.27778	0.30556	0.27778	0.30556	0.33333	**VieAll**
Only NE	0.27778	0.38889	0.27778	0.25000	0.30556	**BL**
Kw and NE	0.27778	0.38889	0.27778	0.25000	0.30556	**BL**
Average	0.27778	0.36111	0.27778	0.26852	0.31481	**BL**

F-Measure

	Nothing	BL	Mak	Vie	VieAll	Top TS
Not Kw nor NE	----------	0.28045	0.24129	0.26678	0.28857	**VieAll**
Only Kw	0.17832	0.28045	0.24129	0.26678	0.28857	**VieAll**
Only NE	0.24514	0.38194	0.25661	0.22320	0.28185	**BL**
Kw and NE	0.24514	0.38194	0.25661	0.22320	0.28185	**BL**
Average	0.22287	0.34811	0.25150	0.23772	0.28409	**BL**

Table 51 Event Tracking results for topic 9

Those are the results of the event tracking for each topic; it is to notice the results of the topics 5 and 8, which will be commented further in the discussion.

4.3.2 Overall event tracking evaluation

The global results of each topic and the average of the proficiency of the temporal similarity algorithms are shown ahead:

Precision

Alg	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Avg
No TS	0.2628	0.5911	0.4533	0.2481	0.7582	0.2654	0.4357	0.5229	0.3172	**0.4283**
BL	0.2259	0.2508	0.3856	0.2989	0.7029	0.2173	0.2583	0.4756	0.4644	**0.3644**
Mak	0.1946	0.4435	0.4028	0.4049	0.7018	0.3045	0.3775	0.5128	0.2668	**0.4010**
Vie	0.2866	0.4789	0.3716	0.3776	0.7018	0.3009	0.3775	0.5128	0.2497	**0.4064**
VieAll	0.3004	0.4736	0.3716	0.3776	0.7018	0.2430	0.3775	0.5128	0.2849	**0.4048**

Recall

Alg	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Avg
No TS	0.2500	0.5833	0.5194	0.3380	0.8503	0.3671	0.4306	0.5157	0.2778	**0.4591**
BL	0.2569	0.4167	0.4289	0.3796	0.8299	0.3188	0.2778	0.4214	0.3611	**0.4101**
Mak	0.2083	0.4444	0.4625	0.4259	0.8231	0.3865	0.4722	0.5723	0.2778	**0.4526**
Vie	0.2326	0.4861	0.4496	0.4491	0.8231	0.3865	0.4722	0.5723	0.2685	**0.4600**
VieAll	0.2361	0.4722	0.4496	0.4491	0.8231	0.3720	0.4722	0.5723	0.3148	**0.4624**

F-Measure

Alg	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Avg
No TS	0.1854	0.5632	0.4305	0.2436	0.7905	0.2830	0.3758	0.4743	0.2229	**0.3966**
BL	0.2256	0.3060	0.3934	0.3212	0.7575	0.2451	0.2454	0.4271	0.3481	**0.3633**
Mak	0.1594	0.4124	0.3886	0.3750	0.7541	0.3162	0.3970	0.5251	0.2515	**0.3977**
Vie	0.1773	0.4459	0.3772	0.3665	0.7541	0.3091	0.3970	0.5251	0.2377	**0.3989**
VieAll	0.1805	0.4373	0.3772	0.3665	0.7541	0.2861	0.3970	0.5251	0.2841	**0.4009**

Table 52 Global results of the event tracking experiments

Furthermore concentrated final results are shown for each temporal similarity algorithm:

Algorithm	Precision	Recall	F-Measure
No Temporal Similarity	0.42829	0.45914	0.39657
BL	0.36440	0.41014	0.36325
Mak	0.40102	0.45257	0.39769
Vie	0.40638	0.46001	0.39887
VieAll	0.40480	0.46239	0.40087

Table 53 Global performance of the temporal similarity algorithms

Finally a ranking was done with the temporal similarity algorithms to show how they performed against each other:

Ranking	Precision	Recall	F-Measure
1st	No TS	VieAll	Vie-All
2nd	Vie	Vie	Vie
3rd	VieAll	No TS	Mak
4th	Mak	Mak	No TS
5th	BL	BL	BL

Table 54 Final ranking of the temporal similarity algorithms performance

As can be seen from the table using no temporal similarity yields better precision but any temporal similarity but the baseline improves recall of the algorithm.

4.3.3 Event tracking and temporal similarity discussion

In the case of the discussion of the event tracking, first the temporal similarity has to be mentioned and discussed:

1. Temporal similarity discussion

The temporal similarity measure alone binds together documents with similar temporal profile together. This can be exploited for example in event tracking however temporal similarity by itself is a wide information retrieval topic worth of complete works by itself.

In the case of this work the temporal similarity is representative of the temporal profile of the documents, however it does not contain all types of named entities or all the cases that might provide useful information.

There were some temporal expressions left behind when computing the similarity namely open-boundary dates, proper way to address them and add them in the temporal similarity was not done in this work. The impact these temporal expression might carry is still a matter of discussion.

In the case of one of the algorithms proposed, durations and frequencies were measured and compared to be part of the temporal similarity, these expressions are known for adding noisy to the systems, however in this work they were considered to convey useful information and surprisingly this algorithm performed slightly better than the others.

The simplest baseline that could be tough of was to use the published date of the documents to compare the similarity, this gives indeed a baseline but in no way it should be considered as comparing the temporal profiles of the documents.

In the other hand Makkonen's algorithm took into account only dates and with a day granularity which did not completely adapted to the requirements of the temporal expressions in the corpus.

