RELATIONSHIP BETWEEN DETECTED EVENTS IN ONLINE MEDIA

Fabrício Raphael Silva Pereira

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Março de 2018

RELATIONSHIP BETWEEN DETECTED EVENTS IN ONLINE MEDIA

Fabrício Raphael Silva Pereira

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Jano Moreira de Souza, Ph.D.

Prof. Alexandre de Assis Bento Lima, D.Sc.

Prof. Sean Wolfgand Matsui Siqueira, D.Sc.

Prof. Eduardo Soares Ogasawara, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2018

Pereira, Fabrício Raphael Silva

Relationship between Detected Events in Online Media/Fabrício Raphael Silva Pereira. – Rio de Janeiro: UFRJ/COPPE, 2018.

XIII, 109 p.: il.; 29, 7cm.

Orientador: Geraldo Bonorino Xexéo

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2018.

Referências Bibliográficas: p. 79 – 109.

1. relationship between events. 2. unsupervised learning. 3. autoencoders. 4. event model. 5. event detection. 6. online media. 7. news. 8. short text. 9. neural networks. 10. deep learning. 11. text processing. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

A minha esposa Karina, meus filhos Felipe e Sophia, e toda minha família, que não mediram esforços para me ajudar e incentivar nessa etapa de grandes ondas e nós na minha vida.

Agradecimentos

Registro aqui meus sinceros agradecimentos a todos que contribuíram para que a conclusão deste trabalho pudesse ser alcançada com sucesso. Em especial, quero agradecer a:

Deus, pela permissão de viver e por todos os acontecimentos, que me proporcionaram crescimento e fortalecimento;

Meus avós, pelo provimento de um referencial moral, bases sólidas do que eu sou; meus pais, Rafael e Ivonete, pelo esforço na minha formação moral e intelectual;

Karina, minha esposa, pelo apoio incondicional e fundamental nesta conquista; meu filho Felipe que soube me esperar por esse longo período; e Sophia pelo último incentivo ao chegar logo no final dessa etapa;

Minha irmã, Anna Rafaela, pela oportunidade de aprimorar a personalidade através da convivência; toda minha família e da minha esposa;

Prof. Geraldo Bonorino Xexéo, meu orientador, não apenas pela orientação acadêmica e suporte, mas também pela agradável companhia e incentivos nas outras empreitadas que enfrentei nesse período;

Professores e alunos do CEFET-RJ no campus Maracanã, em especial o prof. Eduardo Bezerra pela valiosa contribuição com o direcionamento da pesquisa, o prof. Eduardo Ogasawara por sempre reiterar métodos para uma boa pesquisa, e ao prof. Gustavo Guedes por sempre me incentivar;

Professores, funcionários e alunos do IFES no campus de Alegre-ES que colaboraram para a minha empreitada, em especial o prof. Igor Costa que colaborou com a versão do capítulo de análise de eventos;

Professores da UFRRJ no Instituto Multidisciplinar por compreenderem minha posse tardia devido ao final dessa etapa;

Todos os meus amigos, incluindo os amigos da GPE, PESC/COPPE, LINE e LUDES, em especial o Fellipe Duarte e Marcelo Arêas pela amizade construída e incentivo mútuo para concluirmos o doutorado; todos os amigos com os quais tive a oportunidade de estudar ou trabalhar juntos, e que contribuíram para o meu aprendizado técnico e humano;

Todos os meus professores que tive no decorrer da minha formação, principalmente aqueles que desempenharam seu papel não só no aspecto intelectual, mas

principalmente no aspecto moral;

Sukyo Mahikari, por me mostrar que posso ser útil a Deus e à sociedade através do meu trabalho; amigos praticantes da Arte Mahikari, por me incentivarem a sempre colocar Deus em primeiro lugar, e também por terem sido os grandes incentivadores para a conclusão deste trabalho;

CNPq, CAPES, e FAPERJ pelo suporte financeiro durante a pesquisa;

Todos aqueles com quem tive algum contato durante esta etapa.

Meu sincero muito obrigado a todos!

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

RELAÇÕES ENTRE EVENTOS DETECTADOS EM MÍDIAS *ONLINE*

Fabrício Raphael Silva Pereira

Março/2018

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

As mídias sociais permitem que os usuários leiam, publiquem e compartilhem informações sobre eventos do mundo real. Existem várias técnicas para detectar ou descobrir eventos do mundo real a partir de notícias ou publicações online. No entanto, um problema que muitas vezes é ignorado neste contexto corresponde à descoberta de relações entre eventos. Um tipo particular de relacionamento é a semelhança entre dois eventos, que podem ser usada para organizar e filtrar o fluxo de informações fornecidas aos usuários. Esta tese apresenta uma abordagem para identificar relações de similaridade entre eventos previamente detectados em textos curtos das mídias *online*. Dessa forma, é proposto o *Autoencoder Neural Event Model* (*AutoNEM*), um modelo de rede neural não-supervisionado baseado em autoencoder para descobrir relações de similaridade entre eventos estruturados de acordo com o padrão de representação do *5W1H*. Este modelo atende à combinação de um conjunto de requisitos que até então não foram satisfeitos em uma única abordagem na literatura. O *AutoNEM* pode codificar eventos em um espaço latente, incluindo cada atributo 5W1H separadamente, o que permite a busca de relações de similaridade entre eventos através de suas novas representações. Os experimentos usam dados coletados do corpus de notícias *EventRegistry* para validar a abordagem proposta. Os resultados experimentais indicam que o modelo neural proposto para a detecção de relações de similaridade é efetivo, e ao comparar com alguns *baselines* também se demonstra competitivo. E ainda evidencia algum grau de similaridade em outros pares de eventos que não foram evidenciados pelos curadores manuais do *EventRegistry*.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

RELATIONSHIP BETWEEN DETECTED EVENTS IN ONLINE MEDIA

Fabrício Raphael Silva Pereira

March/2018

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

Social media allow users read, post, and share information about real-world events. There are several techniques to detect or discover real-world events from online news or posts. However, a problem that is often overlooked in this context corresponds to discovering relationships between events. A particular kind of relationship is the similarity between two events, which can be used to organize and filter the flow of information provided to users. This thesis presents an approach to identify similarity relationships between events previously detected in short texts from online media. Thus, it proposes the *Autoencoder Neural Event Model* (*AutoNEM*), an autoencoder-based unsupervised neural network model to discover similarity relations between events structured according to the *5W1H* representation standard. This model meets the combination of a set of requirements that have not been satisfied in a single approach in the literature. *AutoNEM* can encode events in latent space, including each *5W1H*-attribute separately, which allows the search for similarity relationships between events through their embeddings. The experiments use data collected from the news corpus *EventRegistry* to validate the proposed approach. The experimental evaluation indicates that proposed neural model for detecting similarity relationships is effective, and by comparing with some baselines is competitive too. The experiments also evidence some degree of similarity in other pairs of events that had not been evidenced by the manual curators of *EventRegistry*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Internet and mobile devices allow us to live in a world where information about events taking place all over the globe are published, shared and commented on in different online media, such as news sites, microblogs and others social networks (CHEN *et al.*, 2017; ZHU & OATES, 2013). In addition to publishing and sharing, the search and consumption of this information is also accentuated by the wide range of mobile applications that are multi-connected to online media, allowing users to interact at any point in their daily lives (LU *et al.*, 2017; MAHATA *et al.*, 2015). All this causes a flood of all kinds of information, in various forms of media (text, image, audio, video), and on various domains (WU *et al.*, 2017; YANG *et al.*, 2015). Such information overload can disperse the user's understanding of what happens around them, or hinder the consumption of information about events that may be of interest to the user (AL-SMADI *et al.*, 2017; ARAPAKIS *et al.*, 2014). Thus, users can fell the necessity to find or receive only news about specific events. Also, news portals can want to upgrade their algorithms to improve user engagement and enriching the consumer experience by suggesting similar news according to some interest events. So the research fields of Information Extraction and the Semantic Web have made several efforts to treat the data available in online media in order to facilitate news consumption and help in understanding events that happen in the real world (EL-KILANY *et al.*, 2017; IGLESIAS *et al.*, 2016; KIM *et al.*, 2013; MAZIERO *et al.*, 2014; ROSPOCHER *et al.*, 2016; UMAMAHESWARI & GEETHA, 2015; XU *et al.*, 2015).

Such kind of research began in a broader area, namely, Topic Detection and Tracking (*TDT*), which is basically concerned with investigating and inferring information about entities and subjects mentioned in online media. ALLAN *et al.* (1998); YANG *et al.* (1998) defined the problem of event detection and tracing as one of the problems within *TDT*, and used daily newspaper news as the main source of data. This niche has further aroused interest of users who directly consume information from online media, as well as from many sectors of society like private,

government, and scientific (BRAHA, 2012; CHEN *et al.*, 2012; JOHNSON *et al.*, 2011; STEPANOVA, 2011).

According to ALLAN *et al.* (1998); YANG *et al.* (1998), an *event* can be defined as *a relevant happening that has some characteristics, such as place and a time period in which it occurred.* On the other hand, a topic is nothing more than a word or expression that defines or summarizes a subject, or that defines an entity. This way, an event can be related to a topic, but they can not be treated as synonyms (KALEEL & ABHARI, 2015).

More broadly, the area of journalism, whose primary activity is the observation and description of events, seeks to report or describe an event by defining answers to the following questions, represented by the acronym *5W1H* (WANG *et al.*, 2007; WOLFE, 2010): "*WHAT*?", "*WHO*?", "*WHEN*?", "WHY?" and "*HOW*?'. The goal is to individually characterize an event by the answers to those questions.

Several researches have begun to treat the task of detecting events from information published primarily on social networks and microblogs (BECKER *et al.*, 2011a,b; NURWIDYANTORO & WINARKO, 2013). There is a special attention to microblogs (ATEFEH & KHREICH, 2015; HASAN *et al.*, 2017) because there is a high probability that people close to a place in which an event occurs publishes about it through mobile applications (LEE, 2012; SHI *et al.*, 2017), such as Twitter. Such information flow may help others avoid or approach the event in their interest (GU *et al.*, 2016; HASAN *et al.*, 2017), or later allow people on the other side of the globe to use the knowledge gained from past events to avoid problems when they face similar ones. However, this task is not trivial since in addition to a huge amount of short publications on diverse subjects, there are several possible ways to express the same meaning and a wide use of informal expressions (CHEN *et al.*, 2017)

Since it is easy for users to consume and share microblog information with their smartphones, virtually every news company publishes the titles of their news in microblogs. Indeed, several works have been studying the detection and processing of events reported in short texts (AL-SMADI *et al.*, 2017), generally grouping together several publications with some common features that evidence an event (ATEFEH & KHREICH, 2015; DOU *et al.*, 2012a,b; HASAN *et al.*, 2017), or in a few cases recognizing an event in each publication individually (SHI *et al.*, 2017).

1.1 Problem Definition

In order to support the computational treatment of events extracted from online media, several other researches have been concerned with the semantic modeling of events (SCHERP & MEZARIS, 2014; TZELEPIS *et al.*, 2016), relating attributes of each event (such as subject, entities, location, period or instant of time, among

others). In general, these works try to fill the answer slots in the *5W1H* for relevant events. More recently, given the growing importance of enriching the consumer experience of online media information in relation to what is happening in the real world and its interests, researchers have built models that try to identify relationships between events. According to SCHERP & MEZARIS (2014), "*human memory organizes experiences in events*", therefore providing an organization among the various events detected, or already known, can contribute even more to the user experience. Such organization can be achieved by relating them in several possible ways, which goes beyond the characterization or summarization of individual events (CUNNINGHAM, 2006; CUNNINGHAM *et al.*, 2005; PISKORSKI & YANGARBER, 2012).

Thus, a problem that is often overlooked in this context corresponds to *detecting relationships between events*. Figure 1.1 update the original view from CUNNINGHAM (2006); CUNNINGHAM *et al.* (2005); PISKORSKI & YANGARBER (2012) with recent researches (AL-SMADI *et al.*, 2017; ARAPAKIS *et al.*, 2014; CHEN *et al.*, 2017; FURLAN *et al.*, 2013; KIM *et al.*, 2013; ROSPOCHER *et al.*, 2016; SHI *et al.*, 2017; WU *et al.*, 2017; XU *et al.*, 2015; YANG *et al.*, 2015; ZHU & OATES, 2013) to portrays direction of the increase of difficulty in the problem of identifying relations between events.

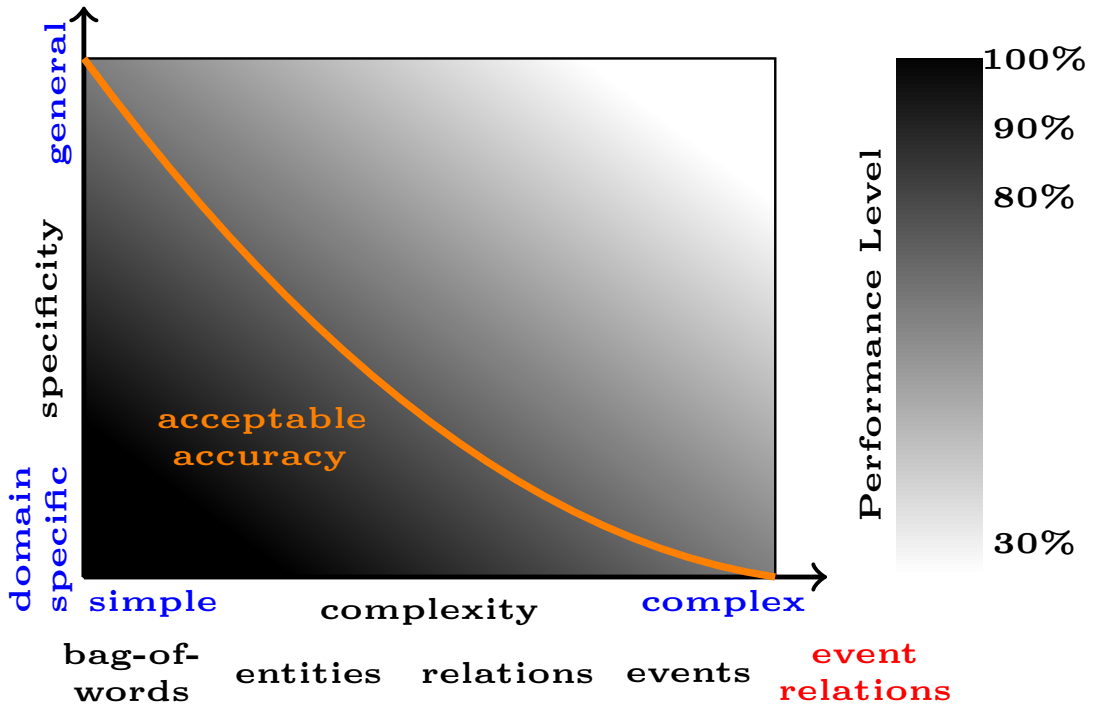


Figure 1.1: Information extraction performance trade-off relative to specificity and complexity – adapted and extended from CUNNINGHAM (2006); CUNNINGHAM *et al.* (2005).

A particular kind of relationship is the similarity between two events, which

can be used to organize and filter the flow of information provided to users (AL-SMADI *et al.*, 2017). This kind of relation also can improve the user engagement through links attaching within news articles according to the current item and the recent news read by the user (ARAPAKIS *et al.*, 2014). Thus, the identifying of this relation can contribute to the idea of Semantic Web by BERNERS-LEE *et al.* (2001), whose purpose is *"driving the evolution of the current Web by enabling users to find, share, and combine information more easily"* (XU *et al.*, 2015).

Although it is a type of simple relationship, it is still a difficult task for short texts (AL-SMADI *et al.*, 2017; FURLAN *et al.*, 2013) – this research defines a short text as a single textual sentence, e.g., news headline or tweets. Among several factors, this is mainly due to the small size of the text, which reflects in sparse or variable data, and the chaotic and huge amount of publications with short texts available for analysis (CHEN *et al.*, 2017; SHI *et al.*, 2017; WU *et al.*, 2017). The approaches often use other information contained in the publication beyond the text to circumvent such difficulties (SHI *et al.*, 2017; YANG *et al.*, 2015), or even they use extra sources of information (AL-SMADI *et al.*, 2017; GAO *et al.*, 2017; LIU *et al.*, 2016b), or delimit a specific domain for analysis (DOS SANTOS *et al.*, 2016; LU *et al.*, 2017). So, this work addresses the problem overcoming these difficulties, i.e, using only the available short text and from any field.

1.2 Aims

The general goal of this work is to discover similarity relationships between previously detected events in short texts from online media, in an unsupervised way (see Figure 1.2). For this, it proposes the *Autoencoder Neural Event Model* (*AutoNEM*), a autoencoder-based unsupervised model to discover similarity relations between structured events (*5W1H*) detected from short texts of online media.

The following specific goals must be satisfied to achieve the general goal:

- To detect events from short texts in online media, and to structure them with attributes according to the *5W1H* representation standard.
- To obtain semantic, compact and dense representations for each structured event and their attributes, embedding each event or attribute in a latent semantic space. This representation must allow a uniform way to process every event or attribute, and also must enable the search for similarity relationships between events.
- To evaluate the performance of *AutoNEM Encoder* to discovery degrees similarity between events against a golden corpus, and compare it with some baseline models.

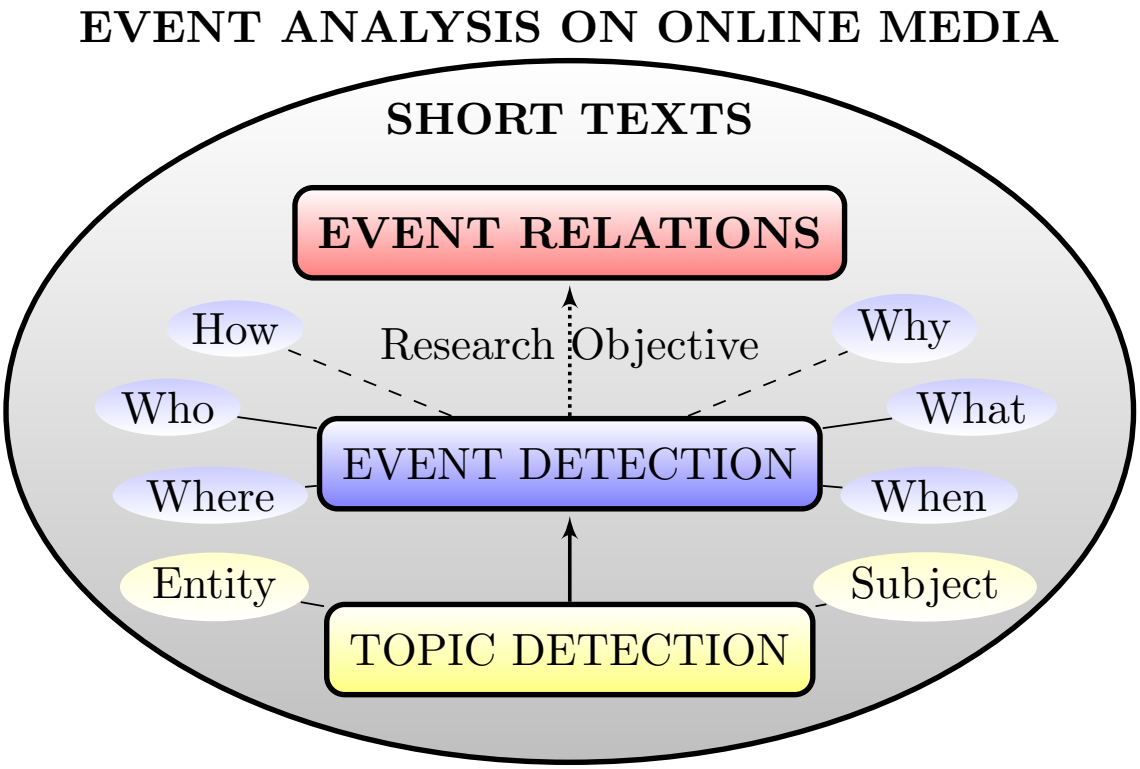


Figure 1.2: Research objective and limits within the area of event analysis on online media.

The main contributions of this work revolve around of entirely unsupervised learning that uses only short texts. Therefore, the proposal of this work arises from the following hypothesis:

General Hypothesis. *With the extraction of attributes (*5W1H*-based) of detected events in short texts from online media, it is possible learning semantic representations in a unsupervised way of each new event and its attributes, which allow the discovery of similarity relations between these new detected events.*

1.3 Contributions

The expected contributions of this work to the area of event analysis in online media are:

1. Identification of a limitations framework targeted at organizing previous works and frame them within it.

2. A neural network architecture based on autoencoder targeted at detecting relationships between events detected in short text from online media in an unsupervised way, called *AutoNEM* (*Autoencoder Neural Event Model*).

3. An unsupervised way to embed events and each of their attributes *5W1H*-

based in a latent semantic space through the encoder phase, called *AutoNEM Encoder*.

4. An evaluation of the degree of similarity between pairs of events representations obtained with *AutoNEM Encoder*, against a golden corpus (subset of *EventRegistry*, a news corpus grouped by unique events) and comparing with traditional methods.

5. An evaluation of the degree of similarity between pairs of events' attributes representations obtained with *AutoNEM Encoder*, against a golden corpus and comparing with traditional methods.

According indicated by LECUN *et al.* (2015), this unsupervised learning way can also contribute to the future of deep-learning with natural language processing.

1.4 Overview

The remaining of this thesis is structured as follows. Chapter 2 presents a review recent conceptual events models under their aspects, including relationships. Chapter 3 presents an overview of event analysis in online media, with the focus in short texts. Chapter 4 briefly describes how the deep learning area handles textual data, including autoencoders. Chapter 5 gives an extensive review of the recent literature dealing with the discovery of the relationship between detected events in online media, framing them under a set of requirements. Also provides a brief description of *EventRegistry*, the tool used to provide the golden corpus of events necessary to evaluate the learning architecture. Chapter 6 discusses the autoencoder-based network architecture used in this work. Chapter 7 presents experimental evaluation along with a corresponding analysis. This chapter also describes how this research built an evaluation data (golden corpus) based on *EventRegistry*, besides how the experiments used each baseline to compare. Chapter 8 concludes and discusses future work.

Chapter 2

Event Models

This chapter contextualizes the proposed event model in this work through the presentation and comparison of some conceptual models applied to various types of media (text, image, audio, video), and that allows the development of some reasoning about the events. This work includes both models already consolidated, as well as those that stand out in state of the art. The emergence of some research fields motivated the formalization of new event models, such as event detection and the processing and recognition of complex events. Besides, these models serve or have the potential to support several areas: sensors, signal processing, pervasive computing, computational context, contextual and/or social awareness, robotics, knowledge extraction, and semantic web.

Event models have the fundamental purpose of understanding both events occurring in a computational or virtual context, and the various documented or reported real-world events in virtual media in the form of text, photo, audio or video. Its purpose is not only to understand, but also extends to an increasingly efficient and effective way of organizing, sharing, retrieving and consuming content from events (SCHERP & MEZARIS, 2014; TZELEPIS *et al.*, 2016), and which are still challenges in this field. The present work focuses on real-world events, that is, events in which there is some direct or indirect human participation.

Chapter 1 mentioned that journalism tries to delineate a given event by describing a set of characteristics that respond to the interrogations known by the acronym *5W1H* (WANG *et al.*, 2007; WOLFE, 2010): *WHAT*, *WHEN*, *WHERE*, *WHO*, *WHY* and *HOW*. However, when dealing with different types of media besides text, having only the description of one or more events under each of these attributes, may not be sufficient to enable greater reasoning about events and provide a high level of interpretation about them (TZELEPIS *et al.*, 2016). Thus, to perform a computational treatment of real-world events and obtain results with a better interpretation level, the event models are being constructed, increased and evolved (YE *et al.*, 2015). In general, the models do not include all the attributes of the event

according to *5W1H*, but only some subset of them (WANG *et al.*, 2007). Also, some of the works began to model the events using other attributes (GONG *et al.*, 2004; SINGH *et al.*, 2004), and still with a different perspective of *5W1H* (JAIN, 2008; WESTERMANN & JAIN, 2006, 2007) – contemplating the aspects temporal, structural, informational, experimental, spatial and causal –, which was being reused and reviewed by SCHERP *et al.* (2009, 2012) – covering aspects participation, mereology, causality, correlation, and documentation.

In addition to modeling under these perspectives, some studies compared event models evidencing the evolution of the reach of these models, starting systemically by GKALELIS *et al.* (2010); SCHERP *et al.* (2009), after this comparison was incremented by SCHERP *et al.* (2012), and more recently by SCHERP & MEZARIS (2014); TZELEPIS *et al.* (2016); YE *et al.* (2015). YE *et al.* (2015) deals with both event modeling and context modeling. In the context modeling, YE *et al.* (2015) considers as requirements the data that synthesize the notions of *WHEN*, *WHERE*, *WHO* and *WHAT*, and for the modeling of events is made a comparison only with models that are in the form of ontologies, under a slight variation from the perspective of SCHERP *et al.* (2012). In a later work, SCHERP & MEZARIS (2014) presents a much more consolidated and comprehensive analysis of event models, including combining the various aspects of this perspective. Through this analysis, the event ontology called *Event-Model-F* is established as state of the art about event models. According to TZELEPIS *et al.* (2016), this state of art still stands as a conceptual model.

Given this scenario briefly explained, this chapter presents the event models following the analysis made by SCHERP & MEZARIS (2014), highlighting the points that have the potential to delineate the model used in the proposal of this work, such as the relationship between events, and including other models not contemplated by previous reviews. In the end, this chapter summarizes some challenges of the models and ends with the detailing of the *NEM* or *Neural Event Model* (DASIGI & HOVY, 2014) for having inspired the proposal of this work.

2.1 Analysis of Event Models

An interesting way of analyzing event models is to do so from some previous perspective on events. As mentioned, a first perspective different from that represented by *5W1H* was formalized by JAIN (2008); WESTERMANN & JAIN (2006, 2007), then evolved by GKALELIS *et al.* (2010), and finally revised by SCHERP & MEZARIS (2014); SCHERP *et al.* (2009, 2012). The initial perspective contemplates the following aspects of an event: *temporal*, *spatial*, *informational* (such as the type of event, or entities and actors involved), *experimental/empirical* (media

containing the captured, documented or reported data during the event, including the metadata), *causal* (causal relationships between events) and *structural* (composite relations of an event, i.e., events composed of sub-events). The intermediate perspective analyzes not only the event itself but also the existing models under the following aspects (GKALELIS *et al.*, 2010): *temporal*, *spatial*, *informational*, *experimental/empirical*, *compositional*, *causal*, *interpretation*, *uncertainty*, *formality*, and *complexity* (the latter two refer only to the models). While the perspective of SCHERP & MEZARIS (2014); SCHERP *et al.* (2009, 2012) considers the following aspects to characterize an event: ***participation***, ***documentation***, ***interpretation***, and relations of ***mereology***, ***causality***, ***correlation***. It still treats the ***temporal*** and ***spatial*** aspects implicitly, considering its relative and absolute representations. It is noteworthy that all these perspectives originated from the effort to model events extracted from multimedia content, mainly images and videos.

SCHERP & MEZARIS (2014) presented a comprehensive analysis within this last perspective under each of aspects and consolidated the requirements for *Event-Model-F*. The current study is a continuation of that. It adds others not mentioned works yet, and mentions essential points already listed for the scope of this work, and eventually adjusting some not clarified points until now. Furthermore, this analysis adds the following aspects: ***contextual*** and the relation of ***similarity*** between events. Table 2.1 presents an overview of the results and comparisons of the study made by SCHERP & MEZARIS (2014), and with the considerations of this work. This table arranges the models according to their nature, purpose or application.

The current research added some papers to this analysis because they complement essential aspects to the thesis, or because they are strongly related to the proposal, which deals with reported events in social media, especially if it is in textual form. The works added to the analysis are enumerated here, which correspond to the model names highlighted in Table 2.1. First, the Macro-Events were defined by CERVESATO & MONTANARI (2000) to support the temporal reasoning between the events and to group them when pertinent. This model is one of the foundations of the *STEEL* language, or *Spatio-Temporal Extended Event Language* (CHAUDET, 2006), which models disease outbreaks from medical reports. Recently it was partially extended to compose the framework established in WANG *et al.* (2014), and together with the concept of *Provenance Logic* aims to obtain the level of confidence of reporting an event, in contrast to others events reports. It is also considered the *TimeML* (PUSTEJOVSKY *et al.*, 2003; SAURÍ *et al.*, 2006) specification that is a markup language for temporal events in a textual document and complemented by MIRZA (2014); MIRZA & TONELLI (2014); MIRZA *et al.* (2014) to annotate causal relations present in a text, by adding the *CLINK* and *C-SIGNAL* annotations. WANG (2012); WANG & ZHAO (2012) presented an event model based on

Table 2.1: Adapting and updating the comparison of event models by SCHERP & MEZARIS (2014).

Models to Events Representation	Participation	Time		Space		Contextual	Relations				Documentation	Interpretation
		Relative	Absolute	Relative	Absolute		Similarity	Mereology	Causality	Correlation		
Logic Knowledge Representation												
Event Calculus (MUELLER, 2008)	–	●	●	–	–	❍	–	–	–	–	–	–
Situation Calculus (LIN, 2008)	–	–	●	●	–	❍	–	–	–	–	–	–
***Macro-Events* (EVANS, 1990)**	–	●	●	–	–	❍	–	●	–	–	–	–
***STEEL* (CHAUDET, 2006)**	–	●	●	●	●	❍	–	●	–	–	–	–
***Provenance Logic* (WANG *et al.*, 2014)**	–	●	●	●	●	●	❍	❍	❍	–	–	–
Ontologies												
Event Ontologies (RAIMOND *et al.*, 2007)	●	●	●	–	●	–	–	❍	❍	–	–	–
LODE (SHAW *et al.*, 2009)	●	–	●	–	●	–	–	–	–	–	●	–
SEM (VAN HAGE *et al.*, 2012)	●	●	●	–	●	❍	–	❍	–	–	–	–
SOUPA & CoBrA (CHEN *et al.*, 2005)	●	●	●	●	●	●	–	❍	❍	–	●	–
CONON (WANG *et al.*, 2004)	●	–	●	❍	●	❍	–	–	–	–	●	–
Situation Ontology (YAU & LIU, 2006)	❍	●	●	–	●	●	–	❍	–	–	–	–
SAWA (MATHEUS *et al.*, 2005b)	●	–	●	●	–	●	–	●	❍	–	–	–
STO (KOKAR *et al.*, 2009)	●	–	●	●	●	●	–	❍	–	–	–	–
Standards												
CIDOC CRM (DOERR *et al.*, 2007)	●	●	●	●	●	❍	–	❍	❍	–	●	–
CAP (OAS, 2010)	–	–	●	●	❍	❍	–	–	–	–	●	–
EventsML-G2 (IPT, 2012)	●	●	●	●	●	❍	●	●	–	–	●	❍
Video												
SsVM (EKIN *et al.*, 2004)	●	●	●	●	●	❍	❍	●	❍	–	●	–
CASE^E (HAKEEM *et al.*, 2004)	●	●	●	●	–	–	–	❍	●	–	–	–
VERL & VEML (FRANÇOIS *et al.*, 2005)	●	●	●	●	●	❍	–	❍	❍	–	●	–
Multimedia												
NMEE (APPAN & SUNDARAM, 2004)	●	●	●	●	●	❍	●	–	–	–	●	❍
GKALELIS *et al.* (2010)	●	●	●	●	●	❍	–	❍	●	–	●	❍
Eventory (WANG *et al.*, 2007)	●	●	●	●	●	❍	❍	❍	❍	–	●	–
Text												
***TimeML* (SAURÍ *et al.*, 2006)**	–	●	●	–	–	–	–	❍	–	–	●	–
***CLINK/C-SIGNAL* (MIRZA, 2014)**	–	●	●	–	–	–	–	❍	●	–	●	–
***NOEM* (WANG & ZHAO, 2012)**	●	●	●	●	●	❍	–	●	●	❍	●	–
***NEM* (DASIGI & HOVY, 2014)**	●	●	●	●	●	❍	–	–	–	–	●	❍
Systemic												
E (SCHERP *et al.*, 2008)	●	●	●	●	●	❍	–	❍	❍	–	●	❍
*E** (GUPTA & JAIN, 2011)	●	●	●	●	●	❍	–	❍	❍	–	●	❍
Event-Model-F (SCHERP *et al.*, 2012)	●	●	●	●	●	❍	–	●	●	●	●	●

● Supported. ❍ Limited, partial or indirect support. – Unsupported.

the perspective of WESTERMANN & JAIN (2007) and which, in a way, is complemented by semantic annotations related to a subset of *5W1H*. This model called *NOEM* (*News Ontology Event Model*) aims to describe the entities present in event news (in the Chinese language) and their relations, that is, its initial purpose was the modeling of events reported in textual form. Lastly, the *Neural Event Model* or *NEM* defines a structured event model to detect events considered anomalous (DASIGI & HOVY, 2014). The following sections discuss all these researches under each of the aspects.

2.1.1 Participation

The *participation* aspect refers to the presence of objects or entities (living or non-living) that play a role in the event. About this first aspect, the most of the models analyzed by SCHERP & MEZARIS (2014) provide support for the modeling of such objects. Models that do not support this support are the *Common Alerting Protocol* or *CAP* (OAS, 2010), of the models that have the purpose of representing knowledge, of the *TimeML* (PUSTEJOVSKY *et al.*, 2003; SAURÍ *et al.*, 2006) and its complementation with *CLINK/C-SIGNAL* (MIRZA, 2014; MIRZA & TONELLI, 2014; MIRZA *et al.*, 2014), and of the Situational Ontology (YAU & LIU, 2006) that partially covers this aspect. The models added to the analysis and that support participation are *NOEM* (WANG, 2012; WANG & ZHAO, 2012) and *NEM* (DASIGI & HOVY, 2014).

Here we already correct the analysis of the SCHERP & MEZARIS (2014), concerning the Situational Calculus (LIN, 1996, 2008; MCCARTHY & HAYES, 1969), since it does not contemplate the modeling of the entities that act on the events, but only the state of objects that undergo some action. It is easy to fall into this misunderstanding of the model, and several works confuse a property (which is the state of an object or situation) with an entity that acts on the event or suffers the event action.