For that reason first the Vie algorithm was constructed changing the granularity of Makkonen's for minute granularity, enough to include most of the temporal expressions in the corpus.

And the algorithm Vie-All was built to take into account the durations and frequencies in the corpus.

Finally it is important to mention that the temporal similarity was calculated and then considered as a feature in the event tracking experiments, however different way to represent the temporal similarity as features was not tested, for example different similarity vectors for types of temporal expressions, etc.

2. Event tracking discussion

In the event tacking task there is a few remarks to do before jumping to the analysis of the results.

Regarding the experiment set up:

First the reasons why Java was selected as the main language is that Java is a very document language and there are several NLP tools and libraries available to help in the work, Python was also considered however many pieces of code were already available in Java and that's why the decision was done.

Second the reasons why the machine learning tool, Weka was selected are, because it is platform highly configurable and contains a graphical explorer in which to initially try the experiments, and in the other hand this experimented were easy ported to a Java program using the same Weka classes as the Weka explorer does.

Regarding the feature selection:

The first thing to point out is the extensive number of features, the keywords are a big concern given their numbers and the lack of a strong normalization dictionary to eliminate repetitive keyword mainly as bigrams and trigrams at the same time.

Also the named entities would be less far numerous as features if normalization of the entities was done before the actual mapping as features.

Al these things translate as said into large numbers of feature which in turn makes a multidimensionality problem in which machine learning algorithms are very sensitive.

Regarding the experiments:

The experiments were done actually testing both the performance of the actual event tracking and the impact of different or no temporal similarity features in it, effectively indirectly measuring the performance f each temporal similarity algorithm.

As shown in the methodology the configuration of the experiments was done in such way to also have track of the impact of named entities, keywords or both together in the event tracking experiments.

Now regarding the results:

First about the event tracking results per topic it is easy to notice that in different topics, algorithms behave completely different and in many cases the no temporal similarity has better precision, this can be for many reasons, first that the event "other" is very wide and can contain far more documents that any other events, second in other topics the durations and frequencies only occur in a handful of documents thus it does not show a considerable difference.

Furthermore the difference in each topic between temporal similarity algorithms is only due to a very small number of documents classified different probably just one or two, hoe ever given the small of the corpus, a single document classified different can be considered at least slightly representative.

Now regarding the overall results there are a few things to mention. First that the best precision was attained by the experiments with no temporal similarity involved, and the difference is very representative; the explanation for this is that the temporal similarity provides a temporal profile of the documents against each other which means that adds similarity not directly related to its content itself, it rather helps distribute the documents in time and thus more recall is achieved at cost of losing some precision.

The recall and F-measure achieved by the VieAll algorithm was unexpected, at a first glimpse it was thought that the frequencies and durations would provide far more noise than useful information however they proved to make an influence and made it the best algorithm of the four tested.

Finally the Vie algorithm proved to be a little bit better than Makkonen's both in precision and recall, meaning that the minute granularity was better suited for this particular task.

In the case of the final ranking, we can say that the Vie algorithm as slightly better precision than Vie-All, and that Makkonen's algorithm does not obtain a higher score than any of the Vie algorithms.

To finish the discussion of the results it is important to notice that first the ceiling of the system was at around 74%, and the best result of the experiments was 40%. The difference can be explained considering all the errors that are carried from the gathering of the corpus itself, the text not completely cleaned, all the annotations, classifications and normalization and of course the human and programming bugs. Even though during each step of the process carefulness was maintained to try to provide the better possible results, human errors were not avoidable. Also is very important to consider that for now the systems cannot perform better than the human evaluators.

Another important concern is the high class imbalance of the events and topics this was maintained trying to keep the experiments the closest possible to a real case scenario and not a perfect laboratory corpus.

5. Conclusion

The conclusion on the work done is described first in the context of the corpus gathering, then for the temporal similarity and then for the event tracking experiments.

Finally an overall conclusion is presented followed by the suggested future work in temporal similarity, event tracking and of named entities and temporal expressions.

5.1 Conclusion on the corpus and the annotation

Despite the fact of the human mistakes, the high classes unbalance in the topics and the events and all the trouble that annotation and classification causes, the corpus proved to be robust and a great starting point for the developing of this work.

The newspapers used as sources proved to be writing about the same topic and events during the days the corpus was gatherer, furthermore important events developed those days in the country giving plenty events and topics on which to experiment.

The annotation of named entities surpassed the expectations of quality, furthermore the great number of them provide a solid profile of what it is discus in the news piece, however normalization and further cleaning should be done to extrac the full potential of the named entities as features.

In the case of temporal expressions all types but the frequencies are well represented in the corpus, both classifications even though they are not perfect make a solid feature of the entities.

Furthermore the results of the normalization are respectable and the normalized temporal expressions may well be used for further development in the area. The normalization of temporal expressions proved to be a very complicated and time consuming task, explaining partially the reason why there are not reliably tools available for NER in Spanish.

In the case of classification in topics there is to notice that the human annotators did not perform extremely well in classifying the news pieces which gives a clear idea of the subjectivity of the tasks and that is highly dependent of the annotator knowledge and familiarity with the topics.

The event classification proved to be even a harder task for the annotators, given that all news pieces are on-topic and many news pieces describe two or more events in them, also the possible events where selected and given to the annotators which implies that the annotators have to stick to the given classes, this eventually proved to be the hardest problem to overcome.

Finally is worth mention that the reliability of the corpus, the annotators and the annotation are very good for the tasks they were given, effectively providing a robust corpus able to be used in further research on the field.

5.2 Conclusion on temporal similarity

Temporal similarity is a challenge by itself worth of entire works, however from the starting point of this work it can be said that temporal similarity does have an impact in event tracking, and that further work is needed to fully understand the temporal similarity, but concluding from what was done in this work:

The fact that not all temporal expressions were used makes the question that if these temporal expressions convey important information and if it does then how to measure their similarity.

In the case of the used algorithms, the baseline not fully represents a document profile but is a good indicator of how the published date influences the classification of the news pieces. Furthermore the baseline fulfills its function to provide a bottom performance of the temporal similarity to have something to measure it against

Makkonen's algorithm was the major guideline for all of the algorithms, which implies that in the corpus the less than day granularity of temporal expressions had low occurrences.

The algorithms *Vie* and *VieAll* were the new ones proposed, it was proven that a minor granularity as small as the normalization considers performs better, second that in some way frequencies and durations do have impact in the performance of the system.

Regarding the analysis of Makkonen's algorithm against the proposed ones it can be said that minor adjustment have to be made in Makkonen's algorithm to achieve the results of the other algorithms, and that even so with minor changes the temporal similarity could be improved.

A final conclusion regarding temporal similarity is that it is not conclusive that frequencies and durations improved the event tracking given the low amount of frequencies and durations in the corpus.

Regarding the aim of this project towards temporal similarity it can be said that it was fulfilled, not one but two new proposals of new algorithms were done, and both of them surpassed Makkonen's algorithm in the test, therefore the first aim of the project, to propose new and improved algorithms for temporal similarity was fulfilled.

5.3 Conclusion on Event Tracking

In the event tracking experiments several things can be concluded, the role of temporal similarity, the machine learning tools and evaluation of the results.

The machine learning algorithm used K-nearest neighbors proved to be somehow reliable even though it weakness to multidimensionality and the short set of training data during the cross validation.

The feature selection proved challenging trying to reduce the number of features as much as possible without losing the core features of the text. However even though there were several features the system seemed to perform as expected in reasonable training and evaluation time.

The sets of experiments even if they were designed to measure the performance of the temporal similarity algorithms, firmly express the performance of the system around 40% F-measure in the best case. The same kind of experiments would have been done regardless if analyzing the effects of temporal similarity or not, and in fact it was proven that not using temporal similarity improves precision.

Finally to say that the aims of the project regarding the event tracking were fulfilled as the system is capable classify news pieces into events, using or not temporal similarity features and other features.

5.4 Overall conclusion

The development of this project solved most of the research question and aims proposed in the beginning, the aims of this project are oriented not only to explore the temporal similarity algorithms that were proposed but to provide a whole platform for future research for future work.

A robust corpus was built and annotated which provided the base for the experiments. The annotation of said corpus was of good quality and was perform followed guidelines to achieve the best annotation possible.

In the case of temporal similarity the aim of the project was fulfill although not completely which was to propose new improved temporal similarity algorithms; even though better results were achieved that with Makkonen's algorithm, the small difference of the results it's not enough to consider it a decisive improval.

The same applies to the fact of using or not frequencies and durations in the temporal similarity, an improval is noted on the results but it is still inconclusive to state that frequencies and durations categorically improve event tracking.

In the explicit case of event tracking the results of the classification show a performance around 40% which is not that bad comparing it with the possible ceiling 70% and considering all the human errors and considerations of subjectivity discussed before.

The aim of developing a system capable of classifying news pieces was achieves with the said results providing a good starting for further research on event tracking, capable of using different types of features.

Furthermore the configurability of the system and experiments set up effectively makes this project a platform for both temporal similarity analysis and event tracking fulfilling the last aim of the project.

Finally the conclusion is that even though the project achieved considerable results it is still needed research and work on the field to improve these. The role of temporal expressions in event tracking is confirmed but not completely conclusive for improval of the event tracking. It is hoped that this platform provides an important tools or a guidelines for further development of the work of temporal similarity and event tracking

5.5 Future work

The development of this system proved to be far bigger than expected; the delimitation of temporal expressions, named entities, their classification and further normalization provides a big challenge to be resolved.

Furthermore work in temporal similarity and in event tracking is also highly susceptive to improval and will be a topic of research in the near future.

5.5.1 Future work on named entities and temporal expressions

In the case of the named entities, the big occurrences of the corpus could be used to train a machine learning system for delimitation and classification for Spanish especially for periodicity documents.

The same could be applied for temporal expressions delimitation, classification and normalization.

These hypothetical systems trained and develop using the corpus developed in this work could be important tools for Spanish for working with these expressions.

In the case of temporal expressions, more effort should be done in the complicated cases but that still convey important information and more work on recognizing the expressions that lacks information.

The temporal expressions normalization could be upgraded to a web-service to aid the normalization of temporal expressions in this and others lines of work.

Finally the annotated corpus itself proves to be a valuable tool for anyone interested in these kinds of expressions thus it should be normalized, annotated in xml and with the use of existing annotation standards to make it available.

5.5.2 Future work on temporal similarity

In the case of temporal similarity there a few lines of work proposed to be continued, these are:

1. A more detailed analysis of temporal similarity between durations and frequency expressions, also about expressions with open boundaries

   There is still doubt if frequencies and durations convey important information in the temporal similarity of documents and if so if this significant or not, the same for open-boundary expression, this hypothesis has to be researched.