2.1.2 Temporal and Spacial

The *temporal* and *spatial* aspects can be modeled by their absolute values or their relative concepts. As stated by SCHERP & MEZARIS (2014), it is must take some position in the face of philosophical discussions about the concept of an event confronted with the concept of objects (CASATI & VARZI, 2015; CHANDY *et al.*, 2007). Thus, in SCHERP *et al.* (2009, 2012) the following statement was adopted: events occur or happen at some instant or period while existing objects occupy some space, independent of the moment. In this way, the event has the temporal attribute directly, while the spatial attribute is linked to the event indirectly through the objects that participate in the event, i.e., through the location information of the event participants.

The present work is not intended to be limited in this way, allowing an event to have its spatial property directly, without necessarily depending on the participants. However, about these aspects, nothing changes on the works already analyzed by SCHERP & MEZARIS (2014). Regarding the works added to the analysis, *STEEL* (CHAUDET, 2006) made an extension on Event Calculus with spatial and temporal aspect modeling, and the *Provenance Logic* reused this proposal (WANG *et al.*, 2014). Macro-Events (CERVESATO & MONTANARI, 2000;

EVANS, 1990), *TimeML* (PUSTEJOVSKY *et al.*, 2003; SAURÍ *et al.*, 2006), and *CLINK/C-SIGNAL* (MIRZA, 2014; MIRZA & TONELLI, 2014; MIRZA *et al.*, 2014) only support the temporal aspect. *NOEM* (WANG, 2012; WANG & ZHAO, 2012) and *NEM* (DASIGI & HOVY, 2014) support both aspects.

2.1.3 Contextual (or Situational)

The studies from SCHERP & MEZARIS (2014); TZELEPIS *et al.* (2016) did not contemplate the *contextual* (or *situational*) aspect. According to context and situation definitions of YAU & LIU (2006), it includes the representation of any instantaneous, detectable and relevant properties of the environment, system, or objects of a situation that involves or is involved by the occurrence of an event. Although time and location may also form the context of a situation, this analysis highlights the contextual aspect of the previous two (temporal and spatial). Hence, the analysis considers supporting this aspect only if the event model in question enables the representation of other contextual properties in addition to time and space. Besides, according to WANG *et al.* (2007), a distinction must be made between the context of the description of the event and the context of the event. The description context of the event is related to the context involved by the event, i.e., the attributes of an event. The context of the event refers to the context that involves a specific event, i.e., other events that support the semantic understanding of the event in focus, therefore this aspect may be related to the response of the motivation of an event occurrence.

The models for the representation of knowledge about events almost always represent the context in some minimal way. *Provenance Logic* (WANG *et al.*, 2014) provides explicit context representation to allow a means of determining logical consistency between various event reports. In research of WANG *et al.* (2014), the provenance data collected by mobile devices at the time of reporting fed the context, such as speed, as well as the traffic situation on a public road. The *SOUPA & CoBrA* (CHEN *et al.*, 2003, 2005), *Ontologia Situacional* (YAU & LIU, 2006), *SAWA* (MATHEUS *et al.*, 2003, 2005a,b,c) and *STO* (KOKAR *et al.*, 2009) also support an explicit representation of the context. Most other models such as *E* (JAIN, 2008; SCHERP *et al.*, 2008; WESTERMANN & JAIN, 2006, 2007), *E** (GUPTA & JAIN, 2011), *NOEM* (WANG, 2012; WANG & ZHAO, 2012), *Event-Model-F* (SCHERP *et al.*, 2009, 2012), and *NEM* (DASIGI & HOVY, 2014) have a possibility of representing the context if such models are extended for this purpose, although it is not explicit for any contextual property.

2.1.4 Relationships of Mereology, Causality and Correlation

This section discusses three aspects of relationships between events (and only between events) already mentioned in SCHERP & MEZARIS (2014); TZELEPIS *et al.* (2016). Next section adds to the analysis an aspect of the relationship between events. Thus, according to SCHERP & MEZARIS (2014), the aspect of the mereological or hierarchical relationship refers to the relation between the parts and the whole when one has a composition of events. The causal relationship applies when a particular event can be the cause of another event, which would be the effect or consequence – such relation can also be a logical relationship. The relationships by the correlation between events occur when two or more related events are consequences of the same cause.

Among the models analyzed by SCHERP & MEZARIS (2014), support for these three types of relationships is mostly limited, in which only *Event-Model-F* supports them (SCHERP *et al.*, 2009, 2012). Similarly to the correction made on participation analysis of SCHERP & MEZARIS (2014), the Event & Situational Calculus do not contemplate the relationship of events between themselves, but rather between events and properties, conditions or situations.

Concerning the mereological relationship, SCHERP & MEZARIS (2014) already checked that *SAWA* (MATHEUS *et al.*, 2003, 2005a,b,c), *EventsML-G2* (IPT, 2012), and *SsVM* (*Semantic-Syntactic Video Modeling*), by (EKIN *et al.*, 2004) support it. This analysis verified that this aspect is also satisfactorily supported by *NOEM* (WANG, 2012; WANG & ZHAO, 2012), and well discriminated by *Macro-Events* (CERVESATO & MONTANARI, 2000; EVANS, 1990) and *STEEL* (CHAUDET, 2006). The *Calculus of Macro-Events* characterizes several types of meritorious relations, describing with formulas the following possibilities of event composition: sequence, alternative, parallelism, and iteration. The initial definition purpose from macro-events was the temporal domain relationship, and later, when formulating in the *STEEL* model, CHAUDET (2006) was able to use this definition by combining the temporal and spatial domains.

In addition to contemplating the modeling of event composition through macro-events, CHAUDET (2006) presented another modeling of hierarchical relations, both temporal and spatial. Such modeling is similar to that presented by *TimeML* (PUSTEJOVSKY *et al.*, 2003; SAURÍ *et al.*, 2006) and used in the definitions of *CLINK* and *C-SIGNAL* (MIRZA, 2014; MIRZA & TONELLI, 2014; MIRZA *et al.*, 2014), which are restricted to the temporal aspect. From this type of relationship, CHAUDET (2006) not only models but also provides much more advanced reasoning for events than that presented by the macro-event calculus in CERVESATO & MONTANARI (2000); EVANS (1990).

Support for the causal relationship is even more discreet, being satisfactorily present in the CASE^{E} (HAKEEM *et al.*, 2004), GKALELIS *et al.* (2010), *NOEM* (WANG, 2012; WANG & ZHAO, 2012) and the definitions of *CLINK* and *C-SIGNAL* (MIRZA, 2014; MIRZA & TONELLI, 2014; MIRZA *et al.*, 2014). Because *CLINK* and *C-SIGNAL* are annotations that complement *TimeML* tags, they are limited to events documented in the same text. The objective of MIRZA (2014); MIRZA & TONELLI (2014); MIRZA *et al.* (2014) is not only in the annotation itself but also in the automatic extraction of causal relations from the pairs of events detected in a text already annotated with *TimeML*. It is also important to emphasize that it is common to find differences between the studies in the way to treat this aspect. While some studies deal with the intra-document causality relation, i.e., it considers the presence of the causality relation within a single report of events (MIRZA, 2014; MIRZA & TONELLI, 2014; MIRZA *et al.*, 2014), others deal with the causal relationship between reports of various occurrences (SCHERP *et al.*, 2009, 2012). This work, just as in SCHERP & MEZARIS (2014), is concerned with the relationship between events reported in different publications.

The causal relationship between events can be generalized as a logical relation and is present when an event has logical properties for the occurrence of another event, and thus, this relationship strengthens the trust or credibility of the occurrence of the events in focus. For example, if there was an accident on the road, and next it detects a traffic jam by the presence of vehicles stopped or with low speed, then, it is remarkable that there is a logical coherence between the two occurrences. The relationship is modeled in this way by *Provenance Logic* (WANG *et al.*, 2014), using reported data by users and provenance data collected at the time of publishing to compose the information and properties related to each event. However, WANG *et al.* (2014) limited your model to a particular domain, similar to the example given. The extending made from Event Calculus is another limiting factor to this kind of relationship in WANG *et al.* (2014). It is due to the fact of considering as a *fluent* what, in fact, would be an event (in this case, a traffic jam), and only then could relate two events.

Finally, as verified by SCHERP & MEZARIS (2014), only the *Event-Model-F* (SCHERP *et al.*, 2009, 2012) has full support for the correlation relationship. Although WANG (2012); WANG & ZHAO (2012) assert the supports correlation by *NOEM*, at no time is this statement clarified or proven.

2.1.5 Relationship of Similarity

In addition to the three aspects of the relationship between events already discussed by SCHERP & MEZARIS (2014) and discussed previously, this work raises yet an-

other way of relating events to each other: by similarity. The similarity relation refers to the similarity between two or more events under some criterion, such as co-occurrence, co-location, and other attributes that can be extracted or discovered from the analysis of event documentation (topic/concept, profile/kind of participants, scope, etc.). Besides, in extreme case, this relationship may indicate that two events documented by two different sources refer to the same event, i.e., they are a unique event.

The *EventsML-G2* (IPT, 2012) and *NMEE* (*Networked Multimedia Event Exploration*), by (APPAN & SUNDARAM, 2004), support the similarity relation. Because *EventML-G2* is a markup pattern, it models the similarity using the `<sameAs>` markup and can be used to indicate that two events mentioned in distinct structures are the same event, for example. The *NMEE* (APPAN & SUNDARAM, 2004) defines ways to get the distance between two events documented from different users' points of view. The ways of obtaining the distances can be dependent or not on a specific user profile in front of the considered events. Independents ways are based on the temporal and spatial attributes of the media in the participants, as well as on the concepts or subjects extracted from the text of each event, and then it combines each of these distances in a single distance between two events. To obtain the distance dependent on the context of a user, it first gets the distance between each event and the context of the user. Then, the combination from this with the result of the independent distances allows computing a distance between two events relativized by the context of a particular user (APPAN & SUNDARAM, 2004). As the present work employees an event model that contributes to the processing and discovery of the similarity relation between events, Chapter 5 looks at more recent proposals that approach this goal concerning methods and techniques. Other models have limited or indirect support for similarity relationships, such as *Provenance Logic* (WANG *et al.*, 2014), *SsVM* (EKIN *et al.*, 2004), and *Eventory* (WANG *et al.*, 2007).

2.1.6 Documentation and Interpretation

The most trivial aspect is documentation, which consists of how to document an event. Documentation can be done through media such as photo, video, audio or text (SCHERP & MEZARIS, 2014), and can be further enhanced by annotations and metadata that complement both the description of the event and the context related to the event.

According to SCHERP & MEZARIS (2014), the following models contemplate this aspect: *LODE* (SHAW *et al.*, 2009), *SOUPA & CoBrA* (CHEN *et al.*, 2003, 2005), *CONON* (WANG *et al.*, 2004), *CIDOC CRM* (DOERR *et al.*, 2007;

SINCLAIR *et al.*, 2006), *CAP* (OAS, 2010), *EventsML-G2* (IPT, 2012), *SsVM* (EKIN *et al.*, 2004), *VERL & VEML* (FRANÇOIS *et al.*, 2005; NEVATIA *et al.*, 2004), *NMEE* (APPAN & SUNDARAM, 2004), GKALELIS *et al.* (2010), *Eventory* (WANG *et al.*, 2007), *E* (JAIN, 2008; SCHERP *et al.*, 2008; WESTERMANN & JAIN, 2006, 2007), *E** (GUPTA & JAIN, 2011), and *Event-Model-F* (SCHERP *et al.*, 2009, 2012).

Among the models added to this analysis, those that support the documentation are *TimeML* (PUSTEJOVSKY *et al.*, 2003; SAURÍ *et al.*, 2006) and, by extension, *CLINK/C-SIGNAL* (MIRZA, 2014; MIRZA & TONELLI, 2014; MIRZA *et al.*, 2014), since they tag the text to have an annotated documentation of the events. *NOEM* (WANG, 2012; WANG & ZHAO, 2012) also supports this aspect as it makes use of textual marks to identify elements that characterize an event. Therefore, beyond to supporting the process of extracting terms or expressions corresponding to some of the elements of *5W1H*, it also classifies them according to their semantic nature. However, this automated process requires a previously fed dictionary containing the semantic possibilities of the terms. *NOEM* also supports the documentation enabling the categorization of news in topics. Almost similarly, the *NEM* (DASIGI & HOVY, 2014) extracts short text arguments that correspond to the properties of *5W1H*, but it does so in an unsupervised way through the semantic paper annotation with *SENNA* (COLLOBERT, 2011; COLLOBERT *et al.*, 2011) – a tool considered as state of the art for the semantic role labeling task (SRL), whose semantic tags follow the style of *PropBank* (PALMER *et al.*, 2005).

The last aspect raised by SCHERP & MEZARIS (2014), called interpretation, is an attempt to make subjectivity permeable in all other aspects of an event. Just to make this aspect a bit clearer, it is essential to distinguish the interpretation of a source when reporting or recording an event, from the resulting interpretation from a report or record (manual or automatic) already made about an event. The present aspect may relate to any of the cited cases.

The interpretation is present in *Event-Model-F* (SCHERP *et al.*, 2009, 2012). It is limited in *EventsML-G2* (IPT, 2012), *NMEE* (APPAN & SUNDARAM, 2004), GKALELIS *et al.* (2010), *E* (JAIN, 2008; SCHERP *et al.*, 2008; WESTERMANN & JAIN, 2006, 2007), and *E** (GUPTA & JAIN, 2011). Partial support of this aspect may be considered in the *NEM* (DASIGI & HOVY, 2014) since the model is intended to detect if the reported event is anomalous or not.

2.2 Models Challenges

Finally, an important point, which is still a challenge among the various studies, concerns the automation of filling an instance of the model from existing docu-

mentation about a particular event, and independently of the application domain. Such researches usually present their experiments, or case studies, in a specific field and semi-automated form, or directed through structures previously prepared for a given application. About textual data, fully automated ways are emerging to at least instantiate the fundamental properties of a *5W1H* based model, like the *NEM* as done by DASIGI & HOVY (2014). In addition to the limitations that involve a gap between the conceptual model and the processing of the events perceived by TZELEPIS *et al.* (2016), another deficiency among most models is the definition of process or reasoning in order to detect some of the forms of relationships between two or more events. Chapter 5 discusses more challenges of methods and techniques.

Although *Event-Model-F* is considered state of the art as a model, there was a determined factor that motivated the **non-use** of it in this work: the model presents a high level conceptual, which can make infeasible the desired process over short texts. Unlike *Event-Model-F*, the *Neural Event Model* was shown to be more straightforward when approaching the attributes of *5W1H*, inspiring the current proposal although it does not contemplate the modeling of relations between events, and next section details it.

2.3 *Neural Event Model* (*NEM*)

The *Neural Event Model* or *NEM* (DASIGI & HOVY, 2014) aims to determine if an event reported in a short text is or not anomalous, through semi-supervised deep learning. Initially, each text is submitted to event recognition by extracting the semantic arguments through *SENNA* (COLLOBERT, 2011; COLLOBERT *et al.*, 2011) to compose the event in a structured way (*5W1H*), instead of plain text. The texts used for the experiment are news headlines from the AFE section of English Gigaword LDC2003T05 (GRAFF & CIERI, 2003) – for the training of regular events – and news headlines from the *Weird Events* section of *NBC* – for the training of anomalous events. *NEM* considers only texts containing the mention of a single event, and after a pre-processing identifies the terms of the collection. Before the training of each sentence, the terms are represented by semantic vectors obtained by an unsupervised learning algorithm called *GloVe* (PENNINGTON *et al.*, 2014) – such representations are known as *word-embeddings*, as well as *Word2Vec* (LE & MIKOLOV, 2014; MIKOLOV *et al.*, 2013a,b).

Once the entries prepared and separated by components of the event structure, the training begins in two stages, the first unsupervised and the second supervised. Its first stage, called for *argument composition*, is inspired by SOCHER *et al.* (2010, 2013a,b) and deals with recurrent and recursive networks. This stage performs unsupervised training to determine the best composition in a binary tree with the

recursive network between the terms of each semantic argument of the event to achieve a better vector representation of the argument. The second stage, called *event composition*, combines the representations of the arguments obtained in the previous step through another layer to get a representation of the event. So, this representation passes through the last layer, which classifying the event as anomalous or regular in a supervised way.

Hence, based on *NEM*, the present work seeks to make use of a simple structured event model to find out if there is a relation between several events reported in short texts published in social media. The next chapter focuses on the methods and techniques of event processing in these media.

Chapter 3

Event Analysis on Online Media

As introduced previously in Chapter 1, ALLAN *et al.* (1998); YANG *et al.* (1998) were the first to detect and track events when investigating events reported in the daily news. Such research field came out as a part of a wider problem named topic detection and tracking (TDT), which basically investigates and infers information about entities and topics mentioned in online media. As a result, research cases related to event analysis about online media have become more frequent, with several purposes and applications, about different data type in online media, developing a diversity of possible methods to be employed in different tasks performed in an analysis. Therefore, it is possible to measure this research field under these various dimensions or perspectives – as a matter of fact, they are permeated. On doing so, this chapter briefly presents an overview of event analysis about online media in the Figure 3.1. Most of the research assignments concerning these topics relate the events to the *5W1H* perspective (SZUCS *et al.*, 2013; WANG *et al.*, 2010; ZHAO *et al.*, 2014; ZHENG *et al.*, 2014), even though they implicitly start from some subset of the attributes this acronym represents (KHURDIYA *et al.*, 2012; KUMAR *et al.*, 2012). Section 3.1 presents a more detailed explanation of the research files under this first dimension.

The event analysis may have different applications (NURWIDYANTORO & WINARKO, 2013; STEIGER *et al.*, 2015): disasters, news, disease outbreaks and traffic. Apart from that, they do limitate themselves to the topics mentioned before. They may have more applications, such as the anticipation of civil unrest (COMPTON *et al.*, 2013). Considering the nature or the purpose of the event analysis, several pieces of research have shown different points of view, combining each other sometimes. This thesis points partially in the opposite direction from the tasks concerning the event's semantic XIE *et al.* (2008) discuss about. At first sight, it is more frequent to assume or to know previously the occurrence of the events, or at least some of their features (ZARRINKALAM & BAGHERI, 2016), which bring us to make a decision when it comes to a specific treatment of the events, as MAGDY

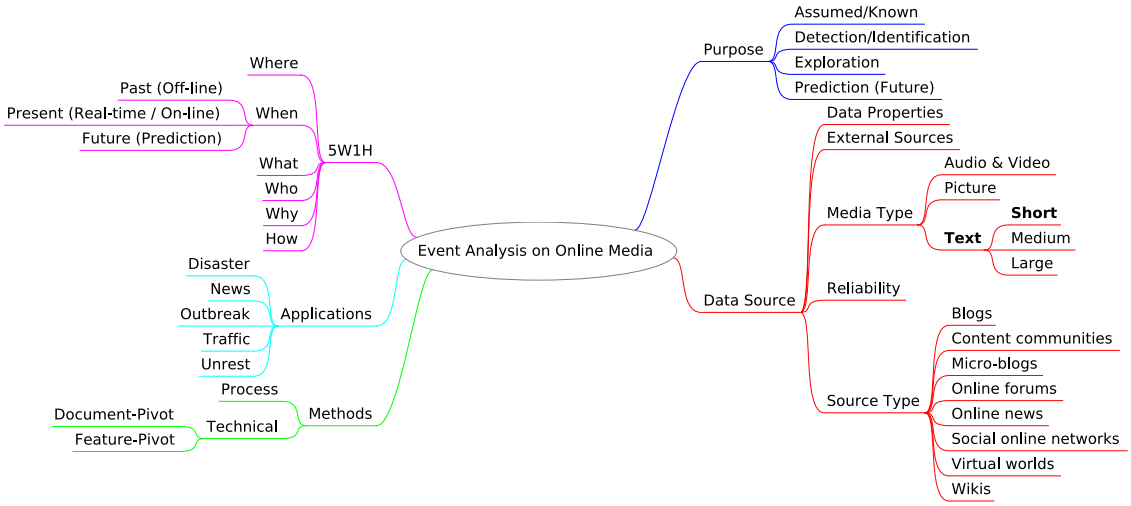


Figure 3.1: An overview of event analysis in online media.

et al. (2014a,b) suggest, pointing the search for information in microblogs from space-time queries. This kind of analysis is usually combined to the exploration of the features of past events (KANHABUA & NEJDL, 2013; SZUCS *et al.*, 2013). The most frequents resolutions in these research assignments are the identification or detection of events that occurred in the past (WANG *et al.*, 2010; WENG & LEE, 2011; ZHENG *et al.*, 2014), or are occurring in real time (ABDELHAQ *et al.*, 2013b; MIDDLETON *et al.*, 2014; WALTHER & KAISSER, 2013), accompanied by the prediction of events that may occur in the future (COMPTON *et al.*, 2013; RADINSKY & HORVITZ, 2013; WANG *et al.*, 2012a). Inside the event analysis area, this chapter focuses on detect events, once it is a necessary task to the purpose of this thesis: to find out relationships among events previously detected considering the extracted attributes. Therefore, the studies collected in this chapter were selected from a systematic search aimed at detecting events in online media that typically publish short texts.

Depending on the aims or the nature of the analysis, it is essential to choose or to develop an adequate method so as to get this purpose. The methods consist of both the processes by which different analysis tasks are applied in several stages, as well as the techniques in each type of process, or in each task. Concerning the techniques commonly used in this research field, ATEFEH & KHREICH (2015); WENG & LEE (2011) divided them in two types: *document-pivot* and *feature-pivot*. As the event analysis develops the treatment of various publications (documents or posts) of online media related to events, some research files choose to deal directly with some form of representing the document, i.e., the document itself is treated as a first instance resource of analysis, what characterize the techniques as *document-pivot*. Otherwise *feature-pivot* techniques use attributes extracted from different documents and grouped them in a suitable way, in order to do a direct analysis on

these attributes, regardless of the individuality of each document. The definition of these two kinds of techniques generalizes the original definition gave by WENG & LEE (2011). Furthermore, IMRAN *et al.* (2015) display a discussion with extensive coverage of the processes and techniques used to analyze online media data for mass emergencies. XIE *et al.* (2008) discuss several aspects on modeling and methods for event mining in multimedia data stream, including text. WEILER *et al.* (2015a,b, 2016) present and evaluate some methods characterized as the state of the art in real-time event detection on Twitter data.

Once the analysis is done on a dataset, it is necessary to know the nature of these data to choose and to apply proper methods, and thus to take advantage of the data extracted about the event. So, it is very important to consider the following questions:

- What properties does the data have? What information does the data consist of? Do they have metadata?
- Will any extra source of data be used to analyze the main data (such as a data dictionary)?
- What media type are the data to analyze (audio, video, image, or text)?
- Is the data accurate and reliable? What factors influence the uncertainty or reliability of the data?
- What kind of online media are the data extracted from?

DAUME *et al.* (2014) label online media into eight classes: blogs, content communities, microblogs, online forums, online news, social online networks, virtual worlds, and wikis. Each of them has specific characteristics, and an instance of online media may receive more than one of these tags, such as Twitter and Facebook, that are at the same time microblogs and social networks. Moreover, each instance operates with one, or more than one, media type. Instagram and Flickr, for example, support pictures with short textual comments or tags, and the data may be constituted by other metadata, like geolocation and time of the moment of the publication. Using this type of media and with certain properties of the data, CHEN & ROY (2009) deal with the publications of Flickr performing a spatial analysis to detect events. As this thesis focuses on short texts – as highlighted in Figure 3.1 – most of the studies listed in this chapter use microblog data, although there are also some studies that deal only with news headlines published online.

As for uncertainty, OZDIKIS *et al.* (2013) rely on the theory of *Dempster-Shafer* (SHAFER, 1976) to treat the uncertainty of Twitter data in order to estimate the location of events detected from this online media. GUPTA *et al.* (2012) analyze

and evaluate the credibility of events on Twitter and explores the possibility of detecting reliable events. WANG *et al.* (2014) deals with the reliability of the data, through the logical confrontation between them. Generally, microblog posts, like Twitter, are brief and limited, offering a great deal of room for uncertainty about the messages. Thus, the analysis on this data type becomes hard without an external support, and hence, researchers make experiments using the help of an external source (AGARWAL *et al.*, 2012; HUA *et al.*, 2013; SAKAKI *et al.*, 2012; SCHULZ *et al.*, 2013), such as a data dictionary (e.g., a dictionary of place names, a dictionary of synonyms, etc.), or even other online media (like news from online newspapers, or wikis). Some studies even cross information from various online media sources to analyze real world events (CROITORU *et al.*, 2013; TONKIN & PFEIFFER, 2013; WANG *et al.*, 2012b).

3.1 Characterization of Events Detected in Media with Short Texts

Due to the aims of this thesis, to analyze media events using short texts (fundamentally microblogs), and as most of the related research cases characterize the events taking some subset of the attributes symbolized by the *5W1H* acronym, this section analyzes these studies in accordance with this perspective. When appropriate, it raises some points referring to the other dimensions that are presented in the Figure 3.1. An example of media that is much mentioned is Twitter, which is not only a microblog but also a social network. The messages published on Twitter are the tweets, short publications limited to 140 characters, in which the user expresses themselves in natural language and often uses various forms of abbreviations due to the constraint imposed. Even so, this microblog platform has become very popular in recent years, with over 500 million tweets created per day[1]. Compared with news sites, event detection and extraction of their attributes in microblogs are much more challenging, once they present imprecise information and an informal language. ATEFEH & KHREICH (2015); HASAN *et al.* (2017); STEIGER *et al.* (2015); ZARRINKALAM & BAGHERI (2016) present these techniques with a different approach than that discussed in this chapter.

3.1.1 *WHERE* attribute

Twitter, like most microblogs, allows users to add geo-location information to tweets at the time of posting. Some studies display that tweets provide valuable real-time

[1]Fonte: `http://www.internetlivestats.com/twitter-statistics` (acessado em 15/01/2018)

information, such as pre-political analysis (TUMASJAN *et al.*, 2010), regional health monitoring (ARAMAKI *et al.*, 2011), or local emergencies detection (STARBIRD *et al.*, 2010). Most of these research files extracts characteristics of the special aspect related to some tweets based on the geographical marking (*geotag* or *geotagging*) to respond "*WHERE*?" of the *5W1H* (STEIGER *et al.*, 2015). However, according to some analyzes (ABDELHAQ *et al.*, 2013a,b; GRAHAM *et al.*, 2014), only between 1% and 5% of all tweets are explicitly georeferenced, that is, their set of metadata contains the geotag. Thus, without being able to determine a location of tweets, 95% to 99% of all tweets cannot be used for such intent. When allowed by the user, Twitter obtains this geographic information in two ways: primarily from the GPS, if it is enabled on the mobile device, and then this information will be very accurate; or through the triangulation of the network to which the device is connected, which returns a position without a high precision, but still distinct. HIRUTA *et al.* (2012); ROZENSHTEIN *et al.* (2014); SZUCS *et al.* (2013); WALTHER & KAISSER (2013); YUAN *et al.* (2013) use only tweets with a *geotag*, and HIRUTA *et al.* (2012) present a distribution of topics in these tweets.

Therefore, to advance in the search for this answer, WATANABE *et al.* (2011) attempt to assign geo-coordinates to non-georeferenced tweets increasing the chances of finding localized events. In order to accomplish this, the study performs a search for place names in the messages and counts the number of keywords that are related to an event that is connected to each place name. This process to find names of geographic references in a text and to assign them a geographic coordinate is called *geocoding* (IMRAN *et al.*, 2015). This method, however, is not well-succeed in finding the event localization when there is not a mentioned place in the set of tweets that contain these keywords (ABDELHAQ *et al.*, 2013a,b). Thus, SCHULZ *et al.* (2013) state that simple approaches are not enough or applicable to determine the location of a tweet. For example, location cannot be estimated using the IP address of a user's device, since neither Twitter nor the telecommunications provider allow access to this information to application developers (SCHULZ *et al.*, 2013). Another way to try to determine the location of the tweets is to estimate it from the location information present in the user profile, which can be obtained through the query provided by the *Twitter Search API*[2], but it is often incomplete and incorrect (HECHT *et al.*, 2011).

On the other hand, the extraction of location information only via texts published in a microblog is a challenge, primarily due to the limited nature of a tweet, which contains a very plain information about something (at most 140 characters) and often uses abbreviated expressions with many variations, which makes the task of named entity recognition very hard. The false positives that may exist and need to

[2] `https://dev.twitter.com/rest/public`

be somehow circumvented hamper the step named entity recognition, as in KRAFT *et al.* (2013). Another challenge, once it recognize the entities, is to meet the need to identify toponyms (name of places) among these entities, and to map them to real-world locations (SCHULZ *et al.*, 2013), since the mapping of a name to a location faces some problems (LIEBERMAN *et al.*, 2010): the same expression may refer to more than one location (for example, *'Rio de Janeiro'* can allude to the name of a state or to the capital of that state), or the same expression may refer to several types of entities (for example, the term *'Fluminense'* can designate the region of a state, or a soccer team). Then, a disambiguation of the toponym found must be performed in order to discover the location it refers to. Disambiguation and mapping are a great challenge, sometimes referred to as toponym resolution (LEIDNER, 2004) in the *Natural Language Processing* (*NLP*) literature. Thus, if on the one hand the information of the location can be explicit and thus possesses a naturally deterministic term, which allows an easy extraction of the location through the *geotag* or the detection of toponyms, on the other hand, this information can be implicit, so the extraction of the location will also depend on other contextual information (KHURDIYA *et al.*, 2012). Considering these points and the support of *NLP* tools, KHURDIYA *et al.* (2012) extracts event toponyms, but without determining a precise location (in terms of latitude and longitude). YUAN *et al.* (2013) estimates the location of a tweet by its affinity with its own content, and by the user's mobility history.

In view of these challenges, OZDIKIS *et al.* (2013) try to estimate the location of a group of tweets related to an event dealing with the uncertainty inherent in the data extracted from tweet content, user profile and *geotag* (when available). SCHULZ *et al.* (2013) go further, seeking to determine the geolocation of tweets, under a certain precision, from a combination of different approaches, and they present promising results. Each of the approaches combined by SCHULZ *et al.* (2013) are based on the following spatial indicators, that are possibly extracted directly from the tweet, metadata, or user profile: the tweet text message, the user profile location attribute, links or personal website, time-zone in the user's profile (when absent, time-offset can be used, which is less precise) and *geotag*. Thus, compared to other research assignments that focus on only one aspect, this multi-approach can respond to "*WHERE*?" with high accuracy, although it does not aim to detect events, but rather the disambiguation of toponyms. MIDDLETON *et al.* (2014) also present good results and provide a brief analysis of the representativeness of some location indicators related to natural disaster events. Language used in the publication is an information that can be added, distinguishing language and regional aspects of language (BOUILLOT *et al.*, 2012). Most of these indicators are also present in online news.

Once the location of a single publication is recognized, whether by *geotag* or another indicator, this information by itself is not enough to infer the existence of an event. It is important to combine it with other information available in the same post, or to group them with related information in other posts considering some attribute. A way to get this is to search for clustering among publications with close geographic information (ZHENG *et al.*, 2014), and the problem may be viewed as a graph, as confirmed in ROZENSHTEIN *et al.* (2014). Another form to solve this is to direct the event search using a location of interest, as in SHARMA *et al.* (2013), that join the extraction of this attribute with a target spatial search.

In addition, once the geographic location is obtained, one can work varying the granularity of this information as appropriate, through the hierarchical relationship between the localities (for example, street, neighborhood, city, state, region, country and continent), as well as DONG *et al.* (2015); KANHABUA *et al.* (2012b); ZHENG *et al.* (2014) apply it to detect localized events, or to infer the location of a tweet (BOUILLOT *et al.*, 2012). Hence, there are more possibilities to discover other information related to the events.

3.1.2 *WHEN* attribute

As it is not always possible to trust in inferring the occurrence of an event only through the information published in a single tweet, the searches generally resort to clustering information published in various tweets, as well as being able to combine them with information from the user profile to detect an event. In order to perform such clustering and combinations, the research cases apply several methods (probabilistic, data mining, machine learning, etc.) based on some principle to correlate the various tweets. The most relevant research files that use clustering, such as ABDELHAQ *et al.* (2013a,b); BECKER *et al.* (2011a,b); LAPPAS *et al.* (2012); LEE *et al.* (2011); SAYYADI *et al.* (2009); WALTHER & KAISSER (2013), show almost unanimously that one of the criteria that contributes greatly to the event detection is the "time proximity" around the moments that the publications were made. Depending on the study, this concept is referred to as "time locality" or "time window". Information about the moment of publication can be extracted directly from the timestamp of each post.

Having the factor of the "time locality" between the messages as aggregation criteria is interesting, since the messages related to the event have a high probability of being published within the same time interval (BECKER *et al.*, 2011a,b; LEE *et al.*, 2011), because this period can be close to the time the event occurred, and the number of posts related to the event decreases as time goes by (SAYYADI *et al.*, 2009; YANG *et al.*, 1998). Thus, the time dimension is a characteristic of

the event, and it advances in an answer to the "*WHEN*?" of 5W1H, and further encourages the discovery of other common characteristics through time clustering. In addition, the treatment of a stream of messages published in sequence through the time windowing is not only adequate, but also flexible, once the time window can be moved over this set of sequential posts, contributing to the detection of events from online media that present a continuous stream of information, as performed in LEE *et al.* (2011); SAYYADI *et al.* (2009); SZUCS *et al.* (2013); WALTHER & KAISSER (2013). GUILLE & FAVRE (2014) proposal, called *MABED*, also uses the idea of "time windowing", but it does it with static data.

HE *et al.* (2007); WENG & LEE (2011) work in a different way to find answers to "*WHEN*?". So as to identify and track events, they look at the time distribution of each term resident in a considerable volume of messages as a signal. In spite of this initial similarity, a remarkable difference between the two ways of investigating is the signal processing technique used in each of them. To reveal details about the time trajectory of terms, evidencing probable events related to terms that stand out in a certain period, HE *et al.* (2007) use the *Fourier transform*, whereas DONG *et al.* (2015); WENG & LEE (2011) use the *wavelet transform*. Another difference is the online media type used as data source in the experiments. HE *et al.* (2007) apply their algorithm to actual news streams, while WENG & LEE (2011) apply it to stream of publications on Twitter, which is a microblog. HE *et al.* (2007) method still identifies features that specify the recurrence or not of an event.

Another issue related to the time component of events detection is if they are on the past, on the present or on the future. That is, we can detect events that have already occurred, are occurring, or are likely to occur. Most of the studies cited so far try to detect events that have already occurred, and this type of detection can be applied to a set of static data as well as to a data stream. The research that intents to detect events in real-time applies their methods on a data stream, considering as their main source of time information the timestamp of each publication (STEIGER *et al.*, 2015), as in ABDELHAQ *et al.* (2013b); MIDDLETON *et al.* (2014); SAKAKI *et al.* (2010); TERRANA & PILATO (2013); WALTHER & KAISSER (2013); WATANABE *et al.* (2011) that somehow used "time windowing" over this stream. TERRANA & PILATO (2013) addresses the problem of event detection in a different way, treating the life cycle of an event as an electric field, and a tweet as an electric charge. This approach allows to apply mining techniques in short time intervals, providing an analysis of the life cycles of events in real time.

Nevertheless, if the goal is to try to detect a probable event that will occur, the information extracted from the timestamp is not enough to know when that future moment will be. To solve this, COMPTON *et al.* (2013) investigated the events that are in their planning phase, in which the participants or organizers use

microblogs as a tool for their organization, as civil unrest (SKINNER, 2011). So it is likely that, in these discussions of event planning, expressions related to the time it will happen occurs. From this scenario, COMPTON *et al.* (2013) extract time expressions from the text of the messages and turn them into information that responds when the event will take place, and for that, they combine these time expressions with the timestamp and apply *NLP* and *ML* (machine learning) techniques to identify and filter information. Thus, while COMPTON *et al.* (2013) focus on the task of early detection of events, ZHAO *et al.* (2015) clearly aim to predict future events. The extraction of time expressions, absolute or relative, is also useful for events that occurred in the past (SHARMA *et al.*, 2013; WANG *et al.*, 2010; ZHENG *et al.*, 2014). KHURDIYA *et al.* (2012) also consider the extraction of these time expressions relative to the context of the publication, so that to associate the tweet with a timestamp or an approximate time interval. KRAFT *et al.* (2013) explore in a practical way automated and parallel extraction to annotate time expressions in tweets and provide real-time analysis for past, emerging, or future events. Thus, depending on the objective within the attribute under focus, the nature of the process differs between real-time data analysis, which requires online processing, and a retrospective analysis of data already collected, which can be done offline (IMRAN *et al.*, 2015).

The research of KANHABUA *et al.* (2012a) follow a similar principle to the study of COMPTON *et al.* (2013) in relation to the time aspect, which evaluates the accuracy of the results obtained through the time identification of the events. However, this research file does not use any microblog as a source of information, but rather news about outbreaks and epidemics of diseases created by health organizations. Following their study, KANHABUA *et al.* (2012b) apply their technique on Twitter data, but they do not evaluate their results using these data, showing only the visualization of the time aspect of the extracted events. In KANHABUA & NEJDL (2013), they investigate the time diversity of tweets in periods of disease outbreaks around the world, changing occurrences of terms related to outbreaks of various diseases into time series in an attempt to find indicators for the detection of such events. According to their final evaluation, KANHABUA & NEJDL (2013) identified a high interconnection among topics associated to outbreaks and the time diversity of these tweets.

Additionally, as with geographical localization, it is possible to deal with the granularity of the time aspect (minute, hour, shift, day, month, year) and to get different combinations of information about this aspect (DONG *et al.*, 2015; ZHENG *et al.*, 2014).

3.1.3 *WHAT* attribute

Some research groups have successfully developed detection methods of topics in textual data stream in a time interval. Some popular microblog services, as Twitter, provide this information type. In a generic way, this attribute is extracted or discovered as the detection topics are, considering entities or terms that present a high token frequency (ABDELHAQ *et al.*, 2013a,b; DOU *et al.*, 2012b). In some studies, interest in a topic or its prior knowledge directs the event search (MIRANDA ACKERMAN, 2012), thus including the relevance factor of the events considering the topic (SHARMA *et al.*, 2013; ZARRINKALAM & BAGHERI, 2016).

Nevertheless, focusing only on this attribute, these efforts are not able to offer a high coverage of all themes related to real-world events, from a complete space-time point of view to satisfy users' information needs (LEE *et al.*, 2011). Even so, these research fronts are useful for clarifying what events are about or, at least, the related theme to them, although with the limited definition of event given by the topic (ZARRINKALAM & BAGHERI, 2016).

Despite the limited definition of this attribute, given only by the topic of initial form, some relevant studies initiate the exploration of events in online media starting from the discovery of this attribute combined with the time attribute, being able to follow up and summarize the evolution of probable events (LONG *et al.*, 2011; ZARRINKALAM & BAGHERI, 2016), or even making possible the detection of other attributes of these events. Other research cases explore this attribute after to cluster or classify the tweets under the time and spatial attributes, in order to the terms in each group have their frequency analyzed (KALEEL & ABHARI, 2015; KRAFT *et al.*, 2013; SZUCS *et al.*, 2013; ZARRINKALAM & BAGHERI, 2016). In addition to the frequency verification, other possibilities of recognizing this attribute are the extraction of verbs (CHANLEKHA & COLLIER, 2010; WANG *et al.*, 2010; ZHENG *et al.*, 2014), optionally accompanied by adverbs or prepositions also extracted, substantive expressions indicating actions or occurrences (WALTHER & KAISSER, 2013). These, and other studies (SHARMA *et al.*, 2013) do so, since they define this attribute as the description of change state in the occurrence of the event, so they focus on the extraction of verbs. In addition, WALTHER & KAISSER (2013) considers the use of a small dictionary of terms categorized into 13 categories of events (such as sporting events, entertainment events, musical events, congestion and violence) to assist in verifying the occurrence of an event.

Some studies, such as KHURDIYA *et al.* (2012), interpret this single attribute as the whole characterization of the event, that is, getting a response to "*WHAT*?" is to describe or summarize the entire event. Therefore, KHURDIYA *et al.* (2012) seek to determine a title to the event that is composed of the following textual elements:

subject (agent), action, object (which undergoes action), time and location.

3.1.4 *WHO* attribute

All of the above aspects are present in much of the work within the field of online media event analysis research, therefore the most common ways of working with them have been described. *WHO* attribute, which refers to those involved in an event, is not always explicitly contemplated in the research, although it is not uncommon to find it within the characterization of a topic, or strongly related to some detected topic (DOU *et al.*, 2012b). Even within studies that contemplate it, there is not a simple tendency in how to treat it, once an event can have different types of elements: objects or entities, agents or patients, or mere spectators. In addition, considering the classification of SCHUSTER *et al.* (2013), this aspect may refer to an individual (a personality or celebrity, an organization, a government, etc.), or to groups and communities (a set of individuals, a people, a nation, workers of an organization, sports teams, etc.).

In a platform like Twitter, this information can be extracted in a primary way from the profile of the users who published about an event (SZUCS *et al.*, 2013), by the relationships (direct or indirect) among the users (CROITORU *et al.*, 2013), or by mentions to users contained in the tweet (KRAFT *et al.*, 2013). Another way is to use named entity detection techniques and filter the entities that refer to the ones possibly involved, with the possibility of aiding from some data dictionary. Furthermore, the extraction of agents and patients involved in an action can be done by applying *NLP* tasks, such as *Named Entity Recognition* (*NER*) and *Semantic Role Labeling* (*SRL*), so as to take advantage of the syntactic and semantic structure of sentences (KHURDIYA *et al.*, 2012; SHARMA *et al.*, 2013; WANG *et al.*, 2010; ZHENG *et al.*, 2014). After extracting and clustering all the information, ZHENG *et al.* (2014) still perform calculations of correlation between the elements of distinct groupings to find the best combination between subject (*WHO*), verb (*WHAT*), and object (*WHOM*).

3.1.5 *WHY* and *HOW* attributes

The attributes "*WHY*" and "*HOW*" did not receive much attention in the literature yet, since they are not trivial, once the discovery and extraction of these two attributes is the most difficult task to characterize an event in online media, especially in microblogs (SZUCS *et al.*, 2013), and this information will not always be present. According to SZUCS *et al.* (2013), they refer to emotional information and can be treated with the sentiment analysis. Thus, SZUCS *et al.* (2013) respond to these two questions estimating the reason and the outcome of an event, but they

admit that this is not entirely satisfactory, requiring a more in-depth look at these responses.

In a way close to the extraction of some previous attributes, there is the possibility of syntactic recognition inside the sentences, and they may also be related to verbs and adverbs related to the occurrence of the event. SHARMA *et al.* (2013) propose the extraction of expressions related to "*WHY*" by the search for terms indicating the reason of the event and selecting the expression of the sentence that brings more confidence, although it is not detailed how this can be done, nor proven. The proposal of SHARMA *et al.* (2013) for the answer to "*HOW*" is to note the additional expressions in the sentences containing the "*WHAT*" and "*WHO*" attributes, stating that their presence is an indicator that the sentence contains some information on how the event happened, but it was not demonstrated and validated.

WANG *et al.* (2010); ZHENG *et al.* (2014) set the "*WHY*" extraction aside, mentioning only the "*HOW*" attribute, but no effective contribution is made to this attribute, for the reason that it simply equals it to the attributes previously discussed: "*WHO*" (with its "*WHOM*" variation) and "*WHAT*".

Eventually a text may contain expressions indicating the cause of an event and how it occurred. In a short text such as a tweet, however, it hardly indicates those attributes. On the other hand, in some situations the cause of an event may be another event, or the way it happened may depend on another event (XIE *et al.*, 2008). So how to detect the other event (if there is another event), and to state the cause of the previously detected event? The answer may be in complementing the analysis with other sources or publications. Thus, MIRANDA ACKERMAN (2012) extracts causal relations among topics of various news published in online newspapers, in order to make it easier for users to understand and navigate through the news. What make MIRANDA ACKERMAN (2012) approach efficient is the use of large processed texts, unlike what would happen if he used short texts, as texts of a microblog.

For the present work to reach its goal, it is necessary applying some technique or tool to extract the attributes of events from short texts in an unsupervised way. However, as discussed in this chapter, such task still has limitations, especially concerning the last two attributes ("*WHY*" and "*HOW*"). As event detection is not the focus of this research, the proposed solution defines a detection process by applying a tool for the extraction of some semantic roles as an attempt to capture *5W1H*-attributes (detailed in Chapter 6). Thus, a set of better techniques can substitute this tool for extracting each argument separately.

Chapter 4

Representation Learning of Textual Data with Deep Learning

Several studies have developed methods to learn new data representation of some media (image, video, audio, signs, texts, etc.) and for some purpose (BENGIO *et al.*, 2013; GOODFELLOW *et al.*, 2016; ZHONG *et al.*, 2016). The general purpose of the representation learning is to enable some algorithm to apply mathematical operations systematically and uniformly on the representations and thus reach a computational goal. Among the most common specific purposes is the reduction of dimensionality (VAN DER MAATEN *et al.*, 2009), whose most known precursor is the *Principal Components Analysis* (*PCA*), a linear and unsupervised method but which requires the context of the data (TIPPING & BISHOP, 1999).

RUMELHART *et al.* (1986) began studies for learning representations with neural networks in developing the procedure known as *backpropagation*. Thenceforth, with the ample and recent development of the area of deep learning, it made meaningful progress to learn hidden representations that collaborate with some complex reasoning (LECUN *et al.*, 2015) either with supervised, semi-supervised or also unsupervised learning techniques through *autoencoders* (KAMYSHANSKA, 2013; KAMYSHANSKA & MEMISEVIC, 2013, 2015). While most effort in the area of deep learning focuses on computational vision, other studies have developed ways of learning representations that incorporate textual data semantics (SALAKHUTDINOV & HINTON, 2007, 2009), especially with the use of *word-embeddings* (BENGIO *et al.*, 2001, 2003; MIKOLOV *et al.*, 2013a,b) and *recurrent neural networks* (LECUN *et al.*, 2015; SUTSKEVER *et al.*, 2014). Learning new representations from textual data is indispensable because traditional representations, such as the vector model with *TF-IDF*, produce a large data sparsity and do not incorporate textual semantics (CHEN *et al.*, 2017; DE BOOM *et al.*, 2016; SHI *et al.*, 2017; WU *et al.*, 2017).

As this work employs recurrent neural networks and an autoencoder-based

model, this chapter presents a brief theoretical description of these two deep learning techniques, followed by a small example of how to obtain textual data representation with their application.

4.1 Recurrent Neural Network (*RNN*)

Recurrent Neural Network, or *RNN*, is a class of neural network where there is a directed graph along a sequence formed by connections between units, so, it is suitable for processing sequential data as texts or time series (DE MULDER *et al.*, 2015; GOODFELLOW *et al.*, 2016; LECUN *et al.*, 2015). More friendly way, the *RNNs* works like humans think, in a continuous way, i.e., nobody has unconnected thoughts over time, at each moment the current thinking depends on the previous one. Thus, unlike *feedforward neural networks* (e.g., *convolutional networks*), the *RNNs* allow exhibiting dynamic temporal behavior for a time sequence.

Formally, a *RNN* map an input sequence with elements $\boldsymbol{x_t}$ into hidden states with elements $\boldsymbol{h_t}$ according Equations 4.1, where each $\boldsymbol{h_t}$ depending on all the previous $\boldsymbol{x_{t'}}$ (for $\boldsymbol{t' \leq t}$), $\boldsymbol{U} \in \mathbb{R}^{\boldsymbol{n} \times |\boldsymbol{x_t}|}$, $\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{n}}$ are weight matrices, $\boldsymbol{b} \in \mathbb{R}^{\boldsymbol{n}}$ is a bias vector, $\boldsymbol{n} = |\boldsymbol{h_t}|$ is the total number of units in the layer, and $\boldsymbol{f}$ is a activation function (DE MULDER *et al.*, 2015; GOODFELLOW *et al.*, 2016; LECUN *et al.*, 2015).

$$h_t = rnn(h_{t-1}, x_t) \tag{4.1a}$$

$$h_t = f(\boldsymbol{U}x_t + \boldsymbol{W}h_{t-1} + \boldsymbol{b}) \tag{4.1b}$$

As past contexts influence such networks, it means that their weights store information in that context. Effective RNNs have mechanisms (gates) to control what information must be stored (memorized) or replaced (forgotten). These effectives *RNNs* are called *gated RNNs*, and there are two kinds of them (GOODFELLOW *et al.*, 2016): *Gated Recurrent Units* (*GRU*) and *Long Short-Term Memorys* (*LSTM*). *LSTMs* are widely used and have separate gates to decide what to "remember" and what to "forget" (HOCHREITER & SCHMIDHUBER, 1997). *GRUs* are newer and have a single gate for both decisions (CHO *et al.*, 2014).

It also is possible classifying *RNNs* according to their design pattern, and they can make new designs when combined GOODFELLOW *et al.* (2016); SUTSKEVER *et al.* (2014):

- *Encoding RNNs*: it take a variable length list as input to produce a output (e.g., to encode a sentence as input);
- *Generating RNNs*: take a input to produce a list of outputs (e.g., to generate

words sequence as output);

- *General RNNs*: take a variable length list as input to produce a list of outputs, making predictions in a sequence (e.g., to translate a sentence from a language to another language); and

- *Bidirectional RNNs*: each hidden state is dependent on both past and future context (it can help the paraphrase identification task).

Beyond this *RNNs* designs above, there is too the *Recursive RNNs*. It' is another generalization to RNNs which connections between units are structured as a deep tree, enabling recursive computing (GOODFELLOW *et al.*, 2016). The area of natural language processing (*NLP*) has taken advantage of this type of RNN, since the sentences have hierarchical structures (DE MULDER *et al.*, 2015; PHONG LÊ, 2016; SOCHER *et al.*, 2010, 2013a,b; TAI *et al.*, 2015).

In practice, the *RNNs* compound the layer soon after the input to receive the textual sequence, in which each term is represented by an index or by a vector (one-hot representation or a semantic word-embedding). RNN also can form a layer soon before the output, when it generates out a sequence.

4.2 Autoencoder (*AE*)

An Autoencoder (*AE*) is a neural network trained in a unsupervised way to reconstruct its input to its output, and it can learn new representations of the data through some hidden layer that encode the input (GOODFELLOW *et al.*, 2016; KAMYSHANSKA, 2013; KAMYSHANSKA & MEMISEVIC, 2013, 2015). Basically, the autoencoders have two parts: an *encoder* function $\boldsymbol{h = enc(x)}$ (see Equations 4.2) that encodes the input $\boldsymbol{x}$ to a hidden representation $\boldsymbol{h}$, and a *decoder* function $\boldsymbol{r = dec(h)}$ (see Equations 4.3) that produces a reconstruction $\boldsymbol{r}$ from input $\boldsymbol{x}$.

$$h = enc(x) \tag{4.2a}$$
$$h = f(\boldsymbol{W_h}x + \boldsymbol{b_h}) \tag{4.2b}$$

$$r = dec(h) \tag{4.3a}$$
$$r = g(\boldsymbol{W_r}h + \boldsymbol{b_r}) \tag{4.3b}$$
$$r = g(\boldsymbol{W_r}f(\boldsymbol{W_h}x + \boldsymbol{b_h}) + \boldsymbol{b_r}) \tag{4.3c}$$

In Equations 4.2 and 4.3, $\boldsymbol{W_h}$ and $\boldsymbol{W_r}$ are weight matrices, $\boldsymbol{b_h}$ and $\boldsymbol{b_r}$ are bias vectors, and $\boldsymbol{f}$ and $\boldsymbol{g}$ are activation functions. As the decoder activation function $\boldsymbol{g}$ will reconstruct the input, so it must be a suitable function according the input properties (KAMYSHANSKA, 2013; KAMYSHANSKA & MEMISEVIC, 2013, 2015). Thus, an autoencoder tries learn the weights to fit $\boldsymbol{x = dec(enc(x))}$ by minimizing reconstruction errors through the loss function $\boldsymbol{L(x, dec(enc(x)))}$ (GOODFELLOW *et al.*, 2016). Usually, the loss function used in autoencoders is the squared reconstruction error, or *sum of squared errors* (*SSE*) (see Equations 4.4).

$$\mathcal{L}(x, r) = \|x - r\|_2^2 \tag{4.4a}$$

$$\mathcal{L}(x, r) = \sum_{1 \leq j \leq |x|} (x_j - r_j)^2 \tag{4.4b}$$

Deep autoencoders are typically nonlinear models, and there are some representatives variations of autoencoders (GOODFELLOW *et al.*, 2016; KAMYSHANSKA, 2013; KAMYSHANSKA & MEMISEVIC, 2013, 2015):

- *Undercomplete AE*: an autoencoder whose hidden dimension is smaller than input dimension, so it presents a bottleneck between encoder and decoder. Then, the undercomplete representation captures more salients features according to training data.

- *Sparse AE*: this autoencoder presents a sparsity penalty on the hidden units during the training, so it activates a small amount of hidden units for each input pattern. It learns representations from data training to use in another task, like in classification.

- *Denoising AE*: it learns to reconstruct the original input $\boldsymbol{x}$ from the noising input $\tilde{\boldsymbol{x}}$, i.e, this autoencoder learns the weights to fit $\boldsymbol{x = dec(enc(\tilde{x}))}$. The learned representations are robust to the noises in inputs data, so it can fix corrupted data or fill missing data.

- *Contractive AE* (*CAE*): an autoencoder with an explicit regularizer in its loss function to learn robust representations to small variations around of input values. Thus, this additional element in loss function penalizes undesired representations, prevailing only good representations.

Some recent studies (PHONG LÊ, 2016; SOCHER *et al.*, 2011a,b) have applied recursive *RNNs* in autoencoders to encode sentences from their syntactic structure in a semi-supervised way. For an autoencoder works with text, its first and last layers must have *RNNs* units. Beyond, for learning semantic representations is suitable to use word-embeddings in the input data.

4.3 Sequence to Sequence Learning

Several NLP tasks applying the previously discussed techniques, like topic or text classification, sentiment analysis, question answering, semantic role labeling, text summarization and language translation (DE MULDER *et al.*, 2015; LE & ZUIDEMA, 2014; LECUN *et al.*, 2015; PAULUS *et al.*, 2018; XU *et al.*, 2017). Usually, there is not an explicit use of learned hidden representations in such tasks. The explicit use of these hidden representations is more common in solutions that perform a sequence to sequence learning (CHO *et al.*, 2014; DAI & LE, 2015; LI *et al.*, 2015; SUTSKEVER *et al.*, 2014) to transfer learning between tasks (BENGIO, 2011, 2012; BENGIO *et al.*, 2013).

To express a simple example of the use of discussed techniques for representation learning of textual data, the Equation 4.5 defines a neural network architecture with two layers to learn representations from sentences (SUTSKEVER *et al.*, 2014). So, considering a sentence with the words sequence $(x_1, ..., x_T)$ and a set of pre-trained word-embeddings $\overrightarrow{x_i}$ for each word x_i in the vocabulary, a simple autoencoder with *RNN* computes the reconstruction from input sequence $(\overrightarrow{x_1}, ..., \overrightarrow{x_T})$ to the output sequence $(\overrightarrow{x'_1}, ..., \overrightarrow{x'_T})$, according to the following equation:

$$\overrightarrow{h_t} = sigm(\boldsymbol{U}\overrightarrow{x_t} + \boldsymbol{W}\overrightarrow{h_{t-1}} + \boldsymbol{b}) \tag{4.5a}$$

$$\overrightarrow{x'_t} = \boldsymbol{W'}\overrightarrow{h_t} + \boldsymbol{b_r} \tag{4.5b}$$

Thus, it is possible employing the hidden representation $\overrightarrow{h_T}$ from a sentence $(x_1, ..., x_T)$ in others tasks, like paraphrase identification (SOCHER *et al.*, 2011a).

Chapter 5

Related Works

The area of the relationship between events in different types of media has been developed both in event model, as methods for the discovery of such relationships TZELEPIS *et al.* (2016). SCHERP & MEZARIS (2014) developed a model, called *Event-Model-F*, to represent events reported at some types of media as text, picture or video, and it is the state-of-the-art of these models. It models essential attributes of an event as participant, time and space, besides of documentation, interpretation and three relations types: mereological, causal and correlation. Thus, this model has potential to provide developing of some reasoning between events detected with its attributes, although it has not direct support.

Despite the advances, TZELEPIS *et al.* (2016) list three challenges that still exist in this field of research. The first is finding for an optimal trade-off between the complexity of the event model and the performance of the methods for the processing of events. It then points to the high dimensionality of data that affects both storage and processing time. In a way related to the two previous ones, there is also the semantic distance between the meanings inferred by the automatic extraction of information and the human understanding. These challenges have been faced with deep learning techniques with significant results, like done by DASIGI & HOVY (2014) with the *NEM* (*Neural Event Model*). The *NEM* is a semi-supervised neural network, based on the works of SOCHER *et al.* (2010, 2013a,b), making use of an event model composed of semantic arguments (*5W1H*) to classify events reported in headlines considered anomalous. As stated in Chapter 2, the architecture model proposed in this work is inspired by *NEM*.

There are others works besides modeling events, further model relationships between events or even propose approaches to relate events, such as *MEP* ZHU & OATES (2013), *NEMo* DO CARMO *et al.* (2013) and *ECKGs* ROSPOCHER *et al.* (2016). As this work is focused on a simple model with arguments from *5W1H* and an unsupervised way to find relations between reported events at short texts from online media, so the following discussion is about works closer to this.

Previous works have proposed, directly or indirectly, to relate events, either by similarity or causality. In this section, we analyze these works under the perspective of the requirements presented previously. The research began with a systematic search for methods to discover relationships between events, whose query string brought 30 relevant results (from almost 500 references related to the area of computing) of the last 6 years (2012 to 2017) from the IEEE[1], ScienceDirect[2] and ACM[3] databases, in addition were found some referrals with ad-hoc queries.

To investigate whether and how the requirements were met by the literature, some critical aspects are listed in Table 5.1 to evidence the current limitations towards the purpose of this work. The first characteristics analyzed were: **(i)** regarding the mode of treatment of the instances to be compared that can be *flat text* or as a *structured event* (representation formed by attributes corresponding to *5W1H*); **(ii)** if the desired relationship is between textual (*textual relation*) or between known events (*event relation*), even if they have not been represented in a thoroughly structured way; and **(iii)** whether the learning method type is *unsupervised*. Other aspects revolve around the data and how they were treated: **(iv)** *cross-document* (or *inter-document*) is satisfied when the relationship objects belong to different documents; **(v)** *text only* when the process was restricted to the textual data; *inner only* if no other auxiliary data external to online media has been used in the process (such as dictionaries or taxonomies); **(vi)** *text length* examined in each document, indicating short (**S**, typically no more than one sentence), medium (**M**, a few sentences like one paragraph), or large (**L**, two or more paragraphs) text; and **(vii)** what online media was used as a *data source*.

Other characteristics were not mapped in Table 5.1, but are discussed in some works, when relevant: **(i)** the technique used can be *document-pivot* – if each document has individual features – or *feature-pivot* – when features are extracted from set of related documents (as a cluster by temporal or another aspect) – ATEFEH & KHREICH (2015); WENG & LEE (2011), and **(ii)** the *cardinality* between the documents and the events. These two characteristics are close to cross-document (or inter-document) from the Table 5.1, but the cross-document is more synthetic to show because when it is satisfied over short texts, it usually uses a document-pivot technique with an event by each document. Next, this chapter presents each referenced work with the analysis of aspects weighted in the table and other requirements (like domain independent or dense representation) when relevant to the contrast to the proposal of this work.

LEE (2012) presents a two-phase framework to analyze events reported on Twit-

[1] `http://ieeexplore.ieee.org/`
[2] `http://www.sciencedirect.com/`
[3] `http://dl.acm.org/`

Table 5.1: Summary of related work.

Referenced Work	Mode		Relation		Unsupervised Learning	Data & Tests				
	Flat Text	Structured Event	Textual	Event		Cross / Inter-Doc	Text Only	Inner Only	Text Length	Source
LEE (2012)	–	❍	–	●	–	●	–	●	S	Tweets
MEP (ZHU & OATES, 2013)	–	●	–	●	❍	–	–	–	L	News
SynRank (KIM *et al.*, 2013)	–	❍	●	–	●	●	❍	●	L	News
LInSTSS (FURLAN *et al.*, 2013)	●	–	●	–	●	●	●	●	S	News-SR
LEPA (ARAPAKIS *et al.*, 2014)	–	❍	–	❍	●	●	●	●	M	News
CST (MAZIERO *et al.*, 2014)	●	–	●	–	–	●	●	●	L	News-PT
E-FC (DEVI & GANDHI, 2015)	●	–	●	–	●	●	●	●	S	News
CDFL (YANG *et al.*, 2015)	●	–	●	●	❍	●	–	●	L	Misc
TFIR (MAHATA *et al.*, 2015)	●	–	●	●	❍	●	–	–	S	Tweets
Knowle (XU *et al.*, 2015)	●	–	–	❍	–	●	–	–	L	Misc-ZH
UMAMAHESWARI & GEETHA (2015)	–	●	–	❍	❍	–	●	–	L	News
GAO *et al.* (2015)	–	❍	–	❍	–	–	●	–	L	News-ZH
FERREIRA *et al.* (2016)	●	–	●	–	❍	●	●	–	S	Misc
ECKGs (ROSPOCHER *et al.*, 2016)	–	●	–	●	❍	●	–	–	L	News
NAVARRO-COLORADO & SAQUETE (2016)	–	●	–	❍	❍	●	●	–	L	News
LIU *et al.* (2016b)	●	–	❍	–	❍	●	●	–	S	Misc
RHNB (ZHAO *et al.*, 2016b)	●	–	–	●	–	–	●	●	S	News
LIU *et al.* (2016a)	–	●	●	–	❍	●	●	●	S	T&H-ZH
WEI *et al.* (2016)	●	–	–	❍	❍	●	–	–	L	News
STeller (ZHAO *et al.*, 2016a)	–	❍	–	❍	●	●	–	●	S	News-ZH
SAI-DbB-SLI (DOS SANTOS *et al.*, 2016)	–	❍	–	❍	❍	●	–	❍	I	Indexed
CTCBG (DRURY *et al.*, 2016)	–	❍	–	❍	❍	●	–	●	L	News-PT
EFS (IGLESIAS *et al.*, 2016)	●	–	●	–	–	●	●	●	L	News
LU *et al.* (2017)	–	❍	–	●	–	●	–	–	L	News
WU *et al.* (2017)	●	–	–	❍	●	●	●	●	S	News-ZH
EL-KILANY *et al.* (2017)	–	❍	●	–	❍	–	●	●	M	News
M²DN+ST (GAO *et al.*, 2017)	●	–	–	●	❍	●	–	❍	S	Weibo-ZH
HEE (SHI *et al.*, 2017)	–	❍	–	●	●	●	–	●	S	Tweets
AL-SMADI *et al.* (2017)	–	❍	●	❍	–	●	●	–	S	News-AR
CHEN *et al.* (2017)	●	–	–	●	–	●	●	●	S	Tweets
EventRegistry (LEBAN *et al.*, 2014)	–	❍	–	●	❍	●	–	–	L	News
AutoNEM	❍	●	❍	●	●	●	●	●	S	News

● Supported ❍ Limited/Partial – Unsupported

ter. The first phase performs an unsupervised grouping of features extracted from tweets, thus identifying events and significant features of each grouping (event). The second phase is supervised in order to evaluate the relationship between events under topics selected from the most significant features of each event. However, the paper itself states the experiments of this last phase didn't show such good results compared to the first phase.

ZHU & OATES (2013) relate events within the same story chain according to a specific topic or user query and also selects the two stories being the boundaries of the story chain. The extraction of the structure of the event is performed using external resources (ontology) to evidence semantic relations between entities and expressions. Thus, the method based on *Multi-dimensional Event Profiles* (*MEP*)

is semi-supervised, the data are news texts, and still uses some redundant news to improve results.

KIM *et al.* (2013) propose a framework composed of three phases to obtain a representation from a document. The first phase is the extraction of semantic roles by *SENNA* (COLLOBERT, 2011; COLLOBERT *et al.*, 2011) from each sentence in the form of triples (called frames) and creation of a network between terms and frames. It follows a clustering of frames with a similarity function (*SynRank*), and the most significant frame is extracted from each cluster. As the frame variations cause the sparsity, the proposal overcomes this by making use of the information contained in the network obtained in the first phase. Finally, the documents are represented by a set of frames representative of each cluster, and the paper ends by evaluating these representations in similarity and grouping tasks. The results indicate that the technique is better than previous works, but uses the whole contents of a document, that is, it uses more than one sentence to obtain such results. Besides, it does only with three pre-selected topics, and uses the timestamp of the documents in part of the experiment. At last, while mentioning events, it does not address the concept or modeling of events.

An approach called *LInSTSS*, proposed by FURLAN *et al.* (2013), aims to discover the semantic similarity between pairs of short texts (STSS) in the Serbian language but could be used in any other language. The principle is through use of vector representations of words obtained from matrix-based methods, such as COALS ROHDE *et al.* (2004) and RI SAHLGREN (2005), and then the similarity of sentence pairs is obtained by a combination of the best similarities of word pairs. The work suggests the necessity to increase the corpus (1194 sentence pairs, 70% for training and 30% for tests) to obtain more representative results.

ARAPAKIS *et al.* (2014) propose a system called Linker for *Events to Past Articles* (*LEPA*) to automate the identification of relationships between real-world event news, in order to reduce the work of experts in adding news links in online media and increase chances of consuming more news related to the event. The proposal is unsupervised, dispensing the use of external sources of information, and has two components: *indexer* and *linker*. The indexer works with medium-sized text (title and news abstract), creating an inverted index with term frequency. The linker proposes to search highlighted events (through syntactic analysis and with verbs in the past) within each news, and then search and rank news similar to each highlighted event. Therefore the relationship sought is not between highlighted events, but rather between an event and documents (news). To evaluate the system, news were collected from Yahoo News, from which 75 articles were selected in which it was possible to find prominent events. All of this news was automatically linked by the system, and also manually linked by Yahoo experts. In the end, the two

forms of link identification were submitted to potential users of online media so that they could evaluate carefully according to their perspective, thus to have feedback regarding the user's experience of reading and engaging the suggested links. To do this, several tasks were submitted to users through Amazon's Mechanical Turk, to obtain quantitative and qualitative assessments of automatic and manual links, thus assessing not only accuracy but also other measures derived from the coverage (F1-Measure). The results suggest that automatic link building is comparable to that created by experts, and its best application would be to reduce the effort of the experts, not to replace them. In spite of the excellent way of conducting the experiments and evaluation, the work leaves out in the expressiveness of the results due to the number of articles, which was limited by the necessity of manual evaluation of the links between them.

MAZIERO *et al.* (2014) renew a model called *CST* (*Cross-document Structure Theory*) that proposes to connect passages between different documents on the same topic. This renewal takes place on some fronts, and one of them is the re-categorization of the relationships that can exist between different texts (such as equivalence, contradiction, and translation) in order to reduce the subjectivity of the categories of relations there were previously. The work, therefore, does not explicitly deal with events, seeking a generic approach to texts, both from the perspective of content and structure. After the establishment of these forms of relations, the work proposes a supervised process for the discovery of relations on a corpus with whole news in Brazilian Portuguese, or by the use of classifiers, or by rules created manually, according to the type of relation.

DEVI & GANDHI (2015) propose to find similar sentences by improving the algorithm of fuzzy clustering (*Enhanced Fuzzy Clustering – E-FC*) since this algorithm has the advantage of capturing the relationship of an object in several classes. The method is evaluated in short news texts, however, the results are not expressive since it uses very few sentences, and does not make it clear how the classes were previously determined for the classification.

YANG *et al.* (2015) aim to learn features from multimedia data published in different online media, and for this purpose, they propose *Cross-domain Feature Learning* (*CDFL*) making use of denoising autoencoders. The learning is performed by crossing data from the same type of media (single-modal) from different sources (cross-domain), as well as crossing data of different media (multi-modal) from the same source (single-domain). The types of media are texts and images, collected from various online media with keywords related to a certain event already known. Although the paper claims its approach to be unsupervised, training with autoencoders is done in a targeted manner, matching data collected from a known event. So their approach was classified as semi-supervised. In addition, their approach is

evaluated in three tasks: sentiment classification, spam filtering, and event classification. Moreover, while learning text features and event images, the search did not bother extracting event attributes.