2. Implement a measure to weight open-boundary expressions according to their boundaries and some distribution that represents them better

   Work has to be developed in accounting the open-boundary expression similarity, these could be done using some kind of distribution along time however the limits, the functions, and their weight have to be researched and applied to new temporal similarity algorithms.

3. Develop other ways to represent the temporal similarity between documents to be more appropriate used as features

   Temporal similarity is usually mapped to features to be used in event tracking or other machine learning tasks, the way to represent this similarity whether it is per expression or per document similarity is still unclear and should be researched to optimize the impact of temporal similarity in event tracking.

With that consideration the temporal similarity measure could be greatly improved or at least could be understood better and applied to practical work on the field.

5.5.3 Future work on event tracking

There are many lines of work regarding further development of the event tracking experiments. Even though the topic has been studied heavily it still prone to be upgraded, as new features, algorithms and preprocessing; all this should account for upgrade in the experiments:

1. Better cleaning and normalization of named entities and keywords

   Future work should include a better cleaning and normalization of the features specially named entities and keyword, many named entities refer to the same entity and a proper wither automatic or a dictionary method should be implemented to try to map these entities to their normalized ones, the same concept applies to keywords in other for them to prove to be solid features.

2. Others features should be explored

   Further other feature sets should be explorer, like information contained in the title of the news, the RSS metadata, the source website, reporter's data, etc. These might prove to be reliable features of future event tracking experiment.

3. Feature reduction implementation

   Even so that the features are carefully selected, still their number is very high specially accounting for such a small documents to be classified, which ends up in a multidimensionality problem which makes it harder to solve. For this matter the number of features should be reduced the most possible to try to avoid these kinds of problems at maximum.

4. Other machine learning algorithms

   Other lazy and not lazy learning algorithms should be trained and tried if they fulfill the criteria for these experiments, even though that the algorithm itself is not the most important things, with different algorithms and options better results could achieved.

5. A platform for more diverse experimenting

   An experimenting platform could be easily done with the systems already this project has, this could help to easily add and remove features, design experiments and adapt to the necessities of other projects.

6. Graphical representation

   A graphical representation either in a web application, a desktop application or a mere automatic generated image could provide instant graphical feedback of the event tracking systems, improving the understating of what the system does and how does it work.

A final remark would be that event tracking should be considered always for real case scenarios with real data and the most common mistakes, problems and solutions that it implies. Furthermore an ideal system should be able not only to classify documents in events but create the events automatically when the systems considers appropriate. The event tracking research to be done should be oriented to these real cases and made robust to be able to handle the huge challenge that event tracking posses.

References

1. [Ahn, Fissaha-Adafre., 2005]
   Ahn, D., Fissaha-Adafre, S., Rijke, M., *Extracting temporal information from open domain text; A comparative exploration.* J. Digital Information Management, vol3, no1, pp. 14-20., 2005.

2. [Ahn, et al., 2007]
   Ahn D., Rantwijk J., Rijke M., *A Cascaded Machine Learning Approach to Interpreting Temporal Expressions.*, University of Amsterdam, Netherlands, 2007.

3. [Allan, 2002]
   Allan, J., *Topic Detection and Tracking: Event-based Information Organization.*, Kluwer Academic Publishers, U.S, 2002.

4. [Allan et al., 1998]
   Allan J., Papka R., and Lavrenko V., *On-line new event detection and tracking.* In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval ACM Press, pp. 37-45., Australia., 1998.

5. [Arevalo et al. 2002]
   Arevalo M., Carreras X., Marquez L., Marti M., Padro L,. Simon M., *A proposal for wide-coverage Spanish named entity recognition.*, Procesamiento del lenguaje natural, ISSN 1135-5948, 2002, no 28, pp. 63-80., España. 2002.

6. [Atkenson et al., 1997]
   Atkeson C., Moore A., Schaal S. *Locally Weighted Learning.* Artificial Intelligence Review, 11(1–5), pp. 11–73. Special Issue on Lazy Learning. 1997.

7. [Benjamin et al., 2003]
   Benjamin, H. Gates, D. Levin, Lori. *From Language to Time: A Temporal Expression Anchorer.* 4th Workshop on Inference in Computational Semantics. Carnegie Mellon University, 2003.

8. [Bontempi et al., 1999]
   Bontempi G., Birattari M., Bersini H., *Lazy learning for modeling and control design.* International Journal of Control, 72(7/8), pp. 643–658., 1999.

9. [Chinchor, 1998]
   Chinchor N., *MUC-7 Named Entity Task Definition (version 3.5).In Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

10. [Cucerzan, Yarowsky, 1999]
    Cucerzan Silviu, Yarowsky D., *Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence.*. In Proc. Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora., 1999.

11. [Darwish, 2013]
Darwish Kareem., *Named Entity Recognition using Cross-lingual Resources: Arabic as an Example.,* Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1558–1567, Sofia, Bulgaria, August 4-9 2013.

12. [Erik, et al. 2003]
Erik F. Tjong Kim Sang and Fien De Meulder., *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognitio.,* CNTS - Language Technology Group University of Antwerp., Belgium, 2003.

13. [Ferro et al., 2000]
Ferro S., Sundheim B., Wilson G., *TIDES Temporal Annotation Guidelines - Draft Version 1.0. Technical Report Technical Report MTR 00W0000094, MITRE*, McLean, Virginia: The MITRE Corporation, 2000.

14. [Filannino et al., 2013]
Filannino M., Brown G., Nenadic G. *ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge.*, Manchester University. U.K. 2013.

15. [Friburger, Maurel., 2002]
Friburger N, Maurel D. *Textual similarity based on proper names.*, Laboratoire d'Informatique de Tours, Tours, France, 2002.

16. [Galicia-Haro, Gelbukh, 2009]
Galicia-Haro S-, Gelbukh A., Complex named entities in Spanish texts. Structures and properties. Faculty of Sciences, UNAM, Ciudad Universitaria Mexico City; Center for Computing Research, National Polytechnic Institute, Mexico City., Mexico, 2009.

17. [Grishman, Sundheim., 1996]
Grishman, Ralph and Sundheim, B., *Message Understanding Conference - 6: A brief history.*, Proc. International Conference on Computational Linguistics. 1996.

18. [Gómez, 2010]
Gómez, Daniel., *Sistema de detección y representación de expresiones de tiempo en textos no estructurados.*, Grade Thesis., Carlos III University, Spain, 2010.

19. [Heng, Grishman, 2006]
Heng Ji, Grishman, R.. *Data Selection in Semi-supervised Learning for Name Tagging*. In Proc. joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics. Information Extraction beyond the Document. United States, 2006.

20. [ISO8601:2004]
ISO8601:2004. *Data elements and interchange formats - Information interchange Representation of dates and times*, 2004.

21. [Kumaran, Allan., 2002]
Kumaran G., Allan J., *Text Classification and Named Entities for New Event Detection.*, Department of Computer Science, University of Massachusetts, U.S., 2002.

22. [Llorens et al., 2010]
Llorens H., Saquete E., Navarro B., *TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2*., Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics. 2010.

23. [Makkonen et al., 2002]
Makkonen Juha, Ahonen-Myka Helena, Salmenkivi Marko. *Applying Semantic Classes in Event Detection and Tracking*. University of Helsinki. Finland, 2002.

24. [Makkonen et al., 2003]
Makkonen J., Ahonen-Myka H., Salmenkivi M., *Topic Detection and Tracking with Spatio-Temporal Evidence*., Department of Computer Science, University of Helsinki, Finland, 2003.

25. [Makkonen, 2004]
Makkonen J., *Temporal information in Topic Detection and Tracking.*, Department of Computer Science, University of Helsinki, Finland, 2004.

26. [Mani et al., 2006]
Mani, I., Verhagen, M,. Wellner, B., Lee, C.M., Pustejovksy, J. *Machine learning of temporal relations. En:* Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL-06)., Sydney, Australia. pp. 753–760. 2006.

27. [Manning et al., 2008]
Manning C., Raghavan P., Schütze H., *Introduction to information retrieval.*, Cambridge Univ., U.k., 2008.

28. [Marrero et al., 2009]
Marrero M., Sánchez-Cuadrado S., Morato J., Andreakis G., *Evaluation of Named Entity Extraction Systems.*, Computer Engineering Department, University Carlos III of Madrid, Spain, 2009.

29. [Marsic, 2011]
Marsic Georgiana,. *Temporal Processing of news: Annotation of temporal expressions, verbal events and temporal relations.*, PhD Thesis, University of Wolverhampton, U.K., 2011.

30. [Masnizah, 2010]
Masnizah M., *Design and evaluation of an interactive Topic Detection and Tracking Interface*. PhD thesis, University of Strathclyde, U.K. 2010.

31. [Minkov et al. 2005]
Minkov E., Wang R., Cohen W., *Extracting Personal Names drom Email: Applying Named Entity Recognition to Informal Text.*, In Proc Human Language Technology and Conference on Empirical Methods in Natural Language Processing. 2005.

32. [Montalvo et al. 2007]
Montalvo S., Martínez R., Casillas A., Fresno V., *Bilingual News Clustering Using Named Entities and Fuzzy Similarity.*, URJC, UNED, UPV-EHU, Spain, 2007.

33. [Mota, Santos, 2008]
Mota C., Santos D., *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*: O Segundo HAREM., Online resource on Linguateca, Portugal, 2008.

34. [Nadeau, Sekine, 2009]
Nadeau David., Satoshi Sekine., *A survey of named entity recognition and classification*. National Research Council Canada., New York University., Canada, 2007.

35. [Nadeau et al. 2006]
Nadeau D., Turney P., Matwin, S., *Unsupervised Named Entity Recognition: Generating Gazetteers and Resolving Ambiguity*., In Proc. Canadian Conference on Artificial Intelligence., Canada, 2006.

36. [Nadeau, 2007]
Nadeau David., *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*,. PhD Thesis, Ottawa-Carleton Institute for Computer Science., Canada, 2007.

37. [NIST, 2008]
NIST: *ACE08 Evaluation Plan.*
http://www.nist.gov/speech/tests/ace/2008/doc/ace08evalplan.v1.2d.pdf., 2008.

38. [Narayanaswamy et al. 2003]
Narayanaswamy M, Ravikumar K., Vijay-Shanker K. *A Biological Named Entity Recognizer. In Proc. Pacific Symposium on Biocomputing*. United States, 2003.

39. [Pons-Porrata et al., 2003]
Pons-Porrata Aurora, Berlanfa-Llavori Rafael, Ruiz-Shulcloper José. T*emporal-Semantic Clustering of Newspaper Articles for Event Detection*. University of Castellon. Spain. 2003.

40. [Powers, 2007]
Powers, D., *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation.*, Technical report, School of Informatics and Engineering., Adelaide Australia., 2007.