MAHATA *et al.* (2015) elaborates *TwitterEventInfoRank* (*TFIR*), which aims to rank tweets according to how informative they are about an event. There are two main steps in this approach. The first step, called *TwitterEventInfoGraph*, consists of discriminate which tweet is informative and which isn't for each event making a graph. Much information is considered here, including if a hashtag is popular or not, what clearly means it uses information from outside the text itself. The second step an interactive non-supervised approach scores vertices from the previous step, ranking them. It is tested in a big dataset manually collected with 3.8 million tweets.

XU *et al.* (2015) propose *Knowle*, a system that aims to organize hot topics from news events reported on the web based on the semantic relationships between the concepts/topics mentioned in each news. The method used, called *ALN* (LUO *et al.*, 2011), is supervised and applies it to words extracted from collected news, followed by deriving associations between web resources through mining association rules, and finally between hot topics. The data initially collected is news from events in Baidu News, which also provides words representative of events. With such words, other web pages are collected using Google search. For processing, both the text, the timestamp and an extra feature of Baidu are considered: words representing the events.

UMAMAHESWARI & GEETHA (2015) propose a semi-supervised bootstrapping approach to recognizing event patterns in news texts, using annotated seed news with properties and relationships from UNL (*Universal Networking Language*) from which new relationships are discovered – the method was abbreviated to *SSB-UNL* in Table 5.1. These relationships associate entities (or elements that fill semantic roles in events) with events, and through these primary relations also associate events with each other, however, they are events reported in the same document. Therefore, the objective of the work is to first discover events, and then evidence relationships between events reported in the same document. These two objectives are evaluated with the state of the art on complete news texts and conclude that UNL annotations along with the bootstrapping process can help to discover relationships between events, for example by inferring similar events.

GAO *et al.* (2015) temporally correlate predicate from sentences that report events, according to the order in which they appear in documents. The statistical correlation based on the co-occurrence of the predicates was improved with the use of expressions that indicate some type temporal relation and these terms were obtained from an external dictionary. Therefore, this approach resembles a semi-supervised method. Although events are treated atomically (by sentence), the set of sentence

sequences and their predicates are extracted from a single whole news in the Chinese language, in order to capture the temporal aspect between them.

FERREIRA *et al.* (2016) calculate the similarity between sentences (extending to summaries) using three layers (lexical, syntactic and semantic). The semantic layer extracts semantic roles annotations with the support of the *FrameNet* corpus RUPPENHOFER *et al.* (2016), but such semantic roles do not resemble the attributes of events (*5W1H*). The authors too state that the proposed algorithm is unsupervised. However, this statement is mistaken because the proposal makes use of an external corpus for the extraction of semantic annotations.

ROSPOCHER *et al.* (2016) propose an approach called *Event-Centric Knowledge Graphs* (*ECKGs*), which differs from traditional knowledge graphs by capturing and representing the dynamics of events occurring in the real world. For this, it applies current NLP and Semantic Web techniques to recognize, group and represent events, which process uses tools, metadata, and auxiliary corpus to capture semantics and relationships in long news texts, including between documents in different languages and different domains. The relationships caught between the events are both intra-document and inter-document.

NAVARRO-COLORADO & SAQUETE (2016) aim to order the events reported in various documents under the temporal aspect, by solving and grouping the temporal, lexical-semantic and distributional co-references – the method was abbreviated to *TC-LCV-DSC* in Table 5.1. First, the event components are extracted for each sentence in a document (with the help of available golden corpus tags). The temporal knowledge is extracted primarily within each document, and in a second moment the events co-referenced in several documents are grouped, so the cross-document temporal relation is inferred. Additionally, event co-references are extended to the same motto and/or synonyms – referring to *WordNet* (FELLBAUM, 1998; MILLER, 1995) –, and to the next words according to their distribution in topics, improving the results of temporal ordering. Although there is no explicit characterization in the article, the work can be considered as semi-supervised, since it makes use of ancillary information to the time in which it discovers distributional correlation on the corpus itself. The tests are conducted on a corpus of news divided under pre-selected subjects.

LIU *et al.* (2016b) does not intend to relate events, but rather sentences that characterize an event in a complete and relevant way. Although the goal is not to relate events, it is possible to get relationships between sentences that are selected, but other sentences that could be involved are discarded. Preliminarily, sentences are searched in online media from queries for keywords that are indicative of an already known event, and then the sentences are separated by events using third-party algorithms (considered state of the art). After this preliminary phase, the

proposed algorithm is applied over sentences, now represented by a limited set of keywords and separated by events.

ZHAO *et al.* (2016b) aim to extract relations of causality between two events reported in a single sentence. For this, the proposal uses the sentence's syntactic structure to categorize the usual causal connectives that are used as features for the proposed model, called *Restricted Hidden Naive Bayes* (*RHNB*). The model is trained on a corpus of news sentences with annotated causal relationships.

LIU *et al.* (2016a) propose to recognize contradictions between pairs of sentences (text and hypothesis) through the structuring of each sentence in a graph that relates the semantic attributes of the event, also taking into account the expressions of negations. The proposed model uses manual annotations to extract the semantic roles of the events so that the graph of each phrase is constructed, in which the nodes and edges construct features to be submitted to a classification supervised by the similarity of the sentence pairs. The result of the classifier is used in a second unsupervised step for the recognition of contradiction from the graphs of each sentence pair. The experiments are performed in annotated pairs of text-hypothesis in the Chinese language.

WEI *et al.* (2016) aim to extract and summarize events that are mentioned in several news texts under any specific topic and place, using keywords and an ontology with the locations to monitor. From this initial filtering each sentence is arranged as nodes of a graph, and vertices are inserted when two sections are consecutive in the same paragraph or belong to semantic (vocabulary) and temporary (publication date) documents. Then, heuristics are applied to extract from the graphs the sentences that best summarize a given event, generating a story line from event.

STeller, a process proposed by ZHAO *et al.* (2016a), aims at identifying stories in various news media (in Chinese) through the construction and clustering of a correlation graph from the semantic similarity (using *Word2Vec* with the dot product of Fisher Vector), temporal (timestamp) and co-occurrence (Jaccard) between short texts. To achieve the goal, as an intermediate step, a representation of an event (called a meme) is obtained for each published news item, whose attributes are the text itself, the timestamp of the publication and the context of that event (the work does not clearly define, but apparently would be other news published next). In this way, the events are not defined precisely in the form of semantic attributes that correspond to *5W1H*, and in addition, the ultimate goal is not to relate events but to identify stories in several published news stories.

DOS SANTOS *et al.* (2016) propose to perform three tasks: to identify similar and recurring events in a given region (the first method, called *DbB – Distance-based Bayesian Inference*), to evidence causal relations or the degree of influence between events (the second method, called *SAI – Spatial Association Index*), and

to identify possibilities of the sequence of some events to provoke a future event (the third method, called *SLI – Spatio-logical Inference*). However, the event is not characterized in a structured way, instead, the work characterizes as entities, so the event is a type of entity that has relations with other entities through semantic connections (verbs). The data is restricted to the domain of violent events, and the inputs are not pure texts like sentences, but processed and indexed data from Twitter and *GDELT* (*Global Database of Events, Language, and Tone*), in which indexes, tags, timestamp and locations are considered. In addition, it makes use of the *GSR* (*Gold Standard Report*, from IARPA) dataset to a reliable basis for events in part of the experiments. For these considerations on input data, the general process (the 3 methods together) was considered as semi-supervised, although the author did not make any categorization of his own method.

DRURY *et al.* (2016) propose the construction of a *Causal Topic Centred Bayesian Graph* (*CTCBG*). For this, the initial processing is done within each sentence in a supervised way, in which temporal information and causal relations are extracted by the presence of known lexical clues and or hard-coded list. Thus, although the work states that it detects cause and effect relationships between events, in fact, such links are between topics. From this, these relationships are mapped in a graph, where topics are aggravated in nodes, and cause and effect relations are accounted for and represented as directed edges. Also, a probability calculation of an event (topic) effect occurs in a given period, as well as the estimate of the probability of the type of each cause-effect relationship (positive, negative or neutral). The proposed method is centered on features (topics) founded in several documents, so each document is not treated individually. The experiments use long texts of Brazilian Portuguese news about sugarcane extracted from news sites of the agricultural sector.

IGLESIAS *et al.* (2016) proposes a classifier based on *Evolving Fuzzy Systems* (*EFS*) which the model updates its structure and parameters based on the content of data. This approach doesn't work with events and makes use of the news article's first paragraph, hence it's not working with short sentences, and has basically two main steps, term extraction and evolving classification. The method is tested in a huge dataset since it's intended to big data problems.

LU *et al.* (2017) propose a system with a support interface for experts to create semantic links between events reported in different collections of data, such as news. Since the system is intended to support the exploration of expert hypotheses (like a political science analyst) in the various stages of the process, the proposal can be considered supervised. Among the steps, the former use both *WordNet* to increase the representation of the same semantic concept and collections of validated and well-treated event data (*ACLED*), which is used as a golden corpus, analogously to

NAVARRO-COLORADO & SAQUETE (2016). Events are not characterized in a structured way, but rather by keywords that characterize them, in addition to the timestamp. Thus, the first semantic relation established is the similarity between the groups of these words, characterizing the events better, and at the last moment, it submits the events to *Granger causality tests*, also considering the temporal aspect. With this, the specialist can validate his hypotheses about the events and their possible relationships. The system works with news from one well-defined domain at a time.

WU *et al.* (2017) aim to summarize events reported in the news through a timeline, including news items that do not have temporal information but that are relevant to a particular query. The proposal works with news in Chinese, which are separated into sentences, and when possible are accompanied by the temporal information extracted from the sentence. Then, the sentences are related by grouping with words and by date, and the sentences without timestamp and considered relevant are embedded in the closest clusters already formed. In the end, some more relevant sentences are extracted to characterize the event of each grouping according to the initial query.

EL-KILANY *et al.* (2017) elaborates a non-supervised method that aims to discover relations between entities in a single event. For such a goal, besides the data collection, which ten websites are crawled in two different days, four steps are used. First, the entire body is used clustered in order to find events, then the sentences of each event are clustered. On next step, the important sentences of each cluster are extracted, ranked and used to find related entities. Finally, a relation extractor and generator step are used to generate these relations in a readable format. The proposed method was designed to use the whole news article, so it uses several sentences of the same news article. It has a good recall performance, but a poor precision performance, making an f-measure capable of competing with other state-of-art relation extraction methods, even though the little documents used for training. However, looking at the whole process, one realizes that in fact the method is semi-supervised since it uses unsupervised tools (*Stanford Named Entity Tagger*) for the extraction of entities.

The work of GAO *et al.* (2017) proposes to classify events mentioned in microblogs through a framework, called M^2DN+ST, with two stages. In the first stage (M^2DN, or *Multi-modal Multi-instance Deep Network*) an off-line classification training in texts and images with a deep-learning architecture is performed. In the second stage (ST, or social tracking) additional information retrieved from online media is used to enrich the classification with the formulation of an *Markov Random Field* model. The experiments are conducted on the *Brand-Social-Net* dataset that contains data collected from the Chinese microblog Sina Weibo, with 20 classified

events. Although the work proposes to relate events, it does so use pure text without extracting the attributes of the events.

SHI *et al.* (2017) presents a model called *HEE* (*Hot Event Evolution*) to discover the evolution of events reported in microblogs, like Twitter. The model establishes an unsupervised method that initially filters significant or hot events, then groups the events through the similarity of the keywords that correspond to topics, generating larger texts for each event. For overcome even more the problem of event sparse representation, the proposal uses information from the users of the posts in each event, and thus identifies user interest groups, improving the relationship between events and their evolution. In addition to the keywords of named topics and entity, each event is also characterized by timestamp extracted from the metadata and is used to establish the sequence of events evolution.

AL-SMADI *et al.* (2017) propose to discover paraphrases and textual semantic similarity between pairs of breaking news published on Twitter by portals in Arabic. Short texts from news are processed in a way that extracts lexical, syntactic and semantic features to characterize the events reported in each of them. So, though they do not treat events as simple texts, they also do not come to be thoroughly structured, only for features such as topic modeling and named entity. Part of the preprocessing for feature extraction is supported by tools that use auxiliary dictionaries, such as *NLTK*. The whole process is supervised, both the identification of paraphrase – treated as a binary classification problem – and the analysis of textual semantic similarity – treated as a regression problem with the use of the *SVR* (*Support Vector Regression*) classifier.

The work of CHEN *et al.* (2017) aims to detect and track events reported in tweets through a similarity metric with a deep neural network to relate the text to an event. For this, the proposal seeks to obtain representation for events and low-dimensionality tweets (dense or compact representation) and is effective since it is possible to relate tweets to a given event even when presented with a set with few common words. The network architecture to learn the similarity relationship between two tweets resembles the architecture proposed in the work of SANTORO *et al.* (2017), which seeks to discover relationships between two objects not necessarily textual. The method is supervised, and so the learning of the tweets representation is linked to the learning of similarity, and the event representation is weighting combination of the representations of related tweets. The evaluation is done with human annotation supervision over the 40 events with the highest amount of related tweets and has been shown to be superior to other methods of event detection. This work is somewhat similar to our current work, although the authors do not intend to relate events directly or treat them in a structured way.

LEBAN *et al.* (2014) created *EventRegistry*[4], a system that is capable of gathering news articles from different sources and reporting similar events. News articles are collected from approximately 75,000 news sources. The pre-processing activities are as follows: a named entity recognizer is applied, temporal and spatial information are fetched with regular expressions or from metadata, and the cross-lingual similarity is calculated, since the articles are in several languages. The systems uses external data, like *GeoNames* and *DMoz* taxonomies. Events are constructed by clustering news articles which are in a vector space representation, considering not only in their title, but also on their body and named entities found. Their process is not entirely automatic: in the last phase, experts intervene to manually input entries of future events and to edit events (like to merge pairs of clusters if were the same event).

Finally, the *AutoNEM* stands out regarding the others by satisfying all aspects analyzed until now. Chapter 6 details this approach, and Chapter 7 presents the evaluation of it.

[4] `http://eventregistry.org/`

Chapter 6

Detecting Relationships Between Events with *AutoNEM*

The Chapter 2 presented models of events including the conceptual modeling of both the essential attributes (*5W1H*) and the forms of relations between events. And Chapter 3 addressed the tasks of analyzing events in online media under some dimensions, especially the characterization of events extracted from short texts from the perspective of *5W1H*. This chapter presents the *AutoNEM*, a model with an unsupervised approach that starts from the characterization of events reported in short texts to discover relationships of similarity between such events, thus meeting the purpose of this research (see Figure 1.2).

The limitations identified from the works analyzed in Chapter 5 motivated the construction of the *AutoNEM*, whose elaboration was inspired by the *NEM* (DASIGI & HOVY, 2014). To cover these gaps, a four step process was developed (see Figure 6.2): (**i**) events detection and structuring, (**ii**) *AutoNEM* training, (**iii**) events encoding, and (**iv**) relationship rating between the events representations. Briefly explaining the process, the first step of the process, submits each text to a semantic role labeling tool (unsupervised), detecting and structuring the event with a set of arguments close to *5W1H*, and trains the vector representations of each term with a *word-embedding* technique using the training sentences. The second step takes as input a mass of structured events training combined with the *word-embeddings* of the terms that make up the arguments of each event, so an autoencoder that considers the structure of the event is trained for several times with these inputs. This neural network model whose architecture comprises an autoencoder for structured events is the core of the work, and it's called *AutoNEM* (*Autoencoder Neural Event Model*) – inspired by the work of DASIGI & HOVY (2014) as discussed in the Section 2.3. The third step, as important as the previous one, makes use of the first part of the *AutoNEM* architecture, known as *encoder*, to encode both a new event and its arguments in a dense and compact vector representation. From these

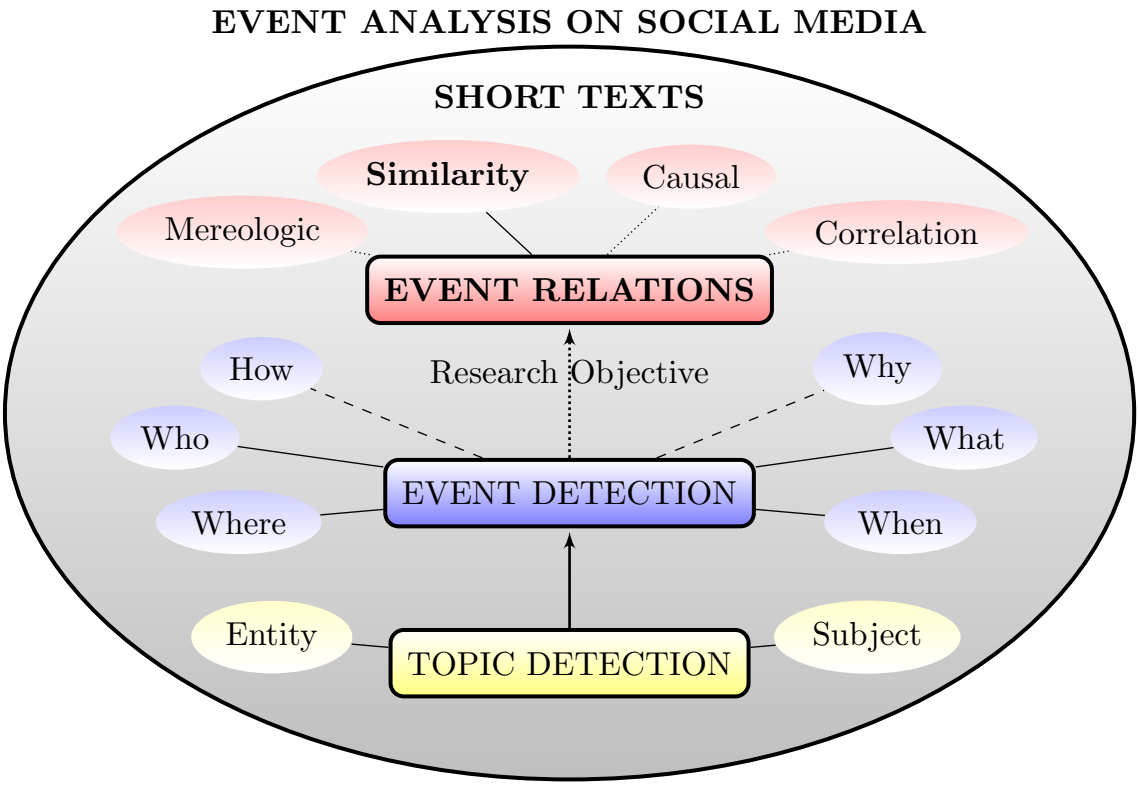


Figure 6.1: Research objective and limits within the area of event analysis on online media.

representations, in the last step, each pair of events is submitted to an evaluation to discover the degree of the similarity relation that may exist between them.

The remaining of this chapter provide details about each step of this process, especially when defining the architecture of the autoencoder and its application.

6.1 Event Detection and Structuring

As the *AutoNEM* should receive as input structured and equally computationally treatable events, the purpose of this step is to recognize events in short texts and to prepare as instances given events, composing them with semantic arguments that match (or approximate) to attributes of the *5W1H* together with the *word-embeddings* of terms that form the arguments. This stage could also be complemented by some more complicated process of event detection, but since this is not the objective of the work, it was decided to use an unsupervised way to recognize and characterize the event by extracting arguments from the text that approach *5W1H*.

As the process involves learning, training data and assessment data must be considered, and the two sets of data are submitted to this first step. Considering a collection $\mathbb{D}$ of sentences (text passages), this step is responsible for detecting

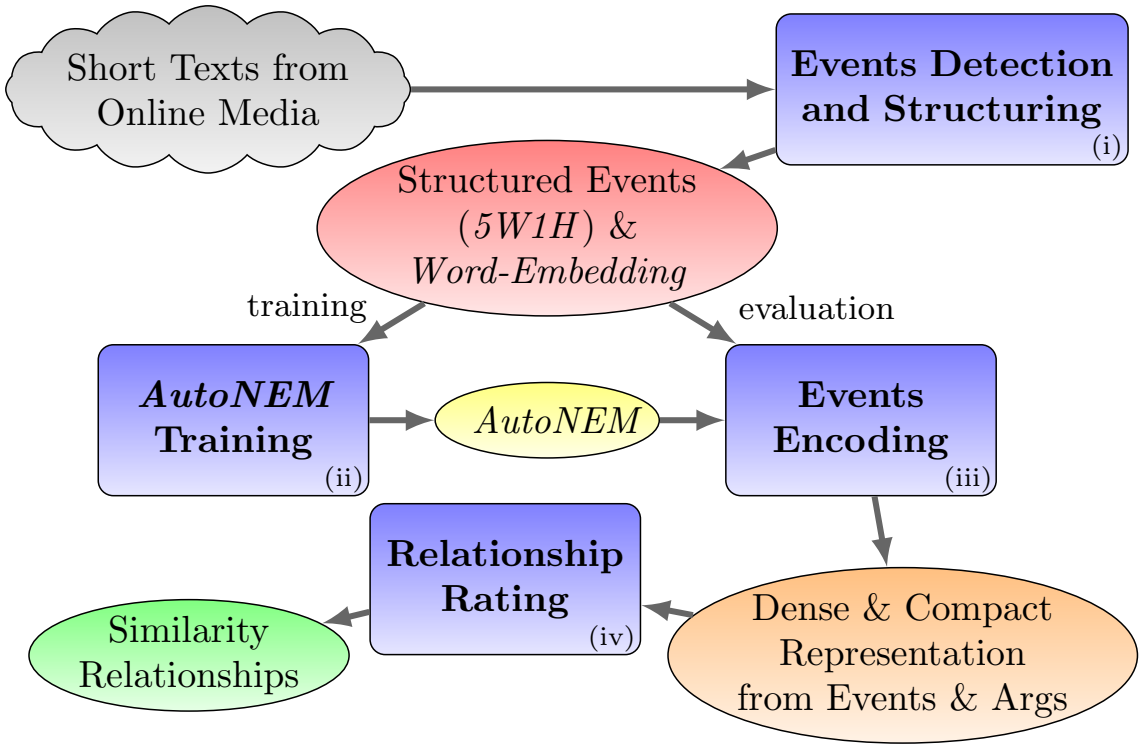


Figure 6.2: Process overview. Step (i) extracts *5W1H* arguments of each text, and produces the *word-embedding* vector representations of each argument. Step (ii) consists of training an autoencoder model in order to generate dense representations of each event. Step (iii) extracts these dense representations. Finally, in step (iv), each pair of events is evaluated in order to discover the similarity degree between them.

the corresponding event in each text passage $s \in \mathbb{D}$ (see Figure 6.3). First, each $s \in \mathbb{D}$ is passed through a tokenizer, which produces a list of corresponding tokens $[t_{s1}, t_{s2}, ...]$. This work applies a tokenization process according to the *Treebank*[1] standard. This list is then simultaneously given as input to two modules, namely, the *Semantic Tagger* and *Semantic Embedder*. This section describe these modules in the following paragraphs.

The *Semantic Tagger* is responsible for doing *semantic role labeling* (*SRL*) of each sentence. First, this module tags fragments t_{si} of s with their corresponding semantic arguments close to *5W1H* setting. Note that more than on token can be tagged together. For example, the *Semantic Tagger* can take the two consecutive tokens *"New"* and *"York"* and tag them as the component WHERE of the corresponding event. This module uses a software tool for (*SRL*) made available by *SENNA* (COLLOBERT, 2011; COLLOBERT *et al.*, 2011). This tool produces state-of-the-art results among the unsupervised *SRL* tools, so far, and the tags used by SENNA follow the *PropBank* tags style (PALMER *et al.*, 2005). Table 6.1 exemplifies the outputs generates by *Semantic Tagger*.

[1] `ftp://ftp.cis.upenn.edu/pub/treebank/public_html/tokenization.html`

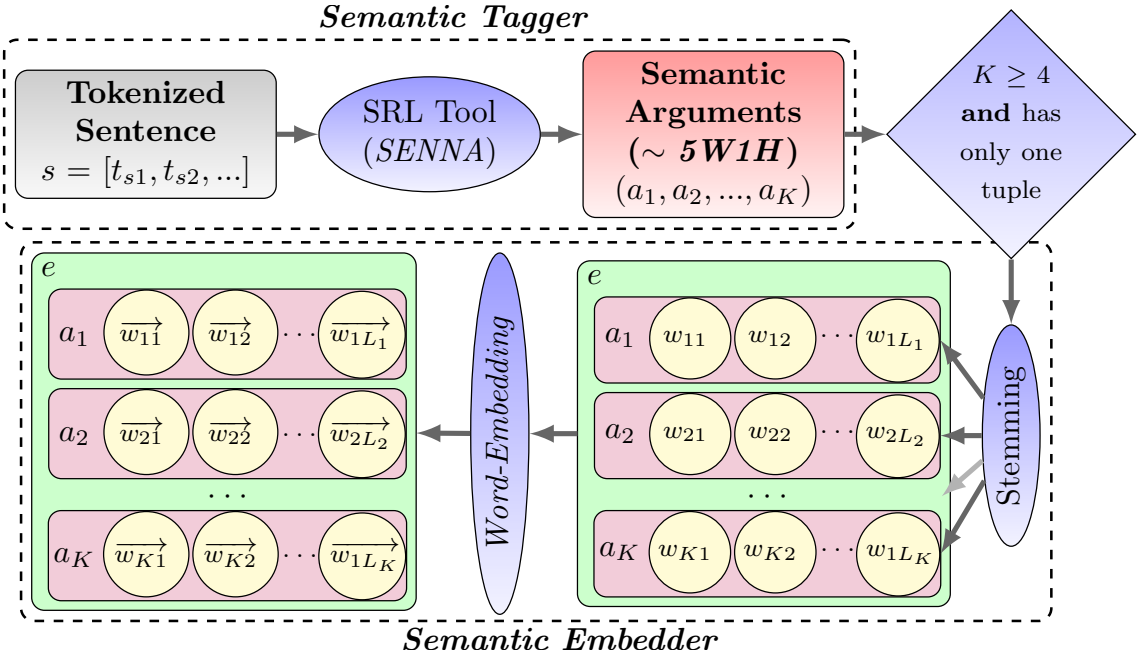


Figure 6.3: Overview of the event detection and structuring step.

With the examples in the Table 6.1, it's possible to see that arguments **V**, **A0** and **A1** revolve around the *WHAT*, *WHO* (and/or *WHOM*) attributes, while arguments **AM-TMP** and **AM-LOC** are closer to *WHEN* and *WHERE* attributes, respectively. Unfortunately, due to the use of *SENNA*, this step inherited limitations from this tool, which does not extract arguments close to *WHY* and *HOW*. Besides, while some sentences have 4 or 5 arguments characterizing the event well, others have few arguments, which can reduce the quality of event characterization. Therefore, the training and evaluation consider only the sentences from which it is possible to characterize a single event with at least four arguments. Longer sentences that evidence more than one event structure are discarded. Only after recognition of the arguments, its original terms pass by stemming.

Formally, the transformation realized by the *Semantic Tagger* is as follows. Each event $\boldsymbol{e}$ corresponding to a sentence $\boldsymbol{s}$ is structured as a tuple of semantic arguments $\boldsymbol{(a_1, a_2, ..., a_K)}$ close to *5W1H*, where $\boldsymbol{K}$ is the number of arguments extracted for each event. Each argument $\boldsymbol{a_k}$ is composed of a non-empty sequence of terms $\boldsymbol{[t_{k1}, t_{k2}, ..., t_{kL_k}]}$, where $\boldsymbol{L_k}$ is the size of the argument $\boldsymbol{a_k}$ (or just $\boldsymbol{L}$ when omitting the subscript does not cause confusion). Thus, each argument $\boldsymbol{a}$ has dimension $\boldsymbol{L}$, and each event $\boldsymbol{e}$ has dimension $\boldsymbol{K \times L}$.

The *Semantic Embedder* comprises a neural network model to produce word-embeddings from each $t_{si} \in [t_{s1}, t_{s2}, ...]$. One possibility for building this model is to use an unsupervised learning algorithm *SGNS* (*Skip-gram with Negative Sampling*) by LE & MIKOLOV (2014); MIKOLOV *et al.* (2013a,b), also known as *Word2Vec*.

Table 6.1: Examples of extracted semantic arguments from sentences with *SENNA*.

Input Sentences	**Extracted Arguments with *SENNA***				
	V	**A0**	**A1**	**AM-TMP**	**AM-LOC**
"Spanish coastguard rescues 600 migrants in 24 hours, amid surge in arrivals via Morocco."	rescues	Spanish coast-guard	600 migrants	in 24 hours	amid surge in arrivals via Morocco
"Pakistan violates ceasefire again in Poonch district."	violates	Pakistan	ceasefire	again	in Poonch district
"Petroleum Minister meets Odisha CM on 18th August 2017 in New Delhi."	meets	Petroleum Minister	Odisha CM	on 18th August 2017	in New Delhi
"Thousands Protest White Supremacy In New Orleans."	Protest	Thou-sands	White Supremacy	–	In New Orleans
"6 police officers were shot in 3 US cities Friday night."	shot	–	6 police officers	Friday night	in 3 US cities
"World leaders condemn Barcelona attack."	con-demn	World leaders	Barcelona attack	–	–
"BPL Union Wants Managers Fired."	Fired	BPL Union Wants Man-agers	–	–	–

Alternatively, this stage can use a pre-trained *word-embedding* model[2]. A *word-embedding* model works as follows: given a set of stemmed words (i.e., a dictionary) $\mathcal{W}$, this model maps each stemmed word $\boldsymbol{w} \in \mathcal{W}$ to a $\boldsymbol{N}$-dimensional feature vector, that is, $\boldsymbol{w} \in \boldsymbol{\mathcal{W}}$ and $\overrightarrow{\boldsymbol{w}} \in \mathbb{R}^{\boldsymbol{N}}$. Thus, given each source token $t_{si} \in [t_{s1}, t_{s2}, ...]$, this module first applies the stemming to get a stemmed word w_{si} and after generates its representation in a low-dimensionality semantic space. So, it is worth mentioning this module applied the stemming to each t_{si} before passing it to the model, because it is usual for both pre-trained and in-house training. The last task of this step is to incorporate *word-embeddings* to represent the terms of the arguments that make up each event.

Formally, the *Semantic Embedder* works as follows. With stemming,

[2]There are many pre-trained models available on the Web: `https://github.com/kudkudak/word-embeddings-benchmarks/wiki`

each argument $\boldsymbol{a_k}$ is composed of a non-empty sequence of stemmed word $[\boldsymbol{w_{k1}, w_{k2}, ..., w_{kL_k}}]$. After, each argument $\boldsymbol{a_k}$ is associated to a sequence of embeddings $[\overrightarrow{\boldsymbol{w_{k1}}}, \overrightarrow{\boldsymbol{w_{k2}}}, ..., \overrightarrow{\boldsymbol{w_{kL_k}}}]$ representative of its stemmed terms. Since term vectors have dimension $\boldsymbol{N}$, each argument $\boldsymbol{a}$ has dimension $\boldsymbol{L \times N}$ ($\boldsymbol{a \in \mathbb{R}^{L \times N}}$), and each event $\boldsymbol{e}$ has dimension $\boldsymbol{K \times L \times N}$ ($\boldsymbol{e \in \mathbb{R}^{K \times L \times N}}$). Each term not found in the word dictionary (i.e., without *word-embedding*) receives a representation close to the origin of the vector space, but not zero. All word vectors are normalized independently using L2-norm (HAKAMI & BOLLEGALA, 2017; LEVY *et al.*, 2015).

As a result of this step, events are structured in semantic arguments (close to *5W1H*) extracted from short texts, and terms are represented by semantic vectors (*word-embeddings*). It is worth noting that all this preparation is performed in an unsupervised way. However, although this initial representation of the event is dense for using *word-embeddings*, it is still not compact because the size of the arguments is variable (not fixed). The next steps overcome this and others limitations mentioned in the previous chapter.

6.2 *AutoNEM* Training

This step corresponds to train the *Autoencoder Neural Event Model* (*AutoNEM* for short). Section 6.2.1 provides a description of the loss function used in the training procedure. Section 6.2.2 details the AutoNEM architecture.

6.2.1 Loss Function

The input to the training procedure is a set of event tuples, structured according to the output produced by the *Semantic Embedder*. Since *AutoNEM* is an autoencoder, the model is trained in such a way that the output is a reconstruction of the input.

Just as most autoencoders train the input reconstruction by minimizing *Sum of Squared Errors* (*SSE*), the training seeks to minimize the *SSE*-based event reconstruction error between word-embeddings of all original input arguments and their reconstruction. Of course, just as the word-embeddings of the input are normalized independently, their reconstructions are also normalized just before calculating the error between them. Also, since each event has no fixed size, that is, it does not have a fixed amount of words in its arguments, the total amount of terms normalizes the reconstruction error in each event.