41. [Pustejovsky, et al., 2005]
Pustejovsky et al. *TimeML: Robust Specification of Event and Temporal Expressions in Text.* [online pdf document] TimeML publications http://www.timeml.org/site/publications/timeMLpubs/IWCS-v4.pdf [last seen 7/12/2012]

42. [Rau, 1991]
Rau, Lisa F., *Extracting Company Names from Text.*, Proc. Conference on Artificial Intelligence Applications of IEEE. 1991.

43. [Ratinov, Roth, 2009]
Ratinov L., Roth D., *Design challenges and misconceptions in named entity recognition.*, Proceeding CoNLL '09 Proceedings of the Thirteenth Conference on Computational Natural Language Learning. pp 147-155., United States., 2009.

44. [Ritter et al., 2011]
Ritter A., Clark S., Mausam, Etzioni O., *Named Entity Recognition in Tweets: An Experimental Study.*, Computer Science and Engineering, University of Washington, United States, 2011.

45. [Saquete, 2003]
Saquete E., *TERSEO: Temporal Expression Resolution System Applied to Event Ordering.*, 6th International Conference, TSD 2003, České Budéjovice, Czech Republic, September 8-12, 2003.

46. [Saquete et al., 2005]
Saquete. E., Muñoz, R., Martínez-Barco, P. *Event ordering using TERSEO system.* Intelligence, no 2807, pp.220-228, 2003.

47. [Saquete, 2005]
Saquete E., *Resolución de información temporal y su aplicación a la búsqueda de respuestas.* PhD Thesis, University of Alicante., Spain, 2005.

48. [Sekine, Eriguchi, 2000]
Sekine S., Eriguchi Y., *Results and analyses of irex-ne.* In Proceedings of The Sixth Annual Meeting of the Association for Natural Language Processing. Japan, 2000.

49. [Shinyama, Sekine, 2004]
Shinyama Y., Sekine S.. *Named Entity Discovery Using Comparable News Articles.* In Proc. International Conference on Computational Linguistics., 2004.

50. [Smeeton, 1985]
Smeeton N.C. *Early History of the Kappa Statistic*. Biometrics Journal 41: 795., 1985.

51. [Steinberger, Pouliquen, 2006]
Steinberger R., Pouliquen B,. *Cross-lingual Named Entity Recognition.*, European commission – Joint Research Center., Italy., 2006.

52. [Song et al., 2011]
Song H., Wang L., Li B., Liu X., *New Trending Events Detection based on the Multi-Representation Index Tree Clustering*., I.J. Intelligent Systems and Applications,3 , 26-32., 2011.

53. [Toribio et al. 2010]
Toribio R., Martínez P. Pablo-Sánchez C., *Evaluation of Named Entity Recognition in Spanish with OpenCalais*., Procesamiento del Lenguaje Natural, ISSN 1135-5948, No 45, pp 287-290, Spain, 2010.

54. [Vicente-Díez et al., 2007]
Vicente-Díez Maria, Pablo-Sánchez César, Martínez Paloma. *Evaluación de un sistema de reconocimiento de expresiones temporales en español.* Procesamiento del Lenguaje Natural N° 39 pp. 113-120. Spain. 2007.

55. [Viera, 2005]
Viera A., Garret J., *Understanding Interobserver Agreement: The kappa Statistic*. Family Medicine (Fam Med 2005;37(5):360-3.)., 2005.

56. [Wang et al, 1992]
Wang Liang-Jyh, Li W., Chang C. *Recognizing Unregistered Names for Mandarin Word Identification.* In Proc. International Conference on Computational Linguistics., 1992.

57. [Wonsever et al. 2012]
Wonsever D., Rosé A., Malcuori M., Moncecchi G., Descoins A., *Event annotation schemes and event recognition in Spanish texts.* In Proceedings of the 13th interactional conference on computational linguistics and intelligent text processing., pp. 206-218., 2012.

58. [Yang et al., 1999]
Yiming Yang, Jaime Carbonell, Ralf Brown, Thomas Pierce, Brian T. Archibald, and Xin Liu. *Learning approaches for detecting and tracking news events.,* IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 14(4):32 – 43, 1999.

Appendix A. Named entities annotation guidelines

The following are the guidelines made for the annotation of named entities; these guidelines were given to the annotators to follow and were done considering the corpus and given that the annotators needed to have Spanish as their mother tongue, they were made in Spanish:

Etiquetado y clasificación de Nombres

Solicitamos que etiquete los nombres que designan diferentes tipos de entidades como de personas, lugares, instituciones, etc. y nombres de desastres naturales e indique el tipo de cada uno de ellos. Deberá encerrar las expresiones entre llaves: *{ }* y colocar la letra correspondiente al tipo de entidad después de las llaves.
Ejemplo:

{Luis Pérez}P
Donde "P" es el tipo de entidad en este caso una persona.

1. Nombres de personas (P)

Los nombres completos de personas, también nombres abreviados que refieran a esa persona. Se deben considerar nombres completos, apellidos, o combinaciones de estos. Por ejemplo:

{José Luis Pérez Rodríguez}P
{Pérez Rodríguez}P
{Luis Pérez}P
{Pérez}P

- Abreviaciones de nombres y acrónimos. Por ejemplo:

  {EPN}P -> Enrique Peña Nieto
  {AMLO}P -> Andrés Manuel López Obrador

- Los nombres de Instituciones, lugares, desastres etc. no deben ser etiquetados como nombres de personas.

  Escuela {Justo Sierra}P /Incorrecto
  {Escuela Justo Sierra}I /Correcto

- Si el nombre de persona viene acompañado con su acrónimo o abreviación, deben ser etiquetados juntos como una sola entidad. Por ejemplo:

  {Andrés Manuel López Obrador (AMLO)}P

- Si una persona es referida apenas con su apodo, un pronombre o su título personal, no debe ser etiquetada:

  {El presidente}P /Incorrecto de la república
  {El}P /Incorrecto comento el caso.