Formally, let $\boldsymbol{e}$ be a structured event with the arguments $\boldsymbol{a_k}$ ($\boldsymbol{k \in \{1, 2, ..., K\}}$, where $\boldsymbol{K}$ is the number of arguments of event $\boldsymbol{e}$) composed of a sequence of word-embeddings $\overrightarrow{\boldsymbol{w_{ki}}}$ ($\boldsymbol{i \in \{1, 2, ..., L_k\}}$, in which $\boldsymbol{L_k}$ is the size of the argument $\boldsymbol{a_k}$) of dimension $\boldsymbol{N}$ (i.e., $\overrightarrow{\boldsymbol{w_{ki}}} = [\boldsymbol{w_{ki1}, w_{ki2}, ..., w_{kiN}}]$, where $\boldsymbol{w_{kij} \in \mathbb{R}}$), and its recon-

struction $\boldsymbol{e'}$ consisting of its arguments $\boldsymbol{a'_k}$ with the word-embeddings $\overrightarrow{\boldsymbol{w'_{ki}}}$. Hence, the training seeks to optimize the cost function $\boldsymbol{\mathcal{L}(e, e')}$ (see development of Equation 6.1).

$$\mathcal{L}(e, e') = \|e - e'\|_2^2 \Big/ \sum_{1 \leq k \leq K} L_k \tag{6.1a}$$

$$\mathcal{L}(e, e') = \left(\sum_{1 \leq k \leq K} \left\| a_k - a'_k \right\|_2^2 \right) \Big/ \sum_{1 \leq k \leq K} L_k \tag{6.1b}$$

$$\mathcal{L}(e, e') = \left(\sum_{1 \leq k \leq K} \left(\sum_{1 \leq i \leq L_k} \left\| \overrightarrow{w_{ki}} - \overrightarrow{w'_{ki}} \right\|_2^2 \right) \right) \Big/ \sum_{1 \leq k \leq K} L_k \tag{6.1c}$$

$$\mathcal{L}(e, e') = \left(\sum_{1 \leq k \leq K} \left(\sum_{1 \leq i \leq L_k} \left(\sum_{1 \leq j \leq N} \left(w_{kij} - w'_{kij} \right)^2 \right) \right) \right) \Big/ \sum_{1 \leq k \leq K} L_k \tag{6.1d}$$

To the model be successful, the training over it must repeatedly feed it with a significant amount of structured events that present at least four semantic arguments close to *5W1H*. Moreover, although the previous step introduces a limitation on the maximum number of arguments (5, due to the use of *SENNA*), *AutoNEM* can be trained with more arguments if it is possible to identify them with other tools or techniques.

6.2.2 *AutoNEM* Architecture

Figure 6.4 shows the architecture of *AutoNEM*, a deep, undercomplete and recurrent autoencoder that encodes the semantic structure of an event, to obtain dense and compact representations of the event and its semantic attributes. Each of its two parts, the encoder and the decoder, are composed of stacked layers and the dimensions of the hidden layers are smaller than the dimensions of the input and output layers. In particular, the hidden layer is the bottleneck between the encoder and the decoder. Despite being inspired by DASIGI & HOVY (2014) work on recursive and recurrent networks, *AutoNEM* does not deal with recursive networks, but Recurrent Neural Networks (*RNN*), in addition to differentiating itself by presenting unsupervised learning.

The two big dashed rectangles in Figure 6.4 are responsible for encoding and decoding. Each one presents its respective stacks of layers. The encoder comprises two layers (besides the input) that are responsible for encoding the arguments and the event as a whole. The layer immediately after the input encodes the word-embeddings sequence of each argument and is formed by $\boldsymbol{K}$ sets of *LSTM* layers whose number of units in each set is determined by the desired dimensionality for

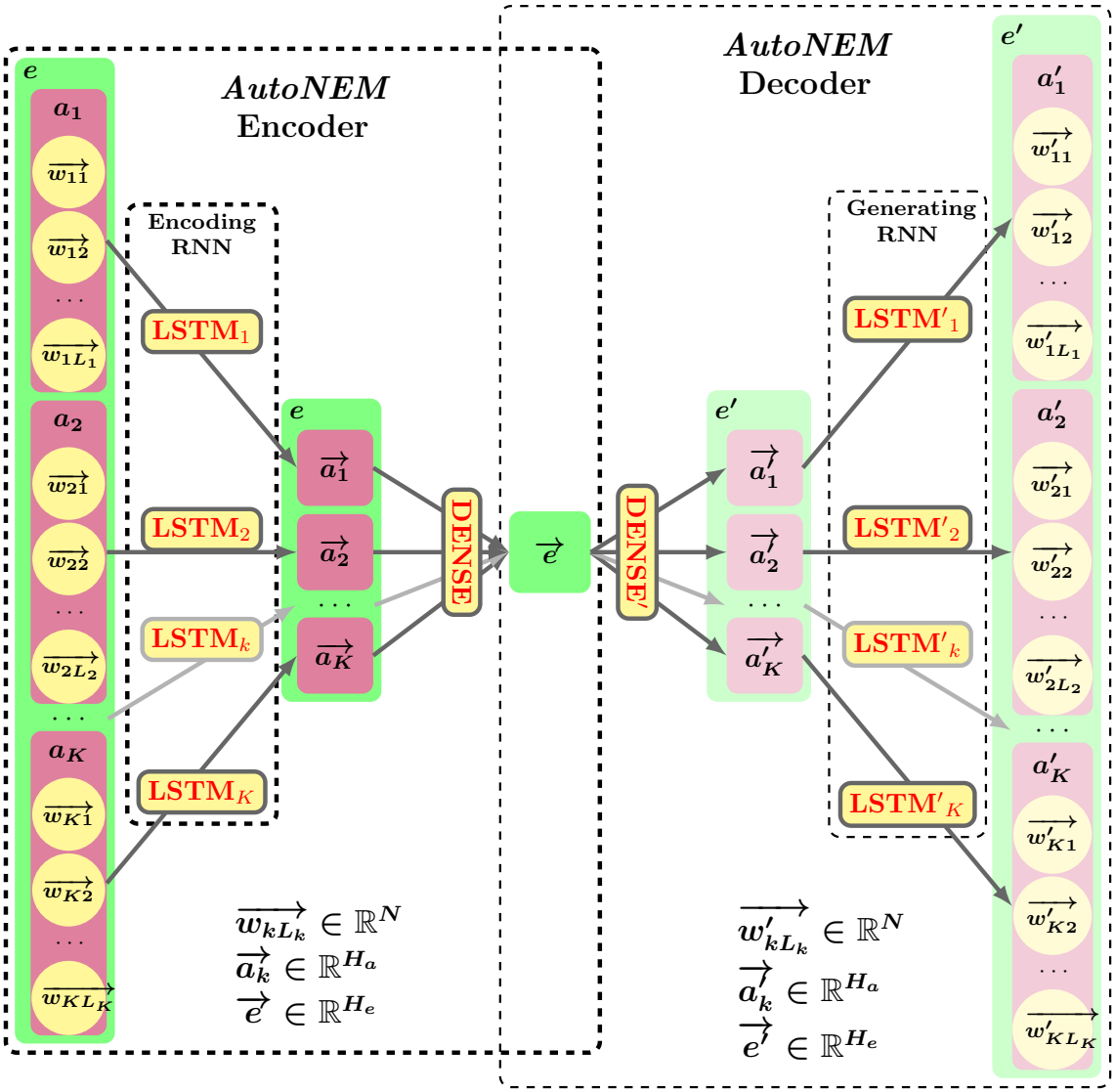


Figure 6.4: *AutoNEM* Architecture

the vector representation of each argument (denoted by $\boldsymbol{H_a}$). By doing this role, the layer is an *Encoding RNN* (LECUN *et al.*, 2015; SUTSKEVER *et al.*, 2014). The next layer is dense, and its goal is to obtain a vector representation for the event that considers the influence of all the arguments. The desired dimensionality determines the amount of units in that layer for the representation of the event (indicated by $\boldsymbol{H_e}$). The decoder looks to mirror the encoder to reconstruct the original event. Therefore, its first layer is dense followed by the layer of *LSTMs*, one for each argument to be reconstructed. This layer of *LSTMs* is characterized as an *Generating RNN* because it generates a sequence from an input, instead of encoding a sequence.

Figure 6.4 doesn't expose some elements of the encoder and decoder, as the activation functions of each layer. In the encoder RNN, the standard hyperbolic tangent is used as activation function in the *LSTM* cells. In the subsequent two dense layers,

the activation function is linear, which allows generating dense representations. The last layer uses a smoothed hyperbolic tangent as activation function. The choice of a smoothed symmetrical function allows the reconstructed values be distributed a little less at the ends, to approach the component features of word-embeddings.

It is worth mentioning that there are rare textual data autoencoders that deal with embeddings directly in the input and output, like those of SOCHER *et al.* (2011a,b). Most receive as input a sequence of terms, in which each word (or even each character) has a one-hot vector representation according to its index in the dictionary (CHEN & ZAKI, 2017; CHO *et al.*, 2014; GAL & GHAHRAMANI, 2016; HILL *et al.*, 2016; SUTSKEVER *et al.*, 2014; YANG *et al.*, 2017). Although it is a discrete representation and therefore easier to direct the training, it is sparse and usually presents a high dimensionality, which forces a limitation on the size of the dictionary due to the high memory consumption, besides not incorporating any semantics to the terms directly. Other works implement autoencoders with entries formed by weights of the terms of the collection, such as *TF-IDF* (YANG *et al.*, 2015), which although it reduces the sparsity, still does not avoid high dimensionality. Except for this last work, already analyzed in Chapter 5, none of the others deal with events, but only with plain texts.

6.3 Events Encoding

After the training of the *AutoNEM* model is finished, anyone can use its encoder part to produce dense representations of events. In this step, when submitting a new event to the *AutoNEM* encoder, the output can be produced on two different representation levels. One of them corresponds to the vector $\overrightarrow{\boldsymbol{e}}$ (see Figure 6.4). The other representation corresponds to the vector $\overrightarrow{\boldsymbol{a_k}}$, one for each argument in the original structured event provided as input. Each of these vector representations is one-dimensional and has a fixed size ($\boldsymbol{H_e}$ for the event and $\boldsymbol{H_a}$ for each argument) – making them simpler compared to the dimensions of the representation of the event. Consequently, these representations can be easily employed in computational tasks where there is a need to treat each event, or each argument, equally and individually.

6.4 Relationship Rating

Once the dense representations of the events (and their arguments) are obtained, the last step in the process is responsible for evaluating the similarity relations that may exist between the instances of the events. In order to do this, it is enough to apply some similarity function to the representations of two events. One possibility is to compute the cosine of the angle formed by the two vectors representations.

Thus, if $\overrightarrow{\boldsymbol{x}}$ and $\overrightarrow{\boldsymbol{y}}$ are two vector representations (either for event, or for its $\boldsymbol{k}$-th argument), then the degree of similarity between them is given by Equation 6.2.

$$\text{Sim}(\overrightarrow{x}, \overrightarrow{y}) = cos(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{||\overrightarrow{x}|| \cdot ||\overrightarrow{y}||} \tag{6.2}$$

In this way, it intends to obtain a degree of similarity between two events detected through their representations, or under the aspect of some of their encoded attributes.

6.5 Properties of *AutoNEM* and its Process

To characterize the present propose against the literature limitations, follow a set of properties (or requirements) that have not yet considered in a single model, technique or approach, but now all of them are present in *AutoNEM*.

- Provide a way to detect similarity relationships between events;
- Use a structured way to model the event from the perspective of *5W1H*;
- Present an approach independent of the field of application or knowledge;
- Present a form of fully unsupervised learning;
- Treat reports of events published within short texts in online media, such as the news headlines;
- Use only text that reports the event while information source, disregarding any other media or information, such as images, timestamp, or any other metadata;
- Provide a dense and compact representation of the event;
- Discover relationships between events extracted from different publications.

Chapter 7

Experimental Evaluation

This chapter presents all the experiments carried out, to confirm the effectiveness of the proposal. As the literature does not manifest any previously and accessible dataset for the specific purpose of the current work, a golden corpus had to be built. For comparison purposes, as no research meets all the properties of the current model, the experiments make use of models consolidated in the literature that provide a compact representation of the data. The core procedures performed in these experiments follow the proposed flow (see Figure 6.2). The sequence of procedures corresponds to the order in which they are described below, beginning with the description of the data used in Section 7.1, succeeded by the core process in Section 7.2, after by the baselines setup in Section 7.3, and finishing with the performance evaluation with results and discussions in Section 7.4.

7.1 Collected Data and Golden Corpus

Due to the purpose of evaluating the proposed process for relationships discovery, this work realized experiments with news headlines collected from different sources in English. The used sources for this step (*AutoNEM* training) are the following: English Newswire of the LDC[1] Online Corpus (contains news in English of all continents), Reuters Corpus[2] (RCV1/2) and Wikinews[3]. As shown in Table 7.1, the collect over these three sources resulted in 6.805.610 news items available to *AutoNEM* training. Because the process relies on an autoencoder-based unsupervised approach, the collected data for training does not need labels.

Regarding the evaluation data, the experiments collected reported events in news in English from anywhere in the world grouped through the *EventRegistry API*[4]

[1] `https://www.ldc.upenn.edu/`
[2] `http://trec.nist.gov/data/reuters/reuters.html`
[3] `https://en.wikinews.org/`
[4] `http://eventregistry.org/`

Table 7.1: Amount of collected data (english news) for training.

Source	News
LDC Online	5,939,654
Reuters	806,791
Wikinews	59,165
TOTAL	6,805,610

(LEBAN *et al.*, 2014). This tool provides these groupings through an automated process, followed by a manual peer review (see end of Chapter 5). Therefore, such clusters can be used as the golden corpus for the evaluation of a method that quantifies the similarity between any two events reported. The free *EventRegistry API* allowed collecting 30,385 news articles with its stream of events method, reporting 477 unique events. Figure 7.1 shows the amount of news reporting each unique event in the real world according to *EventRegistry*, and Figure 7.2 shows the frequency distribution of the amount of news in each unique event.

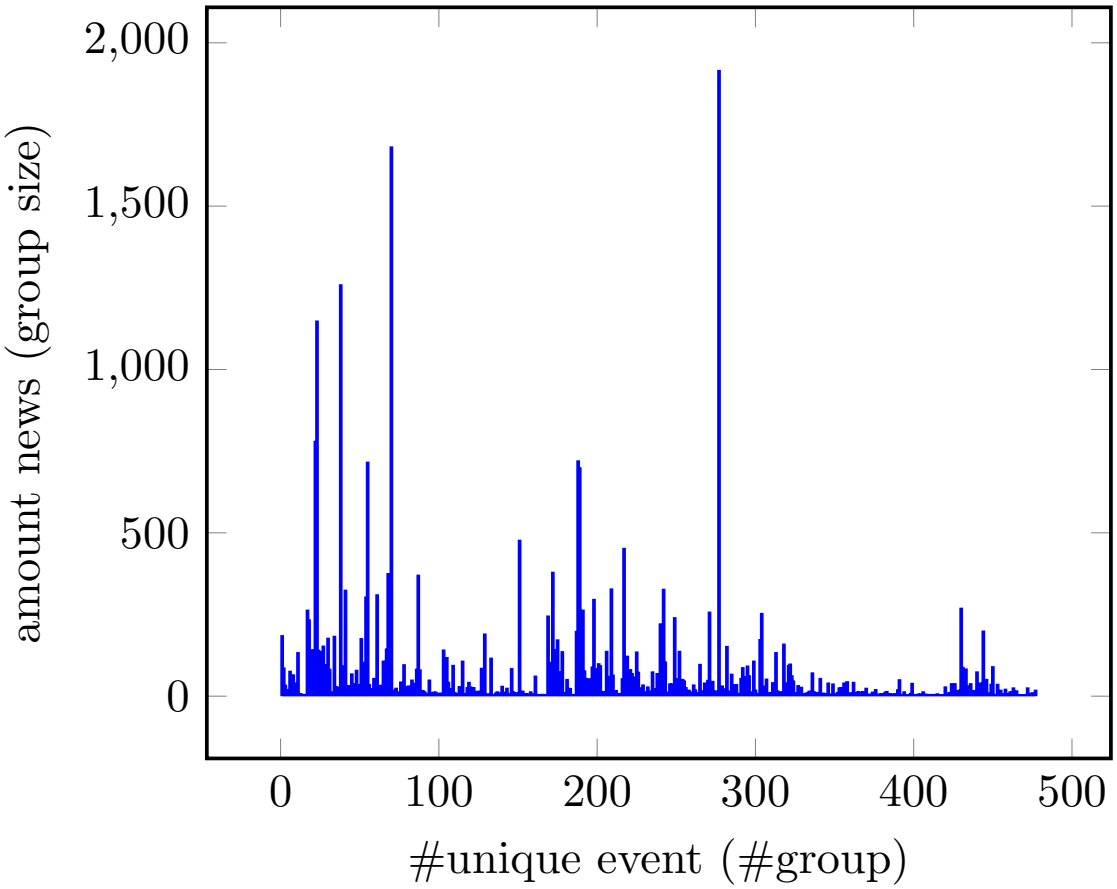


Figure 7.1: Amount news by each unique event in the collected golden corpus from *EventRegistry* for evaluation.

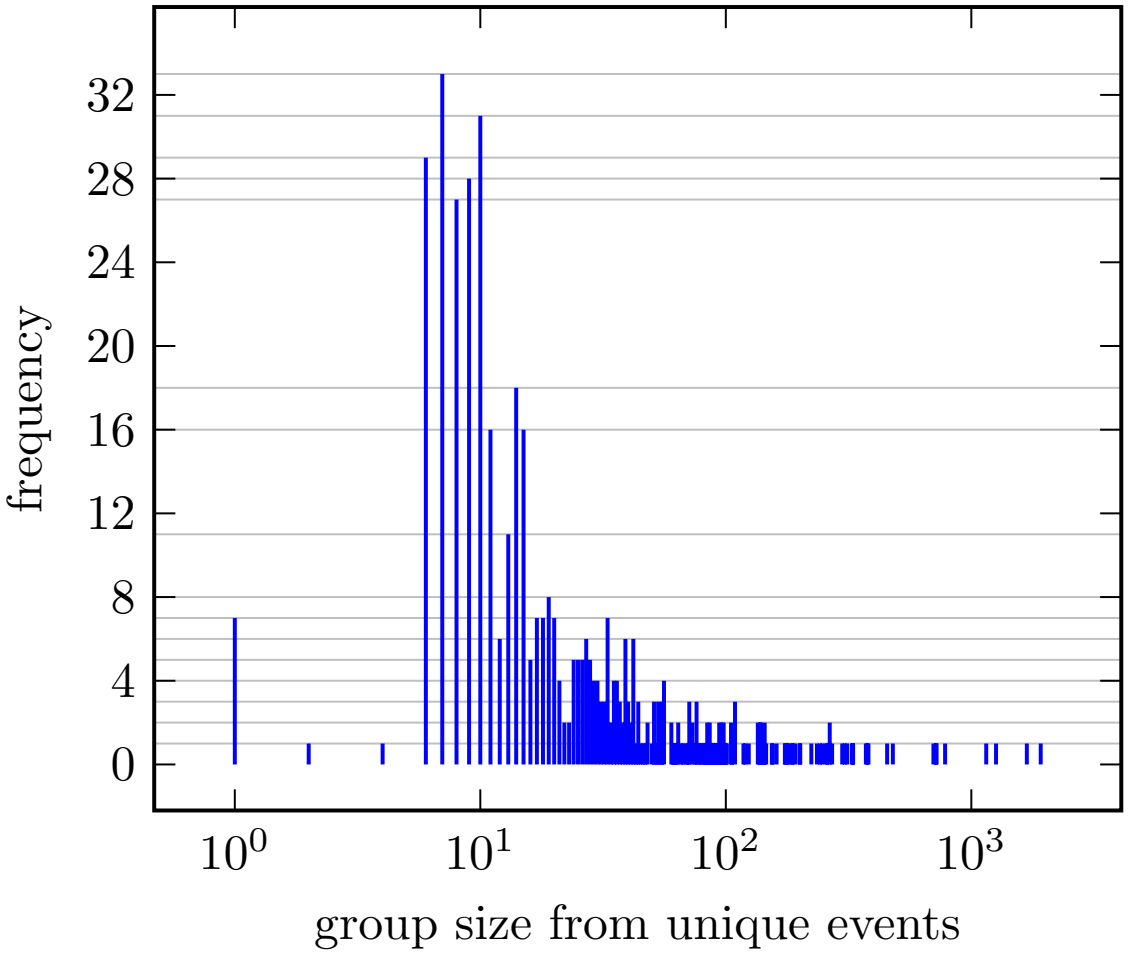


Figure 7.2: Frequency distribution of unique events by news amount in the collected golden corpus from *EventRegistry* for evaluation.

7.2 Experimentation of the *AutoNEM* Process

This section focus on the application of the proposed process in the experiments, and details the employment of each process step: (**i**) Event Detection and Structuring, (**ii**) *AutoNEM* Training, (**iii**) Events Encoding and (**iv**) Relationship Rating.

7.2.1 Event Detection and Structuring

With the data collected from online medias, the first step of the process detected and structured the events from training and evaluation data (as described in Section 6.1). First, the *Semantic Tagger* tokenized each news headline (each headline contains a sentence), and then applied the *SENNA* to extract the arguments and characterize the reported event. Then, the process filtered the sentences that mentioned only one event (i.e., a single argument tuple, or, a hierarchy level extracted from *SENNA*) and detected well-characterized events (tuples consisting of at least four arguments). Thus, it reduced the number of items to 234,451 from the training data, and to

1,438 from the golden corpus evaluation data (see details of filtering and detection in Table 7.2).

Table 7.2: Number of news headlines before and after filtering and detection through the arguments extracted by *Semantic Tagger* with *SENNA*.

Corpus	**Collected**	**One Tuple of Args**	**and Qtd.Args ≥ 4**
Training Corpus	6,805,610	3,928,277	234,451
Evaluation Corpus	30,385	19,397	1,438

Figures 7.3 and 7.4 show the new distribution of filtered golden corpus data (Section 7.4 references these figures to analyze them in order to direct the evaluations).

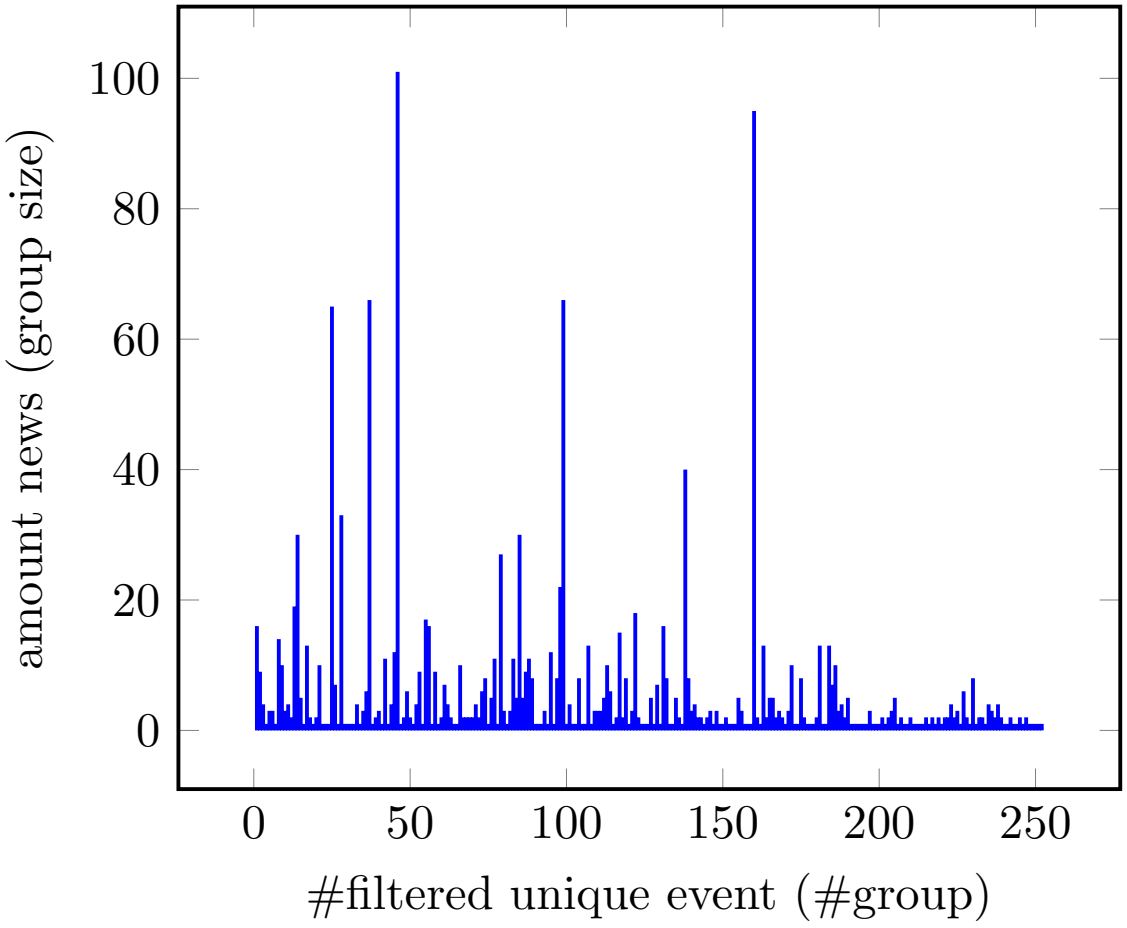


Figure 7.3: Amount news by each unique event in the **filtered** golden corpus from *EventRegistry* for evaluation.

Still in this step, *Semantic Embedder* trained the word-embeddings with the context of all data collected for training (i.e. 6,805,610 news). To do this, it employed the *Word2Vec* implementation provided by the *Gensim API*[5] (ŘEHŮŘEK & SOJKA, 2010), resulting in 168,794 stemmed tokens and its vector representations

[5] `https://radimrehurek.com/gensim/`

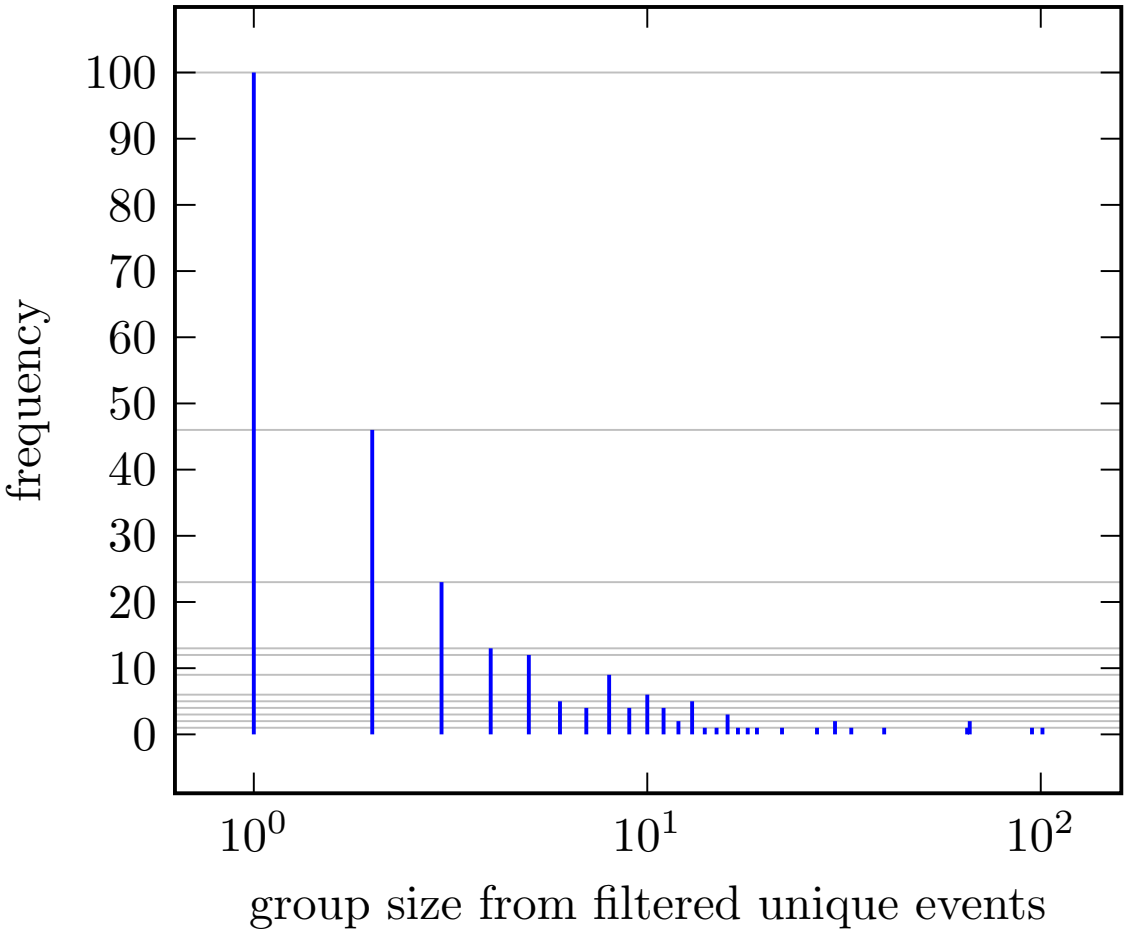


Figure 7.4: Frequency distribution of unique events by news amount in the **filtered** golden corpus from *EventRegistry* for evaluation.

with 100 features (i.e., $|\mathcal{W}| = 168,794$ and $\overrightarrow{w} \in \mathbb{R}^{100}$). Then, *Semantic Embedder* applied the stemming on the terms of the sentences with events detected in the two datasets, at the same time that it structured the events with their extracted arguments. In the end, it applied the embeddings on each stemmed word (when present in the dictionary) obtaining the final structure of each event.

7.2.2 *AutoNEM* Training and Events Encoding

The experiments implemented the *AutoNEM* with *Keras*[6] and *Theano*[7], and it setup the model with 100 units for each hidden layer of the *AutoNEM Encoder*, i.e., for both the layers that encode each argument (*Encoding RNN*) and the dense layer that encodes the event. Formally, the experiments configured the model with the following hyper-parameters: $H_a = 100$ and $H_e = 100$. It also configured the *AutoNEM Decoder* with matching parameters.

Then, the process trained *AutoNEM* (as described in Section 6.2) for 100 epochs.

[6] `https://github.com/keras-team/keras`
[7] `http://deeplearning.net/software/theano/`

After each training epoch, this second step submitted the evaluation data were to the model to gauge the reconstruction error. This allowed us to choose the best model, i.e., the one which produced the least reconstruction error on the evaluation data. Figure 7.5 shows the learning curve of the reconstruction error. All this process took just over four days with the 234,451 events detected for training. Once the process picked this best model, the third step submitted the structured events contained in the evaluation data to the corresponding *AutoNEM* encoder, in order to obtain their respective vector representations, $\overrightarrow{a_k} \in \mathbb{R}^{100}$ to each argument and $\overrightarrow{e} \in \mathbb{R}^{100}$ to the event (according to Section 6.3).

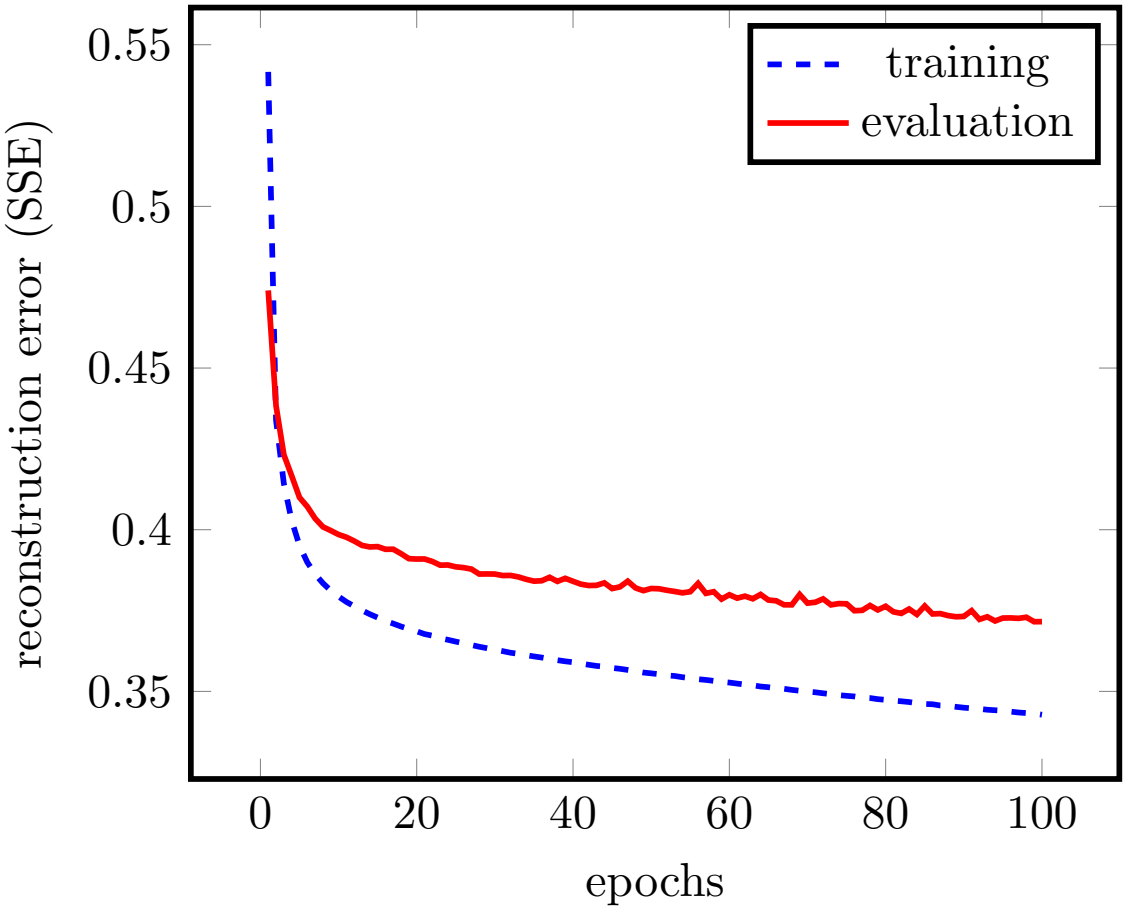
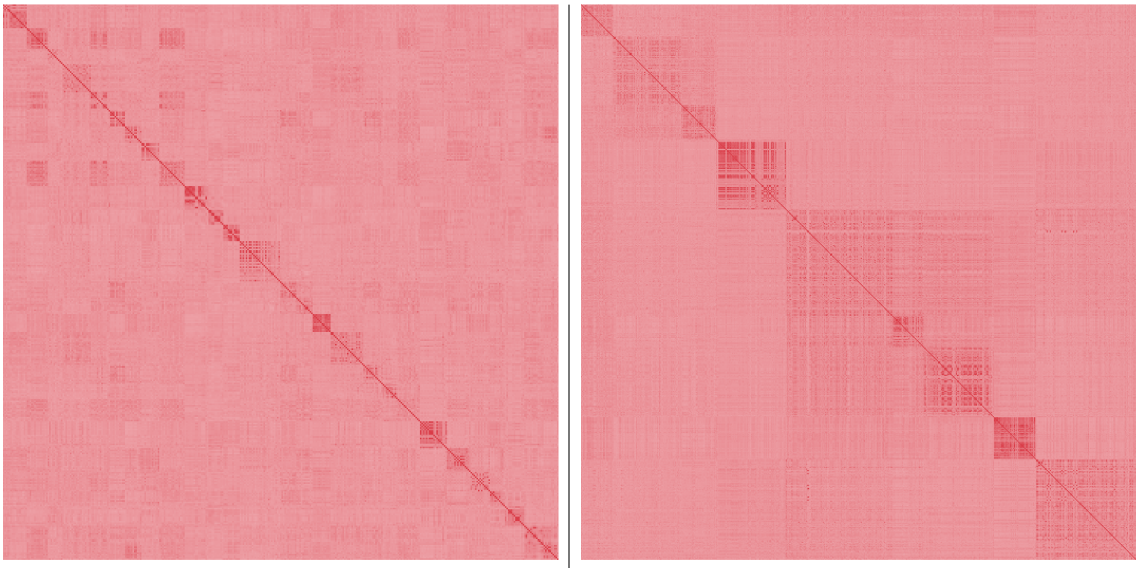


Figure 7.5: Reconstruction learning curve from *AutoNEM*.

7.2.3 Relationship Rating

Taking the vector representations for events and arguments as input, the last step computed the similarities between them (according to Section 6.4). The experiments computed a similarity matrix (by applying Equation 6.2) to store similarities between pairs of vector representations. Figure 7.6 shows heatmaps of similarity matrices for two subsets from encoded events (Section 7.4 explains the partitioning

and choice of these subsets). By observing the components close to the main diagonal, can see that the degrees of similarity obtained between the representations that refer to the same event approaches the ideal, varying the degree among the different events.



(a) Data subset with mention to the same event (according to *EventRegistry*), which has at least 10 news and in less than 30 news.

(b) Data subset with mention to the same event (according to *EventRegistry*), which has at least 30 news.

Figure 7.6: Heatmaps of similarity matrices between events encoded by *AutoNEM*.

7.3 Baselines Setup

Although Figure 7.6 indicate the effectiveness of the model, it is still not enough to confirm it. For this, it is necessary to compare the proposal with other similar models. Since there are no other event models that meet the same requirements of *AutoNEM*, the experiments used as baseline some models consolidated in the literature and that provide a compact representation with the same dimensionality obtained from *AutoNEM* Encoder. With this, it is possible to position the *AutoNEM* in front of these baselines. These experiments chose linear and nonlinear models as in VAN DER MAATEN *et al.* (2009), since *AutoNEM* is nonlinear because this model has overlapping layers and bottlenecks. It employed the *PCA* (TIPPING & BISHOP, 1999) and the *Truncated-SVD* (HALKO *et al.*, 2011) as linear baselines, further the *KernelPCA-GaussianRBF* (SCHÖLKOPF *et al.*, 1999) and the *Eigenmaps* (JIANBO SHI & MALIK, 2000; NG *et al.*, 2001; VON LUXBURG, 2007) as nonlinear baselines. *Scikit-learn API*[8] provided the implementation of such models.

[8] `http://scikit-learn.org/`

It configured these models to return a compact representation with 100 components, similar to the *AutoNEM* configuration.

The experiments applied each baseline both over the sentences in the filtered golden corpus and over each attribute extracted from those sentences. Thus, each baseline provided a representation for the sentence and one for each argument. As it's typical in the literature, these models received as input the *TF-IDF* based representation obtained from the sentences or the arguments, according to the desired element to get and analyze the similarity with others. It didn't use the training data to prepare these four baselines, but only the filtered golden corpus for evaluation.

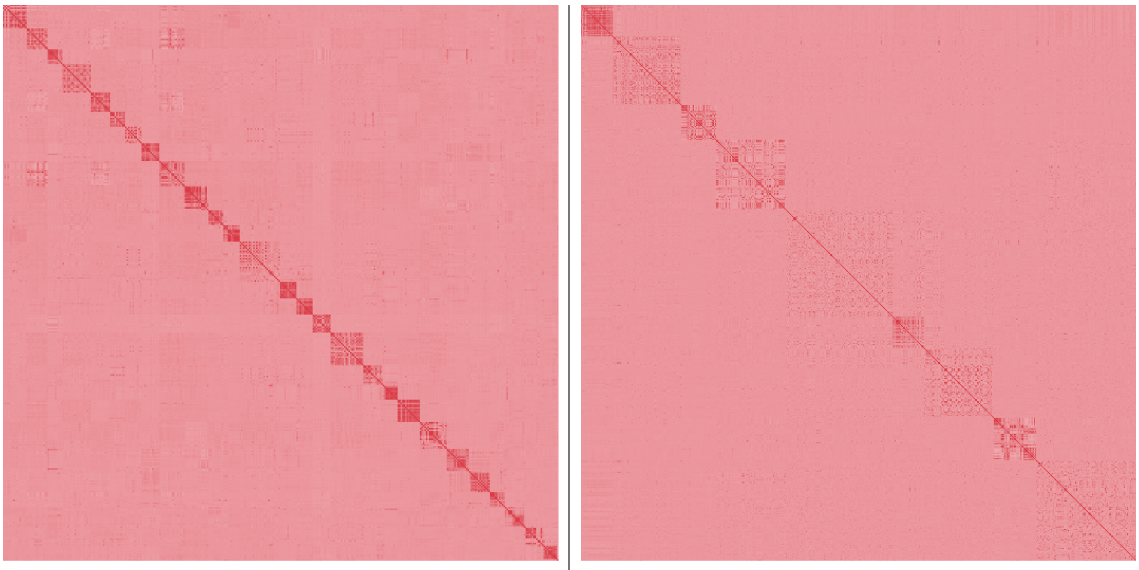
In addition to these baselines, another baseline was created and added to the experiments. To compare *AutoNEM* performance, it created another autoencoder with the same number of layer levels and received the whole sentence with word-embeddings as input and output. This baseline called *AutoFlat* is a kind of simplified version of *AutoNEM* whose only argument is the sentence itself. Thus, while *AutoNEM* works only with the extracted arguments that characterize the events mentioned in the sentences (i.e., only part of the information contained in the sentence), *AutoFlat* sees the whole sentence, bearing more information is considered by the model. The performing of *AutoFlat* training was in the same way as *AutoNEM*. It defined the dimensionality of the sentence representation, obtained as output from the dense layer, with 100 features.

Since all baselines encode the entire sentence, then it is possible to get the similarity between these encodings easily. However, since *AutoFlat* is the only baseline from which it is not possible to get representation for each argument, the similarity between the corresponding arguments is only feasible with the other baselines. Figure 7.7 shows heatmaps of similarity matrices for two subsets from encoded sentences by *KernelPCA-GaussianRBF*.

7.4 Performance Evaluation and Discussions

Finally, the experiments proceeded to evaluate the resulting degrees of similarity calculated against the golden corpus collected from *EventRegistry* and comparing with the results of each baseline, similarly to the evaluation conducted by TAI *et al.* (2015).

For this, it was first necessary to analyze the frequency distribution of the amount of news with the same event, according to the event clusters contained in the filtered golden corpus (see Figures 7.4 and 7.4). From this new distribution, it's can see that there are 100 separate news stories (i.e., news stories mentioning an event which no other news mentions). Besides, most news stories mention a small number of unique events (for example, 101 news stories mention the same event). Therefore,



(a) Data subset with mention to the same event (according to *EventRegistry*), which has at least 10 news and in less than 30 news.

(b) Data subset with mention to the same event (according to *EventRegistry*), which has at least 30 news.

Figure 7.7: Heatmaps of similarity matrices between sentences by *PCA*.

to analyze a large amount of news that has strong similarity between their respective mentioned events, one has chosen to select unique events that are cited in at least 5 news articles. Thus, the number of representations extracted from the events contained in the news fell to 1,125, which mention 70 unique events. For a better results analysis, the experiments divided these data into three subsets according to the group size of each unique event. The first comprising unique events mentioned in less than 10 news (containing 34 unique events and totaling 226 news), the second comprising unique events mentioned at least 10 news and in less than 30 (containing 27 unique events and totaling 373 news), and the third comprising 30 or more news (containing 9 unique events and totaling 526 news). Table 7.3 details this subsets setup.

The heatmaps exhibited earlier in Figure 7.6a and Figure 7.6b show the similarity matrices obtained with the *AutoNEM* over the second and third subsets, respectively. The heatmaps in Figure 7.7a and Figure 7.7b show the similarity matrices obtained with the *KernelPCA-GaussianRBF* over the second and third subsets, respectively.

Comparative evaluation begins with the formation of sentence pairs that mention the same event (according to *EventRegistry*) in each of the three partitions. Thus, for each sentence pair, the experiments calculate the similarity between their respective representations with *AutoNEM* and every baseline. Further, for each of these pairs, it calculates the similarity between the representations of the corresponding attributes with *AutoNEM* and the four baselines firstly presented (since *AutoFlat*

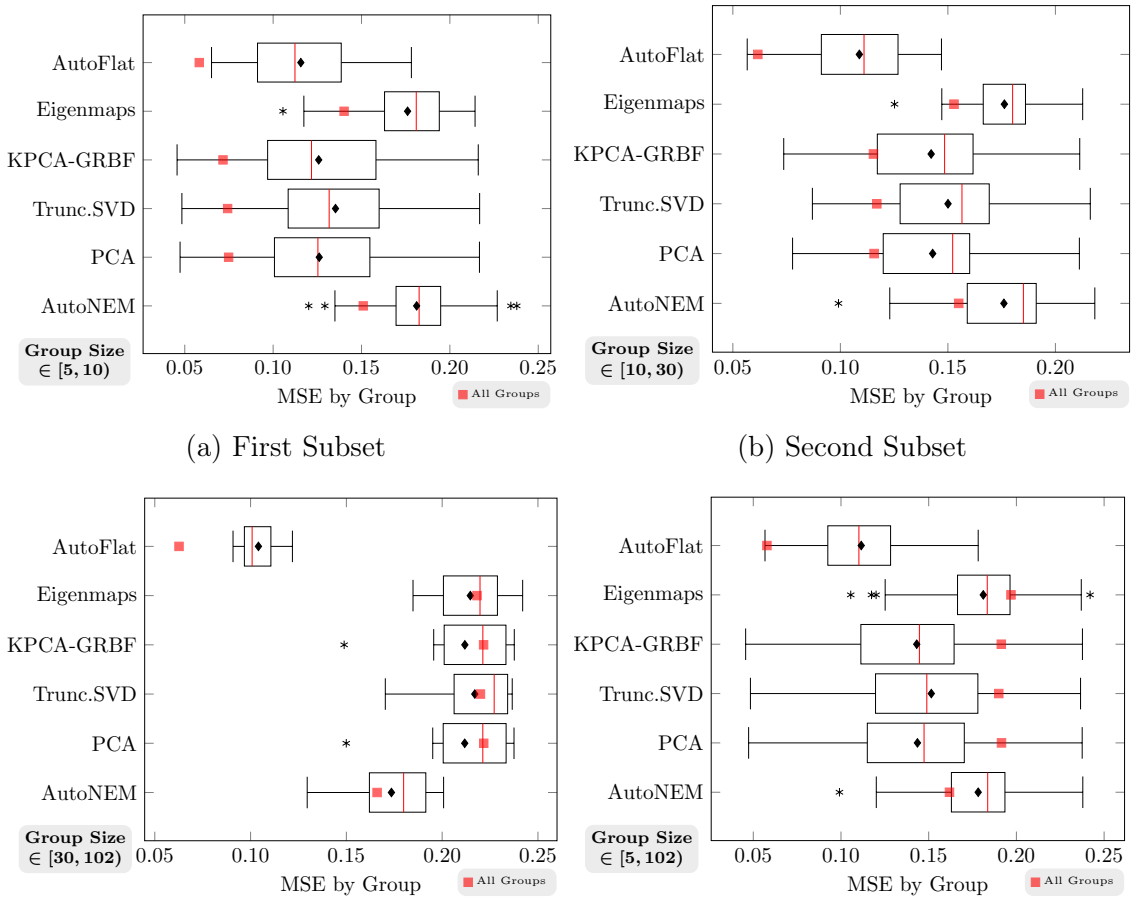
Table 7.3: Subsets setup from the **filtered** golden corpus for evaluation.

Subset	Group Size	Groups (Frequency)	News	Groups by Subset	News by Subset
Not Used	1	100	100	-	-
	2	46	92		
	3	23	69		
	4	13	52		
Subset #1	5	12	60	34	226
	6	5	30		
	7	4	28		
	8	9	72		
	9	4	36		
Subset #2	10	6	60	27	373
	11	4	44		
	12	2	24		
	13	5	65		
	14	1	14		
	15	1	15		
	16	3	48		
	17	1	17		
	18	1	18		
	19	1	19		
	22	1	22		
	27	1	27		
Subset #3	30	2	60	9	526
	33	1	33		
	40	1	40		
	65	1	65		
	66	2	132		
	95	1	95		
	101	1	101		
Total	**28**	**252**	**1438**	**70**	**1125**

does not encode each argument separately). Given the degrees of similarity between the pairs of each group with same event, the procedure calculates the mean square error (*MSE*) between the similarity obtained with each model and the maximum value (1) according to the golden corpus. Besides, for each model, it computes a global MSE for each subset with all events groups. It compute all these measures not only to each subset, but also for the whole dataset selected. With the MSE obtained with each model, the boxplots in the following figures show the distribution of these values of all groups, as well as the global measure against the MSE by group.

Figure 7.8 compares the results obtained as representations of the events with

AutoNEM in relation to the representations of the sentences with each baseline separated by subset, in addition to considering the whole set.



(a) First Subset

(b) Second Subset

(c) Third Subset

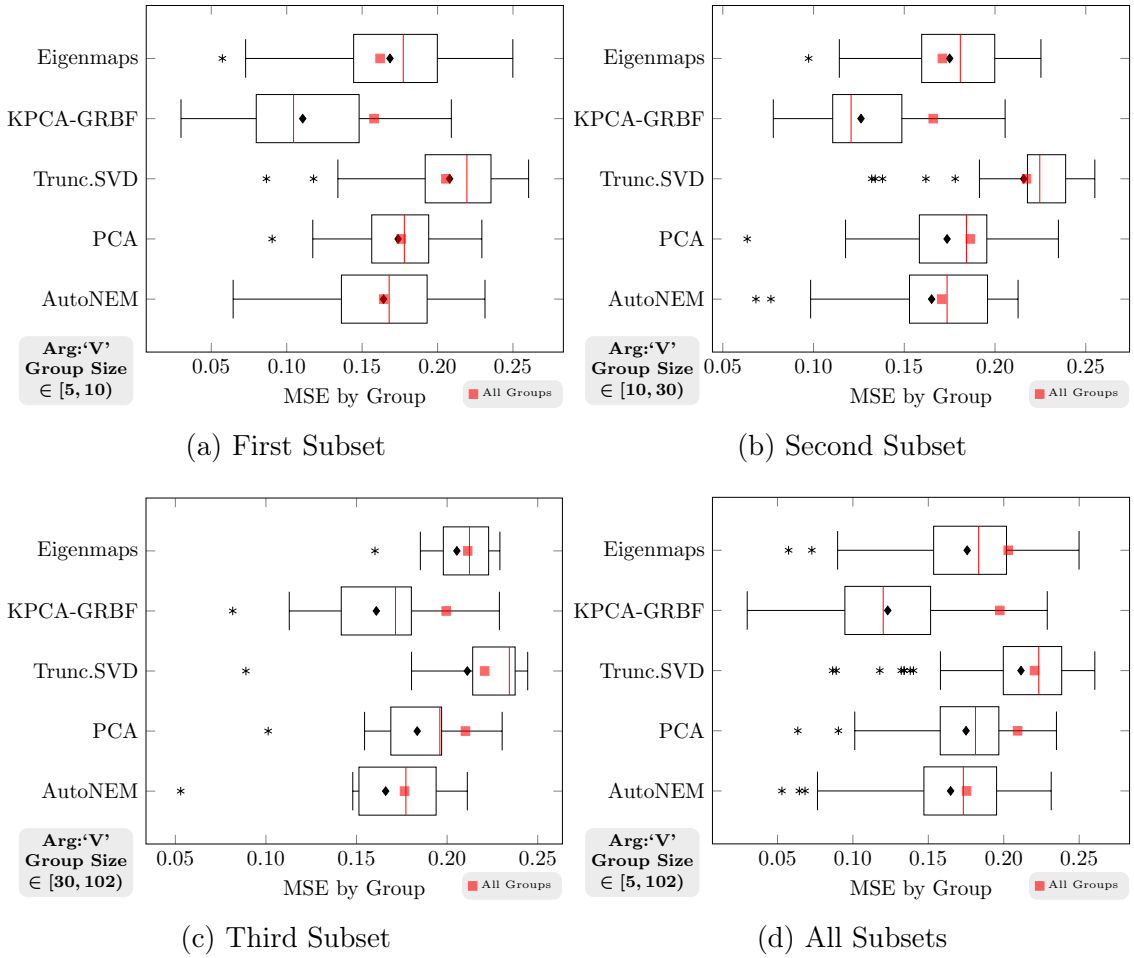
(d) All Subsets

Figure 7.8: Comparisons between the models through MSE from degrees of similarity of the pairs of encoded events (or encoded sentences when the model is a baseline) against the golden corpus.

In each of the comparisons in Figure 7.8, *AutoFlat* overcame all other models including *AutoNEM*, especially in the third subset, composed of the largest groups of events (see Figure 7.8c). In this same subset, *AutoNEM* overcame other linear and nonlinear baselines. In the others subsets (see Figures 7.8a and 7.8b), the *AutoNEM* equaled to the nonlinear *Eigenmaps* model, and, although it did not surpass the others, the *MSE* of some groups presented values close to the *AutoFlat* median. When considering each group of all subsets (see Figure 7.8d) the behavior was similar, but when it observes the *MSE* of all groups, *AutoNEM* exceeds the baselines of the literature. This fact is due to this global measure taking into account the size of each group of events, once *AutoNEM* was better in the subset of larger groups.

Figures 7.9, 7.10, 7.11, 7.12 and 7.13 compares the results obtained as representations of each argument with *AutoNEM* in relation to the its representations with

each baseline separated by subset, considering the whole set too.



(a) First Subset

(b) Second Subset

(c) Third Subset

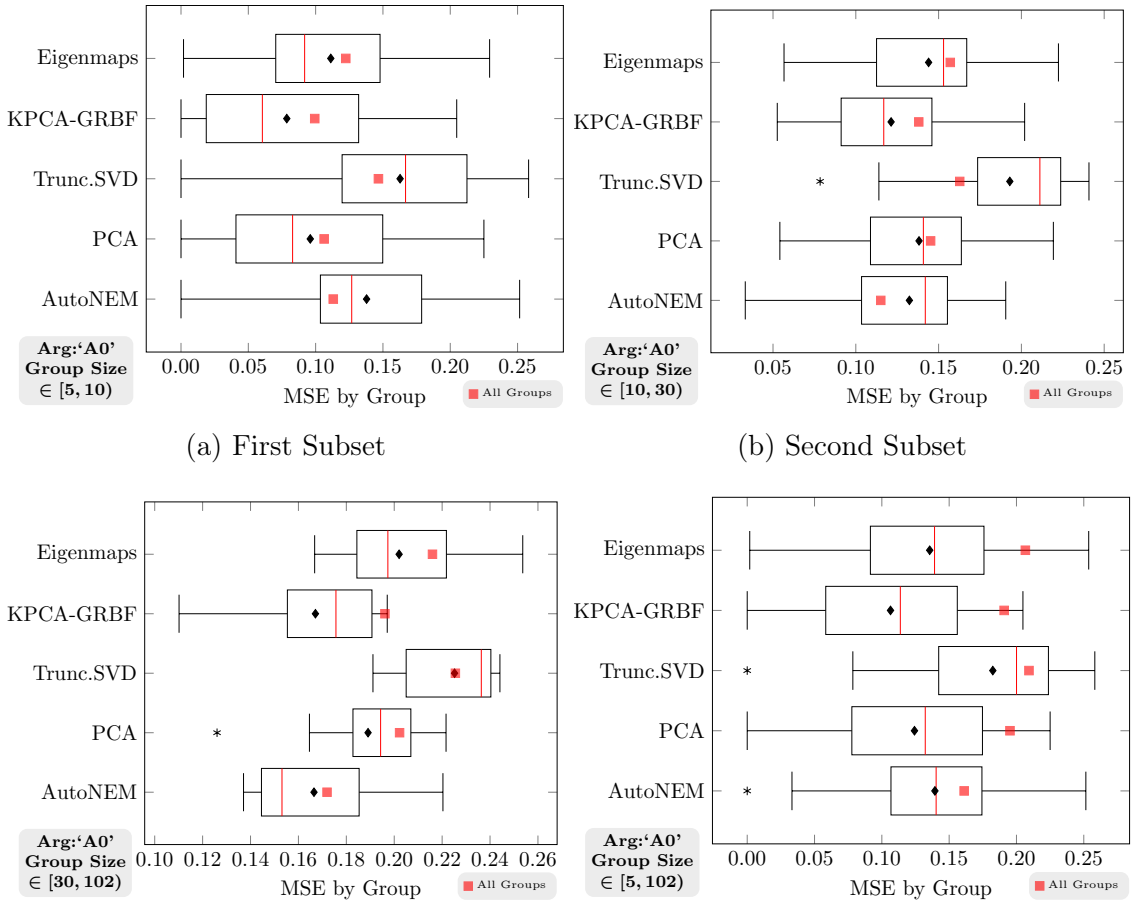
(d) All Subsets

Figure 7.9: Comparisons between the models through MSE from degrees of similarity of the pairs of encoded argument **'V'** against the golden corpus.

On the argument **'V'** (see Figure 7.9), *AutoNEM* equals or even exceeds almost all baselines, with the exception of the nonlinear *KernelPCA-GaussianRBF* model. When looking only at the global measure, *AutoNEM* overcomes the *KernelPCA-GaussianRBF* in subset with larger groups (see Figures 7.9b and 7.9c) and in the set with all groups (see Figure 7.9d), while in the first subset these measures are almost equivalent (see Figure 7.9a).

With the argument **'A0'** (see Figure 7.10), *AutoNEM* was within the baselines distribution (see Figures 7.10a, 7.10b and 7.10d), the only small difference was in the subset with larger groups, where it just was not better than the *KernelPCA-GaussianRBF* (see Figure 7.10c). About the global measure, *AutoNEM* overcame baselines in almost all situations.

With the argument **'A1'** (see Figure 7.11), *AutoNEM* also was within the baselines distribution (see Figures 7.11a, 7.11b and 7.11d), and it was a little better than the *KernelPCA-GaussianRBF* in the subset with larger groups (see Figure 7.11c). Concerning the global measure, *AutoNEM* overcame all baselines in the subset with



(a) First Subset (b) Second Subset

(c) Third Subset (d) All Subsets

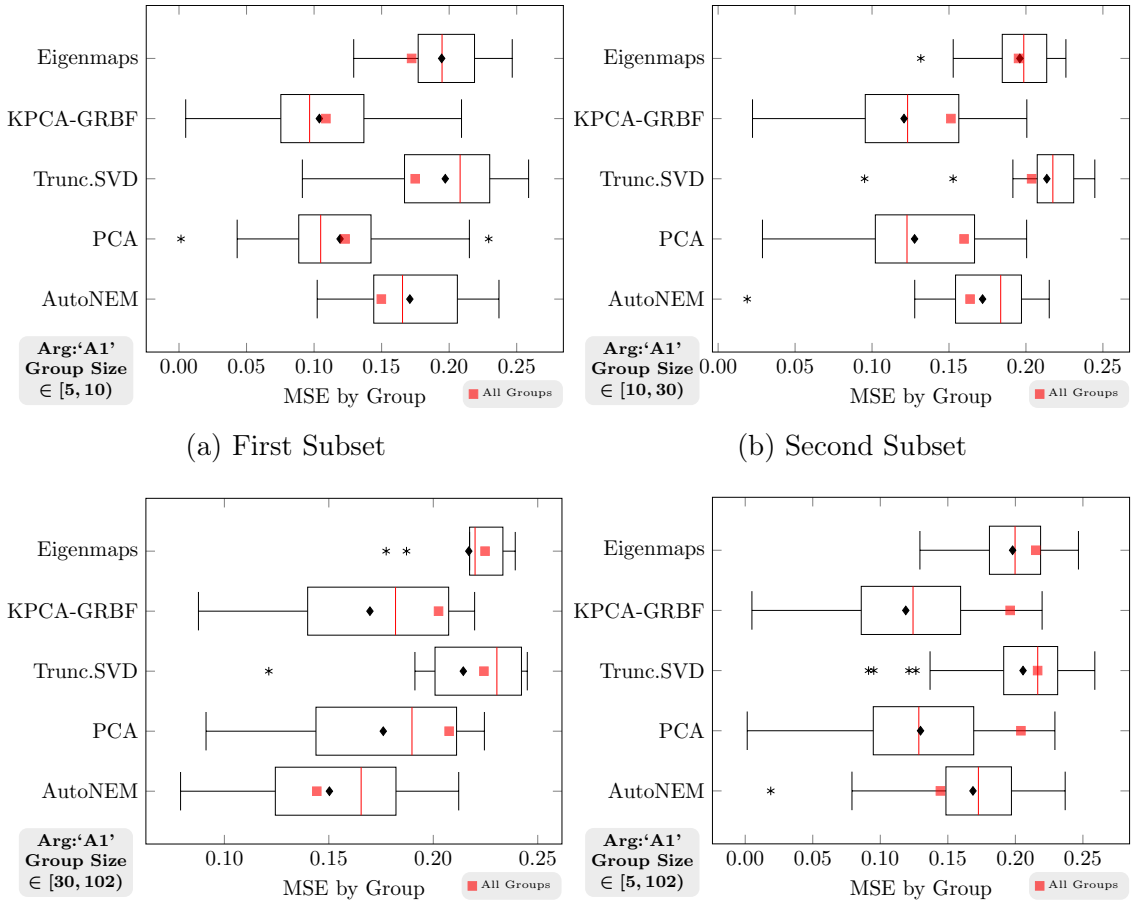
Figure 7.10: Comparisons between the models through MSE from degrees of similarity of the pairs of encoded argument **'A0'** against the golden corpus.

larger groups and in whole set (see Figures 7.11c and 7.11d).

With the argument **'AM-LOC'** (see Figure 7.12), *AutoNEM* was alongside to the baselines in the most cases (see Figures 7.12a, 7.12b and 7.12d), and it was the best in the subset with larger groups (see Figure 7.12c). About the global measure, *AutoNEM* overcame all baselines in the third subset and in whole set (see Figures 7.12c and 7.12d).

On the argument **'AM-TMP'** (see Figure 7.13), *AutoNEM* was equivalent to the baselines (see Figures 7.12a, 7.12c and 7.12d), and a little better in the second subset (see Figure 7.12b). The *AutoNEM* overcame all baselines in most cases with the global measure, except in the first subset (see Figure 7.12a).

In addition to the measures by argument, the experiments added one more comparison with the average of the degrees of similarity of all the arguments present in each pair of news. Figure 7.14 shows this additional comparison. On this extra similarity resulting from every attribute, *AutoNEM* overcomes the *Truncated-SVD* and closes to some baselines of the first and second subsets and the complete set (see Figures 7.14a, 7.14b and 7.14d). In the third subset, it closes to the *KernelPCA*-



(a) First Subset (b) Second Subset

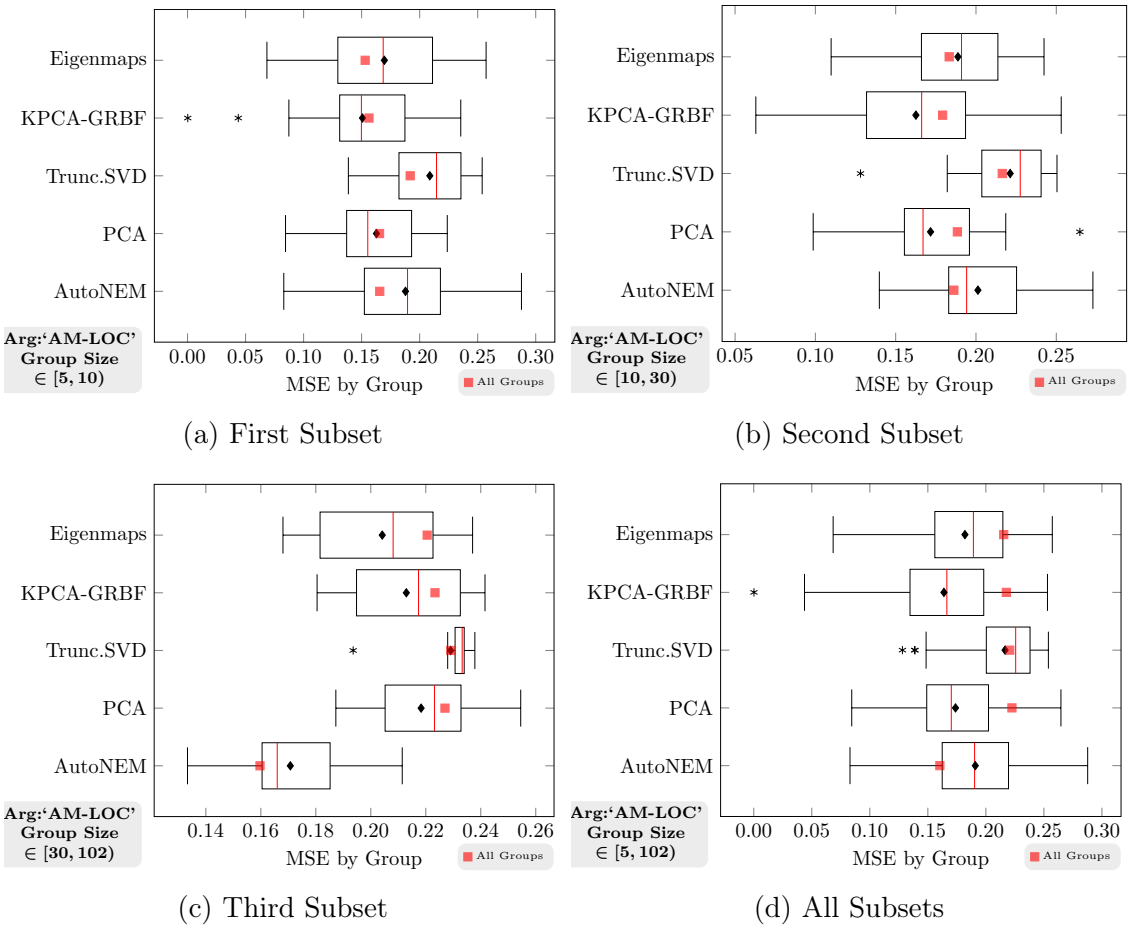
(c) Third Subset (d) All Subsets

Figure 7.11: Comparisons between the models through MSE from degrees of similarity of the pairs of encoded argument **'A1'** against the golden corpus.

GaussianRBF and overcomes other baselines. As for the global measure, *AutoNEM* approximates the best baseline of the second subset (see Figure 7.14b) and outperforms all baselines in the third subset and the whole set (see Figures 7.14c and 7.14d).

In summary, although *AutoNEM* did not exceed all baselines in most situations, in some cases the proposed model was the closest to the golden corpus (especially with the largest groups of events), and in general, the experiments confirmed that the model is competitive with traditional literature approaches such as *PCA*. This competitiveness is accentuated especially by two differentials of *AutoNEM* in relation to the other models. The first is the unawareness of the evaluation data (golden corpus), which characterizes the unsupervised learning. The second is the possibility of overcoming the absence of some of the information contained in the original sentence. These differentials are discussed below.

Because the *AutoNEM* training performs on a separate context of evaluation data, i.e., since it is unsupervised, it is unawareness any data with which it is evaluated. Except for *AutoFlat* that also has unsupervised learning, the other baselines models are prepared directly on the evaluation data (see Section 7.3), and from this



(a) First Subset (b) Second Subset

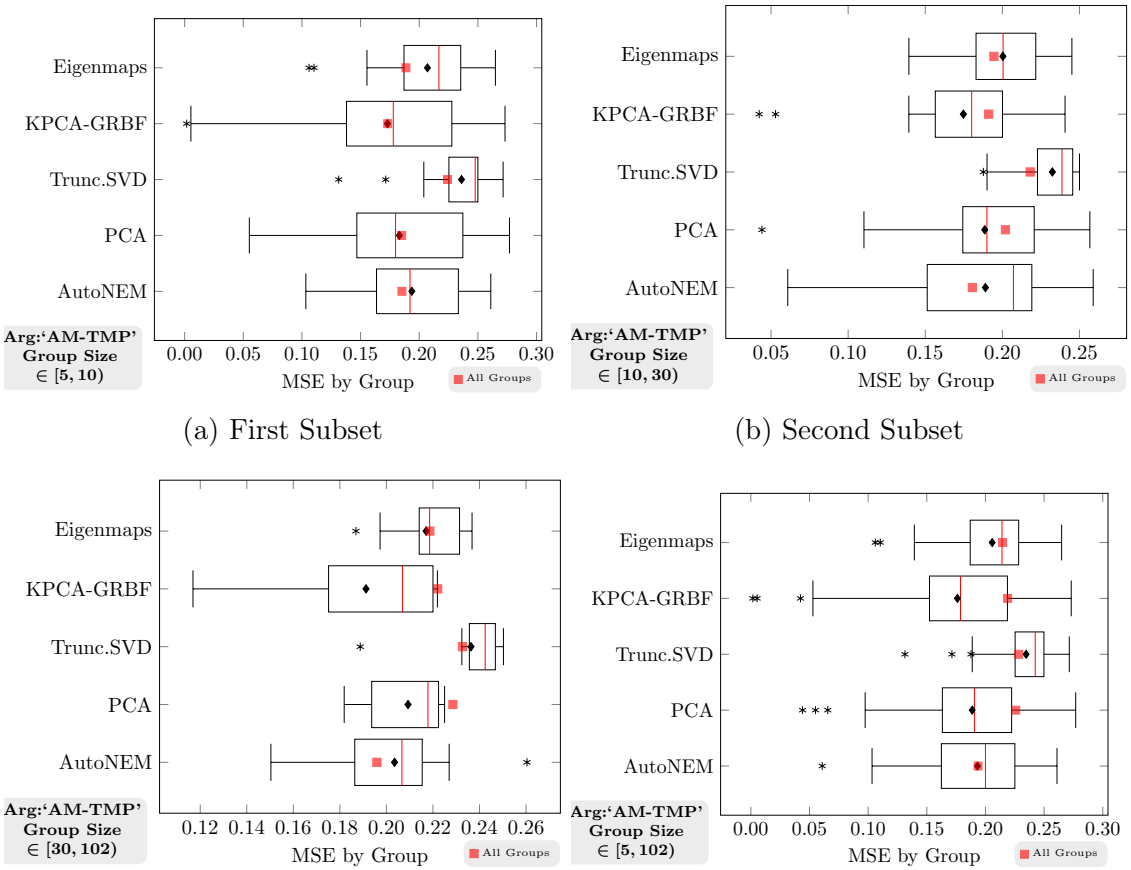
(c) Third Subset (d) All Subsets

Figure 7.12: Comparisons between the models through MSE from degrees of similarity of the pairs of encoded argument **'AM-LOC'** against the golden corpus.

process are obtained the compact representations. So, unlike *AutoNEM*, these four traditional baselines are evaluated under the same context for which they were prepared. Nevertheless, *AutoNEM* was able to get closer and even outperform them in some situations.