- Si el nombre está acompañado de su título personal, solo el nombre debe ser etiquetado:

  El presidente de la República, {Enrique Peña Nieto}P

Ejemplos de anotaciones correctas:

{Jesús Reyna}P
{Díaz}P
{Juan Díaz de la Torre}P
{Jesús Rodríguez Almeida}P
{EPN}P

2. Nombres de Desastres naturales (D)

Principalmente huracanes, tormentas tropicales y fenómenos meteorológicos

- Solamente se deberán etiquetar si tienen nombres propios.
- Si el nombre va a acompañado de el sustantivo "huracán" o "tormenta tropical" o similar, deberán ser etiquetados juntos

  {Huracán Katrina}D

Ejemplos de anotaciones correctas:

{Manuel}D se convierte en huracán
{Ingrid}D se intensifica
El {Huracán David}D toca tierra mañana
{Tormenta Mariana}D provoca lluvias a nivel nacional

3. Nombres de Instituciones (I)

Los nombres y siglas que correspondan a diversas instituciones de gobierno, empresas privadas, etc.

- Nombres completos, siglas, acrónimos, abreviaciones, deberán ser etiquetados

  {Cámara de Senadores}I
  {Segob}I
  {FIL}I (Feria internacional del libro)

- Entidades políticas que también son nombres de lugares, y su función en la oración es de una institución, no deberán ser etiquetadas como instituciones, si no como lugares. Por ejemplo:

  {México}I /Incorrecto voto a favor en las naciones unidas.
  {México}L /Correcto voto a favor en las naciones unidas.

- Entidades que son referidas por su nombre completo y su acrónimo o abreviación deberán ser marcadas como una sola:

  {Feria Internacional del Libro (FIL)}I

- Nombres comunes que refieren a una institución deberán ser etiquetados sí y solo sí son de crítica importancia para la noticia y para la oración en que se encuentren:

  El {gobierno federal}I y los {gobiernos estatales}I deberán demostrarle a la sociedad

- Anáforas y referencias a instituciones con pronombres o nombres simples no deberán ser etiquetadas:

  La {empresa}I/Incorrecto elaboró

Ejemplos de anotaciones correctas:

{SNTE}I
{CNTE}I
{Secretaría de Educación de Veracruz (SEV)}I
{Procuraduría General de Justicia del Estado (PGJE)}I
{Instituto Nacional de Bellas Artes}I
{INBA}I
{Secretaría de Gobernación (Segob)}I
{Escuela Justo Sierra}I

4. Nombres de Lugares (L)

Los nombres de lugares del país o del mundo, incluyendo los nombres completos, cortos, o abreviaciones)

- Solo deben etiquetarse nombres de poblaciones, municipios, estados, países, colonias, y otras entidades políticas, sus abreviaciones y acrónimos, pero no de lugares dentro de una población:

  Por ejemplo:
  {palacio municipal}L /Incorrecto

- Nombres de lugares que describen un accidente geográfico menor deben ser etiquetados

  {Río Bravo}L
  {Cerro de la estrella}L

- Si el nombre, es seguido de otro nombre de una entidad política mayor al primero, deberán ser etiquetados como uno:

  {Tixtla, Guerrero}L

- Igualmente si se encuentra acompañada de su acrónimo o abreviación

  {Chiapas (Chis.)}L

- En el caso de que los nombres se encuentren en una lista o que el nombre contenga el tipo de entidad política o lugar, también debe ser etiquetado junto con el nombre:

  {Municipios de Escárcega}L, {Carmen}L, {Champotón}L, {Calkiní}L y {Hopelché}L del {estado de Campeche}L

Algunos ejemplos de etiquetado correcto son:

{TUXTLA GUTIÉRREZ, Chis}L
{Chiapas}L
{OAXACA, Oax}L
{Quintana Roo}L
{Municipio de Benito Juárez}L
{CIUDAD DE MÉXICO}L
{Santa Cruz Xoxocotlán}L
{Municipio de San Lorenzo Cacaotepec}L
{Tepito}L
{Veracruz-Boca del Río}L

5. Otros nombres (O)

Nombres que no entran en ninguna de esas categorías, pero aún así son considerados de importancia.

Por ejemplo:

{IVA (Impuesto al Valor Agregado)}O
En la {escala Saffir Simpson}O

6. Inclasificables

Son aquellas que es evidentemente el nombre de algo, pero no es fácil clasificarlo en alguna de las categorías anteriores. En este caso únicamente en caso de que sea imposible reconocer a que se refiere el nombre por contexto y por ende no pueda clasificarse en alguna categoría, deberán ser ignoradas.