Another factor not clarified so far is the information portion of each sentence considered by AutoNEM and by each of the baselines. Since the input of *AutoNEM* is an event structured according to *SENNA*'s arguments, it is common for some tokens of each sentence to compose no extracted argument, unlike the well-behaved examples in Table 6.1. By example, in the sentence *"Eagles' Chris Long puts arm around Malcolm Jenkins during anthem protest"* (**V**: *"puts"*, **A0**: *"Eagles' Chris Long"*, **A1**: *"arm"*, **AM-TMP**: *"during anthem protest"*), the SENNA didn't extract the expression *"around Malcolm Jenkins"*. Therefore, while the five baselines make use of all the information available in each news headline to encode the sentence, not always *AutoNEM* uses all the terms to encode the event. Even with this limitation, *AutoNEM* is competitive.

Beyond to all these comparisons, can also see that some pairs of different events



(a) First Subset (b) Second Subset

(c) Third Subset (d) All Subsets

Figure 7.13: Comparisons between the models through MSE from degrees of similarity of the pairs of encoded argument **'AM-TMP'** against the golden corpus.

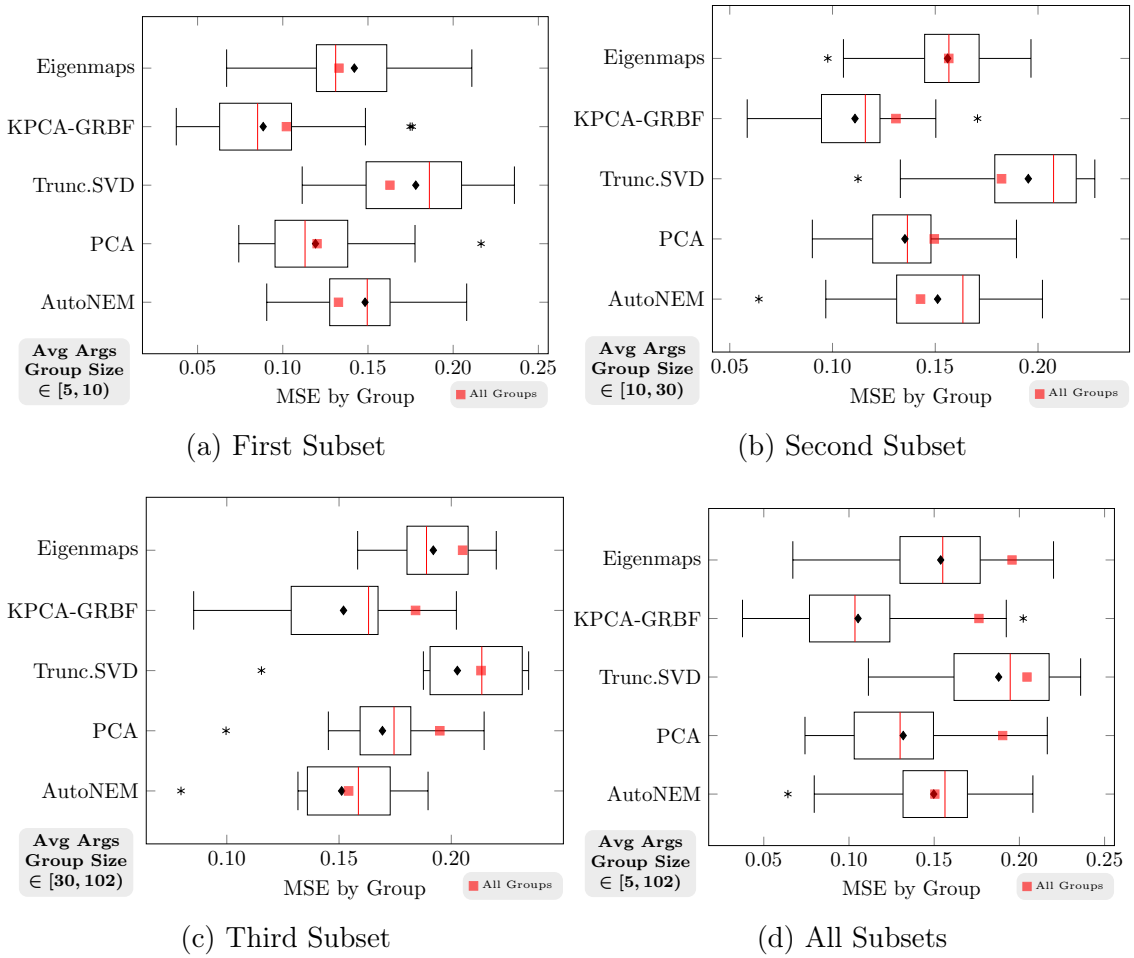
(according to *EventRegistry*) present a degree of similarity relatively higher than expected. For example, consider the events mentioned in the two following news items:

(**i**) *"'Usain Bolt beaten by Justin Gatlin in final 100-meter race."*

- **V**: *"beaten"*
- **A0**: *"by Justin Gatlin"*
- **A1**: *"Usain Bolt"*
- **AM-LOC**: *"in final 100-meter race"*

(**ii**) *"Bolt wins relay heats in his penultimate race."*

- **V**: *"wins"*
- **A0**: *"Bolt"*
- **A1**: *"relay heats"*
- **AM-LOC**: *"in his penultimate race"*



(a) First Subset (b) Second Subset

(c) Third Subset (d) All Subsets

Figure 7.14: Comparisons between the models through MSE from average of similarities of the pairs of the encoded arguments against the golden corpus.

According to *EventRegistry*, these two events are distinct, but the representations of the events obtained with *AutoNEM* allowed us to detect a relevant degree of similarity between them (79% of similarity), which is consistent with a human interpretation. The traditional baselines didn't find a relevant similarity between them. However, for the evaluation of similarity between events beyond the golden corpus, it would be necessary an exhaustive and manually checking all pairs formed by all the unique events, to be carried out in later researches.

All these results evidence that the representations obtained for each event and its attributes received influence from quantity and variety of the training examples, and from word-embeddings applied to these examples. Besides, as 89.8% of golden corpus terms have an intersection with the dictionary of the trained words, so the use of *word-embeddings* as input of *AutoNEM* performed a fundamental role for such results. Thus, since different terms with the same semantics have vector representations very close (i.e., with a high degree of similarity), then the learning process could incorporate to the weights of the autoencoder an approximation of the representations due to the semantic proximity according to the training inputs.

As noted in Section 7.2.2, the training spent much time and may increase according to the number of events detected for the training. However, the process for discovering relationships between new detected events is fast with the trained *AutoNEM Encoder* – sequence of phases: (**i**) detection and structuring of new events, (**iii**) encoding of new events, and (**iv**) relationship rating between new events. Thus, the process allows a relationship analysis between new events detected in real time (online).

Chapter 8

Conclusion and Future

This thesis initially presented a background of event model, analysis event in online media with focus in short text, and deep learning for textual data. After, it reviewed recent works dealing with the discovery of relationships between events, analyzing them under certain requirements. These requirements motivated the development of *AutoNEM*, a neural network model trained in an entirely unsupervised process to discover similarity relationships between structured events (*5W1H*) extracted from short texts of online media. The proposed model is able to provide a compact, semantic and computationally treatable representation for events and their attributes, obtained after the extraction of event attributes with the use of an SRL tool, followed by the training of an auto-encoder with recurrent and dense layers. This representation can be obtained from texts in any domain, but as a consequence of the requirements themselves, it is limited to texts mentioning a single event. From the representations, it is possible to easily apply a similarity function to find out the degree similarity between two events.

The experimental evaluation was realized with data obtained from *EventRegistry*. The experiments used these data as a golden corpus with news grouped into unique events. The results indicated that the degree of relations obtained with representations approaches the golden corpus, in addition to evidencing some degree of similarity in other pairs of events that had not been evidenced by *EventRegistry*.

Besides of applications to *Web Semantic* area, to filtering and organizing information of online media to users, and to improve the engagement of users on news sites, the main contributions of this work are:

1. Identification of a limitations framework targeted at organizing previous works and frame them within it, as well as the framing the proposed approach positively on all desirable properties (requirements).

2. A neural network architecture based on autoencoder targeted at detecting relationships between events detected in short text from online media in an

unsupervised way from an event representation model *5W1H*-based, called *AutoNEM* (*Autoencoder Neural Event Model*). A lot of collected news allowed the training of this model.

3. An unsupervised way to embedding events and each of their attributes *5W1H*-based in a latent semantic space through the encoder phase, called *AutoNEM Encoder*. Such a representation eliminates the difficulty of treating an immense amount of sparse data, besides allows a uniform treatment.

4. An evaluation of the degree of similarity between pairs of events representations obtained with *AutoNEM Encoder*, against a golden corpus and comparing with traditional methods. The results showed that *AutoNEM* is not only effective but also competitive with the traditional baselines, even limited to the terms of extracted attributes.

5. An evaluation of the degree of similarity between pairs of events' attributes representations obtained with *AutoNEM Encoder*, against a golden corpus and comparing with traditional methods. This evaluation confirm that *AutoNEM* allows analyzing the event from the perspective of each attribute separately.

In addition, the proposal still has some limitations and thus opens up opportunities for future contributions, such as:

- It is necessary conducting an additional and expensive experimental evaluation to solidify the model, through the definition of a new golden corpus with the degree of similarity between the pairs of several events according to the human interpretation. Thus, it is possible to make an evaluation of the quality of similarity degree.

- Future analyzes the impact of the number of examples for training can recognize how many examples are needed for the training a good model.

- As mentioned at the end of Chapter 3 and Section 6.2.1, the process of detecting and structuring events can be improved to get a better characterization of the event from the perspective of *5W1H*. Thus, the presented process allows the replacement of the *SENNA* tool used in this phase – which was designed to extract semantic roles and not to characterize an event – by a set of the best techniques for extracting each *5W1H*-attribute.

- With the use of other techniques for the extraction of attributes, variations of the attributes amount K used can be tested to investigate the impact on the trained model, like $K < 4$ or $K > 5$.

- As the proposal is limited to the use of sentences with only a single tuple of arguments, improvements can be proposed on the architecture to support a variable amount of tuples. Besides, it is desired to analyze the effect of some params (e.g., or the number of extracted tuples) and events properties (e.g., differences in the temporal attributes of the training data to the evaluation data) from these two last future improvements, like in KENTER & DE RIJKE (2015).

- This model can be improvement with recursive neural networks, or shared weights between encoder and decoder layers.

- Future works can explore the others possible combinations of representations of arguments and events obtained in intermediate layers of *AutoNEM*, trying to improve the quantifying of the degree of similarity.

- Another natural venue for future work is to investigate other forms of relationships between events (e.g., mereology, causality, correlation). SANTORO *et al.* (2017); ZHENG *et al.* (2016, 2017) can inspire solutions in this directions.

- This work can also support the development or improvement of techniques in adjacent areas, like *process mining* or *complex event processing*. Or simply support common tasks with textual data, like clustering (CHEN & ZAKI, 2017).

- The process can consider an environment to store big data with support to deep learning tools (e.g., *Spark*).

Bibliography

ABDELHAQ, H., GERTZ, M., SENGSTOCK, C., 2013a, "Spatio-temporal Characteristics of Bursty Words in Twitter Streams". In: *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, SIGSPATIAL'13, pp. 194–203, Orlando, FL, a. ACM. ISBN: 978-1-4503-2521-9. https://doi.org/10.1145/2525314.2525354 .

ABDELHAQ, H., SENGSTOCK, C., GERTZ, M., 2013b, "EvenTweet: Online Localized Event Detection from Twitter", *Proc. VLDB Endow.*, v. 6, n. 12, pp. 1326–1329. http://dl.acm.org/citation.cfm?id=2536274.2536307 .

AGARWAL, P., VAITHIYANATHAN, R., SHARMA, S., et al., 2012, "Catching the Long-Tail: Extracting Local News Events from Twitter". In: *Sixth International AAAI Conference on Weblogs and Social Media*, pp. 379–382, Dublin, Ireland. AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4639http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4639/5011 .

AL-SMADI, M., JARADAT, Z., AL-AYYOUB, M., et al., 2017, "Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features", *Information Processing & Management*, v. 53, n. 3 (may), pp. 640–652. ISSN: 03064573. https://doi.org/10.1016/j.ipm.2017.01.002 .

ALLAN, J., PAPKA, R., LAVRENKO, V., 1998, "On-line New Event Detection and Tracking". In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, pp. 37–45, Melbourne, Australia. ACM. ISBN: 1581130155. https://doi.org/10.1145/290941.290954 .

APPAN, P., SUNDARAM, H., 2004, "Networked Multimedia Event Exploration". In: *Proceedings of the 12th Annual ACM International Conference on Multimedia - MULTIMEDIA '04*, pp. 40–47, New York, NY, USA, oct.

ACM. ISBN: 1581138938. https://doi.org/10.1145/1027527.1027536 .

ARAMAKI, E., MASKAWA, S., MORITA, M., 2011, "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter". In: *Proceedings of the 2011 Conference on Emperical Methods in Natural Language Processing - EMNLP'11*, pp. 1568–1576, Edinburgh, United Kingdom. Association for Computational Linguistics. ISBN: 978-1-937284-11-4. http://dl.acm.org/citation.cfm?id=2145600 .

ARAPAKIS, I., LALMAS, M., CEYLAN, H., et al., 2014, "Automatically embedding newsworthy links to articles: From implementation to evaluation", *Journal of the Association for Information Science and Technology*, v. 65, n. 1 (jan), pp. 129–145. ISSN: 23301635. https://doi.org/10.1002/asi.22959 .

ATEFEH, F., KHREICH, W., 2015, "A Survey of Techniques for Event Detection in Twitter", *Computational Intelligence*, v. 31, n. 1 (feb), pp. 132–164. ISSN: 08247935. https://doi.org/10.1111/coin.12017 .

BECKER, H., NAAMAN, M., GRAVANO, L., 2011a, "Beyond Trending Topics: Real-World Event Identification on Twitter". In: *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 1–17, Barcelona, Spain, a. AAAI Press. ISBN: 9781605588896. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2745 .

BECKER, H., NAAMAN, M., GRAVANO, L., 2011b, "Beyond Trending Topics: Real-World Event Identification on Twitter", *Columbia University Computer Science Technical Reports*. http://hdl.handle.net/10022/AC:P:10668 .

BENGIO, Y., 2011, "Deep Learning of Representations for Unsupervised and Transfer Learning", *JMLR: Workshop and Conference Proceedings*, v. 7, pp. 1–20. ISSN: 1938-7228. https://doi.org/10.1109/IJCNN.2011.6033302 .

BENGIO, Y., 2012, "Deep Learning of Representations for Unsupervised and Transfer Learning". In: Guyon, I., Dror, G., Lemaire, V., et al. (Eds.), *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, v. 27, *Proceedings of Machine Learning Research*, pp. 17–36, Bellevue, Washington, USA. PMLR. http://proceedings.mlr.press/v27/bengio12a.html .

BENGIO, Y., DUCHARME, R., VINCENT, P., 2001, "A Neural Probabilistic Language Model". In: Leen, T. K., Dietterich, T. G., Tresp, V. (Eds.), *Advances in Neural Information Processing Systems*, v. 13, MIT Press, pp. 932–938, Vancouver, Canada, dec. http://papers.nips.cc/paper/1839-a-neural-probabilistic-language-model.pdf .

BENGIO, Y., DUCHARME, R., VINCENT, P., et al., 2003, "A Neural Probabilistic Language Model", *Journal of Machine Learning Research – JMLR*, v. 3 (feb), pp. 1137–1155. ISSN: 1532-4435. http://dl.acm.org/citation.cfm?id=944919.944966 .

BENGIO, Y., COURVILLE, A., VINCENT, P., 2013, "Representation Learning: A Review and New Perspectives", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 35, n. 8 (aug), pp. 1798–1828. ISSN: 0162-8828. https://doi.org/10.1109/TPAMI.2013.50 .

BERNERS-LEE, T., HENDLER, J., LASSILA, O., 2001, "The Semantic Web", *Scientific American*, (may), pp. 29–37. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 .

BOUILLOT, F., PONCELET, P., ROCHE, M., 2012, "How and why exploit tweet's location information?" In: *15th AGILE Conference on Geographic Information Science - AGILE'2012*, pp. 24–27, Avignon, France. ISBN: 9789081696005. http://hal-lirmm.ccsd.cnrs.fr/lirmm-00723570 .

BRAHA, D., 2012, "Global Civil Unrest: Contagion, Self-Organization, and Prediction", *PLoS ONE*, v. 7, n. 10, pp. e48596. ISSN: 19326203. https://doi.org/10.1371/journal.pone.0048596 .

CASATI, R., VARZI, A., 2015. "Events". https://plato.stanford.edu/archives/win2015/entries/events/ .

CERVESATO, I., MONTANARI, A., 2000, "A calculus of macro-events: progress report". In: *Proceedings Seventh International Workshop on Temporal Representation and Reasoning. TIME 2000*, pp. 47–58, Cape Breton, NS. IEEE. ISBN: 0-7695-0756-5. https://doi.org/10.1109/TIME.2000.856584 .

CHANDY, K. M., CHARPENTIER, M., CAPPONI, A., 2007, "Towards a Theory of Events". In: *Proceedings of the 2007 Inaugural International Conference on Distributed Event-based Systems - DEBS'07*, pp. 180–187, Toronto, Ontario, Canada. ACM. ISBN: 9781595936653. https://doi.org/10.1145/1266894.1266929 .

CHANLEKHA, H., COLLIER, N., 2010, "Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports." *Journal of biomedical semantics*, v. 1, n. 1, pp. 3. ISSN: 2041-1480. https://doi.org/10.1186/2041-1480-1-3 .

CHAUDET, H., 2006, "Extending the event calculus for tracking epidemic spread", *Artificial Intelligence in Medicine*, v. 38, n. 2, pp. 137–156. ISSN: 09333657. https://doi.org/10.1016/j.artmed.2005.06.001 .

CHEN, F., ARREDONDO, J., KHANDPUR, R. P., et al., 2012, "Spatial Surrogates to Forecast Social Mobilization and Civil Unrests", *Position Paper in CCC Workshop on "From GPS and Virtual Globes to Spatial Computing-2020"*, pp. 1–3. http://people.cs.vt.edu/naren/papers/CCC-VT-Updated-Version.pdf .

CHEN, G., KONG, Q., MAO, W., 2017, "Online event detection and tracking in social media based on neural similarity metric learning". In: *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, IDEAS 2017, pp. 182–184, New York, NY, USA, jul. IEEE. ISBN: 978-1-5090-6727-5. https://doi.org/10.1109/ISI.2017.8004905 .

CHEN, H., FININ, T., JOSHI, A., 2003, "An ontology for context-aware pervasive computing environments", *The Knowledge Engineering Review*, v. 18, n. 3 (sep.), pp. 197–207. ISSN: 1469-8005. https://doi.org/10.1017/S0269888904000025 .

CHEN, H., FININ, T., JOSHI, A., 2005, "The SOUPA Ontology for Pervasive Computing". In: Tamma, V., Cranefield, S., Finin, T. W., et al. (Eds.), *Ontologies for Agents: Theory and Experiences*, Whitestein Series in Software Agent Technologies, Birkhäuser-Verlag, pp. 233–258, Basel. ISBN: 978-3-7643-7361-0. https://doi.org/10.1007/3-7643-7361-X_10 .

CHEN, L., ROY, A., 2009, "Event Detection from Flickr Data Through Wavelet-based Spatial Analysis". In: *Proceedings of the 18th ACM conference on Information and knowledge management - CIKM'09*, pp. 523–532, Hong Kong, China. ACM. ISBN: 9781605585123. https://doi.org/10.1145/1645953.1646021 .

CHEN, Y., ZAKI, M. J., 2017, "KATE: K-Competitive Autoencoder for Text". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, pp. 85–94, New York, New York, USA. ACM Press. ISBN: 9781450348874. https://doi.org/10.1145/3097983.3098017 .

CHO, K., VAN MERRIENBOER, B., GULCEHRE, C., et al., 2014, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Stroudsburg, PA, USA. Association for Computational Linguistics. ISBN: 9781937284961. https://doi.org/10.3115/v1/D14-1179 .

COLLOBERT, R., 2011, "Deep Learning for Efficient Discriminative Parsing". In: Gordon, G., Dunson, D., Dudík, M. (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, v. 15, *Proceedings of Machine Learning Research*, pp. 224–232, Fort Lauderdale, FL, USA, apr. JMLR. http://proceedings.mlr.press/v15/collobert11a.html .

COLLOBERT, R., WESTON, J., BOTTOU, L., et al., 2011, "Natural Language Processing (Almost) from Scratch", *Journal of Machine Learning Research*, v. 12 (nov), pp. 2493–2537. ISSN: 1532-4435. http://dl.acm.org/citation.cfm?id=1953048.2078186 .

COMPTON, R., LEE, C., LU, T.-C., et al., 2013, "Detecting future social unrest in unprocessed Twitter data". In: *2013 IEEE International Conference on Intelligence and Security Informatics*, pp. 56–60, Seattle, WA, USA. IEEE. ISBN: 9781467362139. https://doi.org/10.1109/ISI.2013.6578786 .

CROITORU, A. A., CROOKS, A. A., RADZIKOWSKI, J. J., et al., 2013, "Geosocial gauge: a system prototype for knowledge discovery from social media", *International Journal of Geographical Information Science*, v. 27, n. 12 (dec.), pp. 2483–2508. ISSN: 1365-8816. https://doi.org/10.1080/13658816.2013.825724 .

CUNNINGHAM, H., 2006. "Information Extraction, Automatic". https://gate.ac.uk/sale/ell2/ie/preprint.pdf .

CUNNINGHAM, H., BONTCHEVA, K., LI, Y., 2005, "Knowledge management and human language: crossing the chasm", *Journal of Knowledge Management*, v. 9, n. 5 (oct.), pp. 108–131. ISSN: 1367-3270. https://doi.org/10.1108/13673270510622492 .

DAI, A. M., LE, Q. V., 2015, "Semi-supervised Sequence Learning". In: Cortes, C., Lawrence, N. D., Lee, D. D., et al. (Eds.), *Advances in Neural Information Processing Systems (NIPS)*, v. 28, Curran Associates,

Inc., pp. 3079–3087, nov. http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning .

DASIGI, P., HOVY, E., 2014, "Modeling Newswire Events using Neural Networks for Anomaly Detection". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1414–1422, Dublin, Ireland, aug. Dublin City University and Association for Computational Linguistics. ISBN: 9781941643266. http://www.aclweb.org/anthology/C14-1134 .

DAUME, S., ALBERT, M., VON GADOW, K., 2014, "Forest monitoring and social media – Complementary data sources for ecosystem surveillance?" *Forest Ecology and Management*, v. 316, pp. 9–20. ISSN: 0378-1127. https://doi.org/https://doi.org/10.1016/j.foreco.2013.09.004 .

DE BOOM, C., VAN CANNEYT, S., DEMEESTER, T., et al., 2016, "Representation learning for very short texts using weighted word embedding aggregation", *Pattern Recognition Letters*, v. 80 (sep), pp. 150–156. ISSN: 01678655. doi: 10.1016/j.patrec.2016.06.012. https://doi.org/10.1016/j.patrec.2016.06.012 .

DE MULDER, W., BETHARD, S., MOENS, M.-F., 2015, "A survey on the application of recurrent neural networks to statistical language modeling", *Computer Speech & Language*, v. 30, n. 1 (mar), pp. 61–98. ISSN: 08852308. https://doi.org/10.1016/j.csl.2014.09.005 .

DEVI, M. U., GANDHI, G. M., 2015, "An Enhanced Fuzzy Clustering and Expectation Maximization Framework based Matching Semantically Similar Sentences", *Procedia Computer Science*, v. 57, n. 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), pp. 1149–1159. ISSN: 18770509. https://doi.org/10.1016/j.procs.2015.07.406 .

DO CARMO, R. R. M., SOARES, L. F., CASANOVA, M. A., 2013, "Nested Event Model for Multimedia Narratives". In: *2013 IEEE International Symposium on Multimedia*, pp. 106–113. IEEE, dec. ISBN: 978-1-4799-2171-3. https://doi.org/10.1109/ISM.2013.26 .

DOERR, M., ORE, C.-E., STEAD, S., 2007, "The CIDOC Conceptual Reference Model - A New Standard for Knowledge Sharing". In: Grundy, J., Hartmann, S., Laender, A. H. F., et al. (Eds.), *ER'07: Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling*, v. 83, *CRPIT*, pp. 51–56,

Auckland, New Zealand, nov. Australian Computer Society, Inc. http://crpit.com/abstracts/CRPITV83Doerr.html .

DONG, X., MAVROEIDIS, D., CALABRESE, F., et al., 2015, "Multiscale event detection in social media", *Data Mining and Knowledge Discovery*, v. 29, n. 5 (sep), pp. 1374–1405. ISSN: 1384-5810. https://doi.org/10.1007/s10618-015-0421-2 .

DOS SANTOS, R. F., BOEDIHARDJO, A., SHAH, S., et al., 2016, "The big data of violent events: algorithms for association analysis using spatio-temporal storytelling", *GeoInformatica*, v. 20, n. 4 (oct), pp. 879–921. ISSN: 1384-6175. https://doi.org/10.1007/s10707-016-0247-0 .

DOU, W., WANG, X., RIBARSKY, W., et al., 2012a, "Event Detection in Social Media Data". In: *Proceedings of the IEEE VisWeek workshop on interactive visual text analytics – task driven analytics of social media content*, pp. 971–980, octa. https://webpages.uncc.edu/xwang25/pubs/2012/Dou-EventDetectionTasks-2012.pdf .

DOU, W., WANG, X., SKAU, D., et al., 2012b, "LeadLine: Interactive visual analysis of text data through event identification and exploration". In: *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pp. 93–102, b. https://doi.org/10.1109/VAST.2012.6400485 .

DRURY, B., ROCHA, C., MOURA, M.-F., et al., 2016, "The Extraction from News Stories a Causal Topic Centred Bayesian Graph for Sugarcane". In: *Proceedings of the 20th International Database Engineering & Applications Symposium on - IDEAS '16*, IDEAS '16, pp. 364–369, New York, New York, USA. ACM Press. ISBN: 9781450341189. https://doi.org/10.1145/2938503.2938521 .

EKIN, A., TEKALP, A. M., MEHROTRA, R., 2004, "Integrated Semantic-Syntactic Video Modeling for Search and Browsing", *IEEE Transactions on Multimedia*, v. 6, n. 6 (dec.), pp. 839–851. ISSN: 15209210. https://doi.org/10.1109/TMM.2004.837238 .

EL-KILANY, A., EL TAZI, N., EZZAT, E., 2017, "Building Relation Extraction Templates via Unsupervised Learning". In: *Proceedings of the 21st International Database Engineering & Applications Symposium on - IDEAS 2017*, IDEAS 2017, pp. 228–234, New York, New York, USA. ACM Press. ISBN: 9781450352208. https://doi.org/10.1145/3105831.3105845 .

EVANS, C., 1990, "The macro-event calculus: representing temporal granularity". In: *The 01st Pacific Rim International Conference on Artificial Intelligence – PRICAI'90*, Nagoya, Japan, nov. OHMSHA, LTD.

FELLBAUM, C., 1998, *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, MA, MIT Press. ISBN: 9780262061971. http://mitpress.mit.edu/books/wordnet .

FERREIRA, R., LINS, R. D., SIMSKE, S. J., et al., 2016, "Assessing sentence similarity through lexical, syntactic and semantic analysis", *Computer Speech & Language*, v. 39 (sep), pp. 1–28. ISSN: 08852308. https://doi.org/10.1016/j.csl.2016.01.003 .

FRANÇOIS, A. R. J., NEVATIA, R., HOBBS, J., et al., 2005, "VERL: An Ontology Framework for Representing and Annotating Video Events", *IEEE Multimedia*, v. 12, n. 4 (oct.), pp. 76–86. ISSN: 1070986X. https://doi.org/10.1109/MMUL.2005.87 .

FURLAN, B., BATANOVIĆ, V., NIKOLIĆ, B., 2013, "Semantic similarity of short texts in languages with a deficient natural language processing support", *Decision Support Systems*, v. 55, n. 3 (jun), pp. 710–719. ISSN: 01679236. https://doi.org/10.1016/j.dss.2013.02.002 .

GAL, Y., GHAHRAMANI, Z., 2016, "A Theoretically Grounded Application of Dropout in Recurrent Neural Networks". In: Lee, D. D., Sugiyama, M., Luxburg, U. V., et al. (Eds.), *Advances in Neural Information Processing Systems (NIPS)*, v. 29, Curran Associates, Inc., pp. 1019–1027, Barcelona, Spain. http://papers.nips.cc/paper/6241-a-theoretically-grounded-application-of-dropout-in-recurrent-neural-networks .

GAO, Y., LIU, M., LUO, X., et al., 2015, "News of Atomic Events Time Sequence Relationship Recognition Based on Function Word and Predicate Co-occurrence". In: *2015 Seventh International Conference on Measuring Technology and Mechatronics Automation*, pp. 701–703. IEEE, jun. ISBN: 978-1-4673-7143-8. https://doi.org/10.1109/ICMTMA.2015.174 .

GAO, Y., ZHANG, H., ZHAO, X., et al., 2017, "Event Classification in Microblogs via Social Tracking", *ACM Transactions on Intelligent Systems and Technology*, v. 8, n. 3 (feb), pp. 1–14. ISSN: 21576904. https://doi.org/10.1145/2967502 .

GKALELIS, N., MEZARIS, V., KOMPATSIARIS, I., 2010, "A joint content-event model for event-centric multimedia indexing". In: *Proceedings - 2010 IEEE 4th International Conference on Semantic Computing, ICSC 2010*, pp. 79–84, Pittsburgh, PA. IEEE. ISBN: 9780769541549. https://doi.org/10.1109/ICSC.2010.21 .

GONG, B., SINGH, R., JAIN, R., 2004, "Researchexplorer: Gaining Insights Through Exploration in Multimedia Scientific Data". In: *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval - MIR '04*, p. 7, New York, NY, USA. ACM. ISBN: 1581139403. https://doi.org/10.1145/1026711.1026714 .

GOODFELLOW, I., BENGIO, Y., COURVILLE, A., 2016, *Deep Learning*. MIT Press. http://www.deeplearningbook.org/ .

GRAFF, D., CIERI, C., 2003. "English Gigaword LDC2003T05". jan. https://catalog.ldc.upenn.edu/ldc2003t05 .

GRAHAM, M., HALE, S. A., GAFFNEY, D., 2014, "Where in the world are you? Geolocation and language identification in Twitter", *The Professional Geographer*, v. 66, n. 4 (aug), pp. 568–578. ISSN: 00330124. https://doi.org/10.1080/00330124.2014.907699 .

GU, Y., QIAN, Z. S., CHEN, F., 2016, "From Twitter to detector: Real-time traffic incident detection using social media data", *Transportation Research Part C: Emerging Technologies*, v. 67 (jun), pp. 321–342. ISSN: 0968090X. https://doi.org/10.1016/j.trc.2016.02.011 .

GUILLE, A., FAVRE, C., 2014, "Mention-anomaly-based Event Detection and Tracking in Twitter". In: *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pp. 375–382, Beijing. IEEE. ISBN: 9781479958771. https://doi.org/10.1109/ASONAM.2014.6921613 .

GUPTA, A., JAIN, R., 2011, "Managing Event Information: Modeling, Retrieval, and Applications", *Synthesis Lectures on Data Management*, v. 3, n. 4 (jul), pp. 1–141. ISSN: 2153-5418. https://doi.org/10.2200/S00374ED1V01Y201107DTM019 .

GUPTA, M., ZHAO, P., HAN, J., 2012, "Evaluating Event Credibility on Twitter". In: *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 153–164. https://doi.org/10.1137/1.9781611972825.14 .

HAKAMI, H., BOLLEGALA, D., 2017, "Compositional Approaches for Representing Relations Between Words: A Comparative Study", *Knowledge-Based Systems*, v. 136 (sep), pp. 172–182. ISSN: 09507051. https://doi.org/10.1016/j.knosys.2017.09.008 .

HAKEEM, A., SHEIKH, Y., SHAH, M., 2004, "CASE^{E}: A Hierarchical Event Representation for the Analysis of Videos". In: *Proceedings of the Nineteenth National Conference on Artificial Intelligence, AAAI*, pp. 263–268, San Jose, California, jul. AAAI Press. http://www.aaai.org/Library/AAAI/2004/aaai04-042.php .

HALKO, N., MARTINSSON, P. G., TROPP, J. A., 2011, "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions", *SIAM Review*, v. 53, n. 2 (jan), pp. 217–288. ISSN: 0036-1445. https://doi.org/10.1137/090771806 .

HASAN, M., ORGUN, M. A., SCHWITTER, R., 2017, "A survey on real-time event detection from the Twitter data stream", *Journal of Information Science*, (mar), pp. 016555151769856. ISSN: 0165-5515. https://doi.org/10.1177/0165551517698564 .