Por ejemplo:

(EFE) [Es imposible clasificarla]

Appendix B. Temporal expressions annotation guidelines

Similarly the following are the guidelines made for the annotation and classification of temporal expressions; they were also made in Spanish:

Etiquetado y clasificación de expresiones temporales

Requerimos que delimite y clasifique las expresiones temporales en el texto. Considere estos tres tipos de expresiones: fecha u hora (*DAT*), duración (*DUR*) y frecuencia (*FRQ).*
Deberá encerrar las expresiones entre llaves *{ }* seguidas del tipo de expresión temporal por ejemplo:

- *{El lunes}DAT por la tarde llegará a Panamá*
- *Esto ha sido así {por muchos años}DUR*
- *El pago de gas es {mensual}FRQ*

1. Expresiones de fechas o de hora (*DAT*)

Expresiones que se refieren a una fecha o una hora en particular (implícita o explícitamente): Algunos ejemplos:

- Si se encuentran expresiones del tipo: *{del 15 al 28 de septiembre}*, deberán ser etiquetadas como una fecha y no como una duración

Algunos ejemplos de etiquetado correcto son:

{15/05/2013}DAT
{Martes 15 de mayo de 2013}DAT
{El lunes}DAT
{El mes pasado}DAT
{Ayer}, {mañana}DAT
{Mañana por la mañana}DAT
{Antiguamente}DAT
{En un futuro}DAT
{15:23}DAT
{A las tres de la tarde con veintitrés minutos}DAT
{a las quince horas}DAT
{Cuarto para las cuatro}DAT
{Cuatro menos cuarto}DAT
{A la media noche}DAT

2. Expresiones de duración (*DUR*)

Expresiones que contienen una duración temporal.

Algunos ejemplos de etiquetado correcto son:

{Toda la semana}DUR
{Durante 3 años}DUR
{Por muchos años}DUR
Un viaje {de 2 semanas}DUR al Brasil
Pedro realizó {dos días}DUR de trabajo y después renuncio

3. Expresiones de frecuencia(*FRQ*)

Las expresiones que indican cierta frecuencia en el tiempo.
Algunos ejemplos de etiquetado correcto son:

{Todos los días}FRQ
{Diariamente}FRQ, {diario}FRQ
{Mensualmente}FRQ
{Cada tercer día}FRQ
{Cada quince días}FRQ
{En el primer Domingo de cada mes}FRQ

Excepciones:

No deben ser etiquetadas:

- Las expresiones de edad:
  Tirso Cruz Yuca, {de 46 años}DUR/Incorrecto

- Las expresiones de fecha que determinan un nombre:
  El ciclo escolar {2013-2014}DAT/Incorrecto

Appendix C. Temporal expression normalization tool user guide

A tool was developed to aid in the Timex normalization task, ahead is the instructions of said tool, including the guidelines of normalization of the temporal expressions, this tool and instructions were done in English:

Instructions for the Timex Normalization Tool (TNT)

1. Startup

To start the program run:

java -jar TimexNormalizer.jar

A) Input File

The input file is asked for the system at startup and ideally has to be in the same directory as the TimexNormalizer.jar file.

B) Output File (Normalization.txt)

The output file will be named the same name as the input file but with ending *DONE*

2. Initial Data

If the loading of the input file was successful some data will be displayed:
Num docs parsed: The number of Documents contained in the File and correctly parsed
A Full list of the TIMEX found in ALL documents
Total Timex: Is the total number of TIMEX found in the file.

Additionally if some Documents of the input file were already normalized and saved on the output file, the system will indicate it and skip them.

3. Automatic Save

During the Normalization process every time a document's TIMEX are finished, the program saves the progress in the output file.

4. Normalizing

A) Normalizing helping data

When a new Timex is addressed some data are provided to help the normalization:

Timex: #x of y
The number of the Timex of the total in the CURRENT document
Type: DAT|FRQ|DUR
The type of Timex as it was tagged.
Full line:
The full line in which the Timex appears.
Ref Date Readable:
The date of reference in which the document was published
Raw Timex:
The actual TIMEX to normalize

B) Input

Dates are divided in the start and end date of the actual Date, the user gives the granularity by inputting two dates beginning and end of the actual Timex.
There are two ways of input the normalized date, the complete one in which input of both ending and starting date is needed and the short way in which a Normalizer is used.

B.1) Complete way of normalize

First an input of the beginning date is needed in Format *YYMMdd HH:mm*.
Even if the date does need the *HH:mm* because of its granularity, The *HH:mm* indicate the system that this is a date which is going to be completely (beginning and ending) inputted by the user

Second the system will required the End date also in format *YYMMdd HH:mm*, after the input the Normalization will be completed

Example:

Start: 131022 00:00
End: 131022 23:59
Start: 131022 16:00
End: 131022 18:00

B.2) Simplified way

If the date can be normalized applying the Normalizers (D|W|M|Y) (Day, Week, Month, Year) in which the beginning date would be a given input date and the end will be automatically filled by the system.

The input will be in format: *YYMMdd(D|W|M|Y)*
For example:

131022D, 130101Y, 130201M, etc.

The system will automatically assign the ending date according the Normalizer used.

C) Classification

Also the user is required to enter a boundary classification of the Timex
The input consist of the two or three letters of the abbreviations (DD|ID|ODL|ODR|ODA) only

DD Direct Date
ID Indirect Date
ODL Open Date Left
ODR Open Date Right
ODA Open Date Ambiguous (open both sides)

After this the system automatically will move to the new Timex and after the last Timex of the document, to the next document.

5. Notes on Timex Normalization and common cases

- Dates, frequencies and durations that cannot be normalized at all I.E. The day the people rises. Should be marked with both dates as the first minute of 1950. I.E.
  - *19th of every month:*
    1. Start: *500101 00:00*
    2. End: *500101 00:00*
  - *The day the people rises:*
    3. Start: *500101 00:00*
    4. End: *500101 00:00*
- Durations and frequencies must be settled in the year 1950, to have a reference that the TE is not bounded in the timeline.
  - *19th of every month:*
    1. Start: *500101 00:00*
    2. End: *500101 00:00*
  - *The day the people rises:*
    1. Start: *500101 00:00*
    2. End: *500101 00:00*