HE, Q., CHANG, K., LIM, E.-P., 2007, "Analyzing feature trajectories for event detection". In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, p. 207, Amsterdam, The Netherlands. ACM Press. ISBN: 9781595935977. https://doi.org/10.1145/1277741.1277779 .

HECHT, B., HONG, L., SUH, B., et al., 2011, "Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles". In: *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI'11*, p. 237, Vancouver, BC, Canada. ACM. ISBN: 9781450302289. https://doi.org/10.1145/1978942.1978976 .

HILL, F., CHO, K., KORHONEN, A., 2016, "Learning Distributed Representations of Sentences from Unlabelled Data". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1367–1377, Stroudsburg, PA, USA. Association for Computational Linguistics. ISBN: 9781941643914. https://doi.org/10.18653/v1/N16-1162 .

HIRUTA, S., YONEZAWA, T., JURMU, M. B., et al., 2012, "Detection, classification and visualization of place-triggerd geotagged tweets". In: *UbiComp'12*

- *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp'12, pp. 956–963, Pittsburgh, PA. ACM. ISBN: 978-1-4503-1224-0. https://doi.org/10.1145/2370216.2370427 .

HOCHREITER, S., SCHMIDHUBER, J., 1997, "Long Short-Term Memory", *Neural Computation*, v. 9, n. 8 (nov), pp. 1735–1780. ISSN: 0899-7667. https://doi.org/10.1162/neco.1997.9.8.1735 .

HUA, T., CHEN, F., ZHAO, L., et al., 2013, "STED: Semi-supervised Targeted-interest Event Detection in Twitter". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'13, pp. 1466–1469, New York, NY, USA. ACM. ISBN: 978-1-4503-2174-7. https://doi.org/10.1145/2487575.2487712 .

IGLESIAS, J. A., TIEMBLO, A., LEDEZMA, A., et al., 2016, "Web news mining in an evolving framework", *Information Fusion*, v. 28 (mar), pp. 90–98. ISSN: 15662535. https://doi.org/10.1016/j.inffus.2015.07.004 .

IMRAN, M., CASTILLO, C., DIAZ, F., et al., 2015, "Processing Social Media Messages in Mass Emergency: A Survey", *ACM Computing Surveys*, v. 47, n. 4 (jun.), pp. 67:1–67:38. ISSN: 03600300. https://doi.org/10.1145/2771588 .

2012, *EventML-G2 – IPTC Standard*. IPTC International Press Telecommunications Council. https://iptc.org/standards/eventsml-g2/ .

JAIN, R., 2008, "EventWeb: Developing a Human-Centered Computing System", *Computer*, v. 41, n. 2, pp. 42–50. ISSN: 00189162. https://doi.org/10.1109/MC.2008.49 .

JIANBO SHI, MALIK, J., 2000, "Normalized cuts and image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 8, pp. 888–905. ISSN: 01628828. https://doi.org/10.1109/34.868688 .

JOHNSON, N., CARRAN, S., BOTNER, J., et al., 2011, "Pattern in escalations in insurgent and terrorist activity", *Science (New York, N.Y.)*, v. 333, n. 6038, pp. 81–84. ISSN: 0036-8075. https://doi.org/10.1126/science.1205068 .

KALEEL, S. B., ABHARI, A., 2015, "Cluster-discovery of Twitter messages for event detection and trending", *Journal of Computational Science*, v. 6 (jan), pp. 47–57. ISSN: 18777503. https://doi.org/10.1016/j.jocs.2014.11.004 .

KAMYSHANSKA, H., 2013, *Autoencoder Scoring*. Ph.D. Thesis, Goethe University Frankfurt. https://fias.uni-frankfurt.de/{~}kamyshanska/pubs/aescoring{_}thesis.pdf .

KAMYSHANSKA, H., MEMISEVIC, R., 2013, "On autoencoder scoring". In: Dasgupta, S., McAllester, D. (Eds.), *Proceedings of The 30th International Conference on Machine Learning*, v. 28, *Proceedings of Machine Learning Research*, pp. 1757–1765, Atlanta, Georgia, USA, jun. PMLR. http://proceedings.mlr.press/v28/kamyshanska13.html .

KAMYSHANSKA, H., MEMISEVIC, R., 2015, "The Potential Energy of an Autoencoder", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 37, n. 6 (jun), pp. 1261–1273. ISSN: 0162-8828. https://doi.org/10.1109/TPAMI.2014.2362140 .

KANHABUA, N., NEJDL, W., 2013, "Understanding the Diversity of Tweets in the Time of Outbreaks". In: *Proceedings of the 22Nd International Conference on World Wide Web Companion - WWW'13 Companion*, pp. 1335–1342, Rio de Janeiro, RJ, Brazil. International World Wide Web Conferences Steering Committee. ISBN: 9781450320382. http://dl.acm.org/citation.cfm?id=2487788.2488172 .

KANHABUA, N., ROMANO, S., STEWART, A., 2012a, "Identifying Relevant Temporal Expressions for Real-World Events". In: *Proceedings of the SIGIR 2012 Workshop on Time-aware Information Access - TAIA'12*, Portland, Oregon, USA, a. ACM Press. http://research.microsoft.com/en-us/people/milads/kanhabua-taia2012.pdf .

KANHABUA, N., ROMANO, S., STEWART, A., et al., 2012b, "Supporting Temporal Analytics for Health-related Events in Microblogs". In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM'12, pp. 2686–2688, New York, NY, USA, b. ACM. ISBN: 978-1-4503-1156-4. https://doi.org/10.1145/2396761.2398726 .

KENTER, T., DE RIJKE, M., 2015, "Short Text Similarity with Word Embeddings". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, pp. 1411–1420, New York, New York, USA. ACM Press. ISBN: 9781450337946. https://doi.org/10.1145/2806416.2806475 .

KHURDIYA, A., DEY, L., MAHAJAN, D., et al., 2012, "Extraction and Compilation of Events and Sub-events from Twitter". In: *Web Intelligence and*

Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on, v. 1, pp. 504–508. https://doi.org/10.1109/WI-IAT.2012.192 .

KIM, H., REN, X., SUN, Y., et al., 2013, "Semantic Frame-Based Document Representation for Comparable Corpora". In: *IEEE 13th International Conference on Data Mining*, pp. 350–359. IEEE, dec. ISBN: 978-0-7695-5108-1. https://doi.org/10.1109/ICDM.2013.99 .

KOKAR, M. M., MATHEUS, C. J., BACLAWSKI, K., 2009, "Ontology-based situation awareness", *Information Fusion*, v. 10, n. 1, Special Issue on High-level Information Fusion and Situation Awareness (jan.), pp. 83–98. ISSN: 15662535. https://doi.org/10.1016/j.inffus.2007.01.004 .

KRAFT, T., WANG, D. X., DELAWDER, J., et al., 2013, "Less After-the-Fact: Investigative visual analysis of events from streaming twitter". In: *IEEE Symposium on Large Data Analysis and Visualization 2013, LDAV 2013 - Proceedings*, pp. 95–103, Atlanta, GA. IEEE Computer Society. https://doi.org/10.1109/LDAV.2013.6675163 .

KUMAR, S., MORSTATTER, F., MARSHALL, G., et al., 2012, "Navigating Information Facets on Twitter (NIF-T)". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'12, pp. 1548–1551, New York, NY, USA. ACM. ISBN: 978-1-4503-1462-6. https://doi.org/10.1145/2339530.2339777 .

LAPPAS, T., VIEIRA, M. R., GUNOPULOS, D., et al., 2012, "On the Spatiotemporal Burstiness of Terms", *Proceedings of the VLDB Endowment*, v. 5, pp. 836–847. ISSN: 2150-8097. https://doi.org/10.14778/2311906.2311911 .

LE, P., ZUIDEMA, W., 2014, "Inside-Outside Semantics: A Framework for Neural Models of Semantic Composition". In: *Deep Learning and Representation Learning Workshop - NIPS*, pp. 1–11, Montreal, Quebec, Canada. http://www.dlworkshop.org/30.pdf .

LE, Q. V., MIKOLOV, T., 2014, "Distributed Representations of Sentences and Documents", *Proceedings of the 31st International Conference on Machine Learning*, v. 32, n. 2 (jun), pp. 1188–1196. http://proceedings.mlr.press/v32/le14.html .

LEBAN, G., FORTUNA, B., BRANK, J., et al., 2014, "Event Registry: Learning About World Events from News". In: *Proceedings of the 23rd International*

Conference on World Wide Web - WWW '14 Companion, pp. 107–110, Seoul, Korea. ACM Press. ISBN: 9781450327459. https://doi.org/10.1145/2567948.2577024 .

LECUN, Y., BENGIO, Y., HINTON, G., 2015, "Deep learning", *Nature*, v. 521, n. 7553 (may), pp. 436–444. ISSN: 0028-0836. doi: 10.1038/nature14539. http://www.nature.com/articles/nature14539 .

LEE, C.-H., 2012, "Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams", *Expert Systems with Applications*, v. 39, n. 18 (dec), pp. 13338–13356. ISSN: 09574174. https://doi.org/10.1016/j.eswa.2012.05.068 .

LEE, C. H., YANG, H. C., CHIEN, T. F., et al., 2011, "A novel approach for event detection by mining spatio-temporal information on microblogs". In: *Proceedings - 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011*, ASONAM'11, pp. 254–259, Kaohsiung, jul. IEEE. ISBN: 9780769543758. https://doi.org/10.1109/ASONAM.2011.74 .

LEIDNER, J. L., 2004, "Toponym Resolution in Text: "Which Sheffield is It?"". In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 602, Sheffield, United Kingdom. ACM. ISBN: 1-58113-881-4. https://doi.org/10.1145/1008992.1009147 .

LEVY, O., GOLDBERG, Y., DAGAN, I., 2015, "Improving Distributional Similarity with Lessons Learned from Word Embeddings", *Transactions of the Association for Computational Linguistics*, v. 3, pp. 211–225. ISSN: 2307-387X. https://transacl.org/ojs/index.php/tacl/article/view/570 .

LI, J., LUONG, T., JURAFSKY, D., 2015, "A Hierarchical Neural Autoencoder for Paragraphs and Documents". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1106–1115, Stroudsburg, PA, USA. Association for Computational Linguistics. ISBN: 9781941643723. https://doi.org/10.3115/v1/P15-1107 .

LIEBERMAN, M. D., SAMET, H., SANKARANARAYANAN, J., 2010, "Geotagging with local lexicons to build indexes for textually-specified spatial

data". In: *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pp. 201–212, Long Beach, CA, USA, may. IEEE. ISBN: 978-1-4244-5445-7. https://doi.org/10.1109/ICDE.2010.5447903 .

LIN, F., 1996, "Embracing Causality in Specifying the Indeterminate Effects of Actions". In: *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1*, pp. 670–676, Portland, Oregon. AAAI Press. http://www.aaai.org/Papers/AAAI/1996/AAAI96-100.pdf .

LIN, F., 2008, "Situation Calculus". In: van Harmelen, F., Lifschitz, V., Porter, B. (Eds.), *Handbook of Knowledge Representation*, v. 3, *Foundations of Artificial Intelligence*, Elsevier, cap. 16, pp. 649–669, Amsterdam. ISBN: 9780444522115. https://doi.org/10.1016/S1574-6526(07)03016-7 .

LIU, M., WANG, L., NIE, L., et al., 2016a, "Event graph based contradiction recognition from big data collection", *Neurocomputing*, v. 181, n. Big Data Driven Intelligent Transportation Systems (mar), pp. 64–75. ISSN: 09252312. https://doi.org/10.1016/j.neucom.2015.06.099 .

LIU, W., LUO, X., GONG, Z., et al., 2016b, "Discovering the core semantics of event from social media", *Future Generation Computer Systems*, v. 64, n. C (nov), pp. 175–185. ISSN: 0167739X. https://doi.org/10.1016/j.future.2015.11.023 .

LONG, R., WANG, H., CHEN, Y., et al., 2011, "Towards Effective Event Detection, Tracking and Summarization on Microblog Data". In: Wang, H., Li, S., Oyama, S., et al. (Eds.), *International Conference on Web-Age Information Management - WAIM 2011*, v. 6897, *Lecture Notes in Computer Science*, pp. 652–663, Wuhan, China. Springer Berlin Heidelberg. ISBN: 9783642235344. https://doi.org/10.1007/978-3-642-23535-1_55 .

LU, Y., WANG, H., LANDIS, S., et al., 2017, "A Visual Analytics Framework for Identifying Topic Drivers in Media Events", *IEEE Transactions on Visualization and Computer Graphics*, v. PP, n. 99, pp. 1–1. ISSN: 1077-2626. https://doi.org/10.1109/TVCG.2017.2752166 .

LUO, X., XU, Z., YU, J., et al., 2011, "Building Association Link Network for Semantic Link on Web Resources", *IEEE Transactions on Automation Science and Engineering*, v. 8, n. 3 (jul), pp. 482–494. ISSN: 1545-5955. https://doi.org/10.1109/TASE.2010.2094608 .

MAGDY, A., ALY, A., MOKBEL, M., et al., 2014a, "Mars: Real-time spatio-temporal queries on microblogs". In: *Proceedings - International Con-*

ference on Data Engineering, pp. 1238–1241, Chicago, IL, a. IEEE Computer Society. ISBN: 9781479925544. https://doi.org/10.1109/ICDE.2014.6816750 .

MAGDY, A., MOKBEL, M., ELNIKETY, S., et al., 2014b, "Mercury: A memory-constrained spatio-temporal real-time search on microblogs". In: *Proceedings - International Conference on Data Engineering*, pp. 172–183, Chicago, IL, b. IEEE Computer Society. ISBN: 9781479925544. https://doi.org/10.1109/ICDE.2014.6816649 .

MAHATA, D., TALBURT, J. R., SINGH, V. K., 2015, "From Chirps to Whistles". In: *Proceedings of the ACM Web Science Conference on ZZZ - WebSci '15*, WebSci '15, pp. 1–10, New York, New York, USA. ACM Press. ISBN: 9781450336727. https://doi.org/10.1145/2786451.2786476 .

MATHEUS, C. J., KOKAR, M. M., BACLAWSKI, K., 2003, "A core ontology for situation awareness". In: *Information Fusion, 2003. Proceedings of the Sixth International Conference of*, v. 1, pp. 545–552, Cairns, Queensland, Australia, jul. IEEE. ISBN: 0972184449. https://doi.org/10.1109/ICIF.2003.177494 .

MATHEUS, C. J., BACLAWSKI, K., KOKAR, M. M., et al., 2005a, "Using SWRL and OWL to Capture Domain Knowledge for a Situation Awareness Application Applied to a Supply Logistics Scenario". In: Adi, A., Stoutenburg, S., Tabet, S. (Eds.), *Rules and Rule Markup Languages for the Semantic Web, First International Conference, RuleML 2005*, v. 3791, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Berlin Heidelberg, pp. 130–144, Galway, Ireland, nov.a. ISBN: 978-3-540-32270-2. https://doi.org/10.1007/11580072_11 .

MATHEUS, C. J., KOKAR, M. M., BACLAWSKI, K., et al., 2005b, "SAWA: an assistant for higher-level fusion and situation awareness". In: Dasarathy, B. V. (Ed.), *Proceedings of SPIE*, p. 75, Bellingham, WA, marb. SPIED Digital Library. https://doi.org/10.1117/12.604120 .

MATHEUS, C. J., KOKAR, M. M., BACLAWSKI, K., et al., 2005c, "An Application of Semantic Web Technologies to Situation Awareness". In: Gil, Y., Motta, E., Benjamins, V. R., et al. (Eds.), *International Semantic Web Conference – ISWC 2005*, v. 3729, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 944–958, Galway, Ireland, sepc. ISBN: 978-3-540-29754-3. https://doi.org/10.1007/11574620_67 .

MAZIERO, E. G., JORGE, M. L. D. R. C., PARDO, T. A. S., 2014, "Revisiting Cross-document Structure Theory for multi-document discourse parsing", *Information Processing & Management*, v. 50, n. 2 (mar), pp. 297–314. ISSN: 03064573. https://doi.org/10.1016/j.ipm.2013.12.003 .

MCCARTHY, J., HAYES, P. J., 1969, "Some Philosophical Problems from the Standpoint of Artificial Intelligence", *Machine Intelligence*, v. 4. http://www-formal.stanford.edu/jmc/mcchay69.pdf .

MIDDLETON, S. E., MIDDLETON, L., MODAFFERI, S., 2014, "Real-Time Crisis Mapping of Natural Disasters Using Social Media", *Intelligent Systems, IEEE*, v. 29, n. 2 (mar.), pp. 9–17. ISSN: 1541-1672. https://doi.org/10.1109/MIS.2013.126 .

MIKOLOV, T., CORRADO, G., CHEN, K., et al., 2013a, "Efficient Estimation of Word Representations in Vector Space", *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12. ISSN: 15324435. https://arxiv.org/abs/1301.3781 .

MIKOLOV, T., YIH, S. W.-T., ZWEIG, G., 2013b, "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*, pp. 746–751, Atlanta, Georgia, junb. Association for Computational Linguistics. http://www.aclweb.org/anthology/N13-1090 .

MILLER, G. A., 1995, "WordNet: a lexical database for English", *Communications of the ACM*, v. 38, n. 11 (nov), pp. 39–41. ISSN: 00010782. https://doi.org/10.1145/219717.219748 .

MIRANDA ACKERMAN, E. J., 2012, "Extracting a Causal Network of News Topics". In: Herrero, P., Panetto, H., Meersman, R., et al. (Eds.), *On the Move to Meaningful Internet Systems: OTM 2012 Workshops*, v. 7567 LNCS, Springer Berlin Heidelberg, pp. 33–42, Rome, Italy, sep. ISBN: 9783642336171. https://doi.org/10.1007/978-3-642-33618-8_5 .

MIRZA, P., 2014, "Extracting Temporal and Causal Relations between Events". In: *Proceedings of the ACL 2014 Student Research Workshop*, pp. 10–17, Baltimore, Maryland, USA. Association for Computational Linguistics. http://acl2014.org/acl2014/P14-3/pdf/P14-3002.pdf .

MIRZA, P., TONELLI, S., 2014, "An Analysis of Causality between Events and its Relation to Temporal Information". In: *Proceedings of COLING 2014, the*

25th International Conference on Computational Linguistics: Technical Papers, pp. 2097–2106, Dublin, Ireland. Association for Computational Linguistics. http://www.aclweb.org/anthology/C14-1198 .

MIRZA, P., SPRUGNOLI, R., TONELLI, S., et al., 2014, "Annotating Causality in the TempEval-3 Corpus". In: *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pp. 10–19, Gothenburg, Sweden. Association for Computational Linguistics. http://www.aclweb.org/anthology/W14-0702 .

MUELLER, E. T., 2008, "Event Calculus". In: van Harmelen, F., Lifschitz, V., Porter, B. (Eds.), *Handbook of Knowledge Representation*, v. 3, *Foundations of Artificial Intelligence*, Elsevier, cap. 17, pp. 671–708, Amsterdam. ISBN: 9780444522115. https://doi.org/10.1016/S1574-6526(07)03017-9 .

NAVARRO-COLORADO, B., SAQUETE, E., 2016, "Cross-document event ordering through temporal, lexical and distributional knowledge", *Knowledge-Based Systems*, v. 110 (oct), pp. 244–254. ISSN: 09507051. https://doi.org/10.1016/j.knosys.2016.07.032 .

NEVATIA, R., HOBBS, J., BOLLES, B., 2004, "An Ontology for Video Event Representation". In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)*, p. 119. IEEE Computer Society, jun. ISBN: 0-7695-2158-4. https://doi.org/10.1109/CVPR.2004.27 .

NG, A. Y., JORDAN, M. I., WEISS, Y., 2001, "On Spectral Clustering: Analysis and an Algorithm". In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pp. 849–856, Cambridge, MA, USA. MIT Press. http://dl.acm.org/citation.cfm?id=2980539.2980649 .

NURWIDYANTORO, A., WINARKO, E., 2013, "Event detection in social media: A survey". In: *ICT for Smart Society (ICISS), 2013 International Conference on*, pp. 1–5. https://doi.org/10.1109/ICTSS.2013.6588106 .

2010, *Common Alerting Protocol Version 1.2 - OASIS Standard*. OASIS Emergency Management TC, July. http://docs.oasis-open.org/emergency/cap/v1.2/CAP-v1.2.pdf .

OZDIKIS, O., OGUZTUZUN, H., KARAGOZ, P., 2013, "Evidential Location Estimation for Events Detected in Twitter". In: *Proceedings of the 7th Workshop on Geographic Information Retrieval, GIR 2013*, GIR'13, pp. 9–16, Orlando, FL. ACM. ISBN: 978-1-4503-2241-6. https://doi.org/10.1145/2533888.2533929 .

PALMER, M., GILDEA, D., KINGSBURY, P., 2005, "The Proposition Bank: An Annotated Corpus of Semantic Roles", *Computational Linguistics*, v. 31, n. 1 (mar), pp. 71–106. ISSN: 0891-2017. https://doi.org/10.1162/0891201053630264 .

PAULUS, R., XIONG, C., SOCHER, R., 2018, "A Deep Reinforced Model for Abstractive Summarization". In: *Sixth International Conference on Learning Representations – ICLR*, Vancouver, Canada, apr. https://openreview.net/forum?id=HkAClQgA- .

PENNINGTON, J., SOCHER, R., MANNING, C. D., 2014, "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. http://www.aclweb.org/anthology/D14-1162 .

PHONG LÊ, 2016, *Learning Vector Representations for Sentences – The Recursive Deep Learning Approach.* Amsterdam, ILLC - Institute for Logic, Language and Computation. ISBN: 9789402801811. https://www.illc.uva.nl/Research/Publications/Dissertations/DS-2016-05.text.pdf .

PISKORSKI, J., YANGARBER, R., 2012, "Information Extraction: Past, Present and Future". In: Poibeau, T., Saggion, H., Piskorski, J., et al. (Eds.), *Theory and Applications of Natural Language Processing*, Theory and Applications of Natural Language Processing, Springer Berlin Heidelberg, cap. 2, pp. 23–49, Berlin, Heidelberg, jul. ISBN: 978-3-642-28568-4. https://doi.org/10.1007/978-3-642-28569-1_2 .

PUSTEJOVSKY, J., CASTAÑO, J. M., INGRIA, R., et al., 2003, "TimeML: Robust Specification of Event and Temporal Expressions in Text". In: *New Directions in Question Answering, 2003 AAAI Spring Symposium*, v. 3, pp. 28–34, Stanford, CA, USA. AAAI Press. ISBN: 1577351843. https://www.aaai.org/Papers/Symposia/Spring/2003/SS-03-07/SS03-07-005.pdf .

RADINSKY, K., HORVITZ, E., 2013, "Mining the Web to Predict Future Events". In: *Proceedings of the sixth ACM international conference on Web search*

and data mining - WSDM'13, pp. 255–264, Rome, Italy. ACM. ISBN: 9781450318693. https://doi.org/10.1145/2433396.2433431 .

RAIMOND, Y., ABDALLAH, S., SANDLER, M., et al., 2007, "The Music Ontology", *8th International Conference on Music Information Retrieval - ISMIR 2007*, v. 8, pp. 417–422. http://raimond.me.uk/pubs/Raimond-ISMIR2007-Submitted.pdf .

ŘEHŮŘEK, R., SOJKA, P., 2010, "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, pp. 46–50, Valletta, Malta. University of Malta. ISBN: 2-9517408-6-7. https://is.muni.cz/publication/884893/en .

ROHDE, D. L. T., GONNERMAN, L. M., PLAUT, D. C., 2004, "An Improved Method for Deriving Word Meaning from Lexical Co-Occurrence", *Cognitive Psychology*, v. 7, pp. 573–605. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.3741 .

ROSPOCHER, M., VAN ERP, M., VOSSEN, P., et al., 2016, "Building event-centric knowledge graphs from news", *Web Semantics: Science, Services and Agents on the World Wide Web*, v. 37-38 (mar), pp. 132–151. ISSN: 15708268. https://doi.org/10.1016/j.websem.2015.12.004 .

ROZENSHTEIN, P., ANAGNOSTOPOULOS, A., GIONIS, A., et al., 2014, "Event Detection in Activity Networks". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'14, pp. 1176–1185, New York, NY, USA. ACM. ISBN: 978-1-4503-2956-9. https://doi.org/10.1145/2623330.2623674 .

RUMELHART, D. E., HINTON, G. E., WILLIAMS, R. J., 1986, "Learning representations by back-propagating errors", *Nature*, v. 323, n. 6088 (oct), pp. 533–536. ISSN: 0028-0836. https://doi.org/10.1038/323533a0 .

RUPPENHOFER, J., ELLSWORTH, M., PETRUCK, M. R. L., et al., 2016, *FrameNet II: Extended theory and practice*. Berkeley. https://framenet.icsi.berkeley.edu/fndrupal/the{_}book .

SAHLGREN, M., 2005, "An Introduction to Random Indexing". In: *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*, pp. 1–9, Copenhagen, Denmark, aug. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.96.2230 .

SAKAKI, T., MATSUO, Y., YANAGIHARA, T., et al., 2012, "Real-time event extraction for driving information from social sensors". In: *Proceedings - 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, CYBER 2012*, pp. 221–226, Bangkok. IEEE Computer Society. ISBN: 9781467314213. https://doi.org/10.1109/CYBER.2012.6392557 .

SAKAKI, T., OKAZAKI, M., MATSUO, Y., 2010, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors". In: *Proceedings of the 19th international conference on World wide web - WWW'10*, WWW'10, pp. 851–860, New York, NY, USA. ACM. ISBN: 978-1-60558-799-8. https://doi.org/10.1145/1772690.1772777 .

SALAKHUTDINOV, R., HINTON, G., 2007, "Semantic Hashing". In: *SIGIR workshop on Information Retrieval and applications of Graphical Models*, Amsterdam, jul. ACM Press. http://www.cs.cmu.edu/{~}rsalakhu/papers/semantic{_}final.pdf .

SALAKHUTDINOV, R., HINTON, G., 2009, "Semantic hashing", *International Journal of Approximate Reasoning*, v. 50, n. 7 (jul), pp. 969–978. ISSN: 0888613X. https://doi.org/10.1016/j.ijar.2008.11.006 .

SANTORO, A., RAPOSO, D., BARRETT, D. G. T., et al., 2017, "A simple neural network module for relational reasoning". In: Guyon, I., Luxburg, U. V., Bengio, S., et al. (Eds.), *Advances in Neural Information Processing Systems 30, NIPS*, pp. 4974–4983, Long Beach, CA, dec. Curran Associates, Inc. https://papers.nips.cc/paper/7082-a-simple-neural-network-module-for-relational-reasoning .

SAURÍ, R., LITTMAN, J., KNIPPEN, B., et al., 2006. "TimeML Annotation Guidelines, Version 1.2.1". http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf .

SAYYADI, H., HURST, M., MAYKOV, A., 2009, "Event Detection and Tracking in Social Streams". In: *Third International AAAI Conference on Weblogs and Social Media*. http://aaai.org/ocs/index.php/ICWSM/09/paper/view/170 .

SCHERP, A., MEZARIS, V., 2014, "Survey on modeling and indexing events in multimedia", *Multimedia Tools and Applications*, v. 70, n. 1, pp. 7–23. ISSN: 15737721. https://doi.org/10.1007/s11042-013-1427-7 .

SCHERP, A., AGARAM, S., JAIN, R., 2008, "Event-centric media management". In: *SPIE, Multimedia Content Access: Algorithms and Systems II, Video Analysis and Retrieval I*, v. 6820, pp. 68200C–68200C–15, San Jose, CA. SPIED Digital Library. https://doi.org/10.1117/12.761974 .

SCHERP, A., FRANZ, T., SAATHOFF, C., et al., 2009, "F — A Model of Events based on the Foundational Ontology DOLCE + DnS Ultralite", *International Conference on Knowledge Capturing (K-CAP)*, pp. 137–144. https://doi.org/10.1145/1597735.1597760 .

SCHERP, A., FRANZ, T., SAATHOFF, C., et al., 2012, "A core ontology on events for representing occurrences in the real world", *Multimedia Tools and Applications*, v. 58, n. 2, pp. 293–331. ISSN: 15737721. https://doi.org/10.1007/s11042-010-0667-z .

SCHÖLKOPF, B., SMOLA, A. J., MÜLLER, K.-R., 1999, "Kernel principal component analysis". In: Schölkopf, B., Burges, C. J. C., Smola, A. J. (Eds.), *Advances in Kernel Methods*, MIT Press, cap. 20, pp. 327–352, Cambridge, MA, USA. ISBN: 9780262283199. http://cognet.mit.edu/book/advances-kernel-methods .

SCHULZ, A., HADJAKOS, A., PAULHEIM, H., et al., 2013, "A Multi-Indicator Approach for Geolocalization of Tweets". In: *Seventh International AAAI Conference on Weblogs and Social Media*, pp. 573–582, Boston, Massachusetts USA. AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6063 .

SCHUSTER, D., ROSI, A., MAMEI, M., et al., 2013, "Pervasive social context: Taxonomy and survey", *ACM Transactions on Intelligent Systems and Technology*, v. 4, n. 3 (jun.), pp. 1. ISSN: 21576904. https://doi.org/10.1145/2483669.2483679 .

SHAFER, G., 1976, *A Mathematical Theory of Evidence*. Princeton University Press. ISBN: 0691081751 069110042. http://nla.gov.au/nla.cat-vn1156754 .

SHARMA, S., KUMAR, R., BHADANA, P., et al., 2013, "News Event Extraction Using 5W1H Approach & Its Analysis", *International Journal of Scientific & Engineering Research - IJSER*, v. 4, n. 5, pp. 2064–2067.

SHAW, R., TRONCY, R., HARDMAN, L., 2009, "LODE: Linking Open Descriptions of Events". In: *The Semantic Web - Fourth Asian Conference, ASWC 2009*, v. 5926, *Lecture Notes in Computer Science*, pp. 153–167,

Shanghai, China, dec. Springer Berlin Heidelberg. ISBN: 978-3-642-10871-6. https://doi.org/10.1007/978-3-642-10871-6_11 .

SHI, L.-L., LIU, L., WU, Y., et al., 2017, "Event Detection and User Interest Discovering in Social Media Data Streams", *IEEE Access*, v. PP, n. 99, pp. 1–1. ISSN: 2169-3536. https://doi.org/10.1109/ACCESS.2017.2675839 .

SINCLAIR, P., ADDIS, M., CHOI, F., et al., 2006, "The use of CRM Core in Multimedia Annotation". In: *First International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*, p. 14, Edinburgh, Scotland, jul. http://eprints.soton.ac.uk/262828/ .

SINGH, R., LI, Z., KIM, P., et al., 2004, "Event-based Modeling and Processing of Digital Media". In: *Proceedings of the 1st international workshop on Computer vision meets databases - CVDB '04*, pp. 19–26, Paris, France. ACM. ISBN: 1581139179. https://doi.org/10.1145/1039470.1039478 .

SKINNER, J., 2011, "Social Media and Revolution: The Arab Spring and the Occupy Movement as Seen through Three Information Studies Paradigms", *Sprouts: Working Papers on Information Systems*, v. 11, pp. 169. http://sprouts.aisnet.org/11-169 .

SOCHER, R., MANNING, C. D. C., NG, A. Y. A., 2010, "Learning Continuous Phrase Representations and Syntactic Parsing with Recursive Neural Networks". In: *Proceedings of the Deep Learning and Unsupervised Feature Learning Workshop of NIPS 2010*, pp. 1–9. http://ai.stanford.edu/{~}ang/papers/nipsdlufl10-LearningContinuousPhraseRepresentations.pdf .

SOCHER, R., HUANG, E. H., PENNIN, J., et al., 2011a, "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection". In: *Advances in Neural Information Processing Systems (NIPS)*, v. 24, Curran Associates, Inc., pp. 801–809, Granada, Spain, a. https://papers.nips.cc/paper/4204-dynamic-pooling-and-unfolding-recursive-autoencoders-for-paraphrase-detection .

SOCHER, R., PENNINGTON, J., HUANG, E. H., et al., 2011b, "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 151–161, Edinburgh, Scotland, UK, julb. Association for Computational Linguistics. http://www.aclweb.org/anthology/D11-1014 .

SOCHER, R., BAUER, J., MANNING, C. D., et al., 2013a, "Parsing with Compositional Vector Grammars". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013*, pp. 455–465, Sofia, Bulgaria, auga. http://www.aclweb.org/anthology/P13-1045 .

SOCHER, R., PERELYGIN, A., WU, J. Y., et al., 2013b, "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pp. 1631–1642, Seattle, Washington, USA, octb. http://www.aclweb.org/anthology/D13-1170 .

STARBIRD, K., PALEN, L., HUGHES, A. L., et al., 2010, "Chatter on the Red: What Hazards Threat Reveals About the Social Life of Microblogged Information". In: *CSCW'10 Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 241–250, Savannah, Georgia, USA. ACM. ISBN: 9781605587950. https://doi.org/10.1145/1718918.1718965 .

STEIGER, E., DE ALBUQUERQUE, J. P., ZIPF, A., 2015, "An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data", *Transactions in GIS*, v. 19, n. 6 (dec), pp. 809–834. ISSN: 13611682. https://doi.org/10.1111/tgis.12132 .

STEPANOVA, E., 2011, *The Role of Information Communication Technologies in the "Arab Spring" - Implications beyond the region.* Relatório Técnico 159, The George Washington University Elliott School of International Affai, Washington, US. http://www.gwu.edu/~ieresgwu/assets/docs/ponars/pepm_159.pdf .

SUTSKEVER, I., VINYALS, O., LE, Q. V., 2014, "Sequence to Sequence Learning with Neural Networks". In: Ghahramani, Z., Welling, M., Cortes, C., et al. (Eds.), *Advances in Neural Information Processing Systems (NIPS)*, v. 27, Curran Associates, Inc., pp. 337–42, Montreal, Quebec, Canada, sep. ISBN: 1409.3215. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural .

SZUCS, I., GOMBOS, G., KISS, A., 2013, "Five Ws, one H and many tweets". In: *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pp. 441–446. https://doi.org/10.1109/CogInfoCom.2013.6719287 .

TAI, K. S., SOCHER, R., MANNING, C. D., 2015, "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks". In:

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, v. 1: Long Pa, pp. 1556–1566, Beijing, China, jul. http://aclweb.org/anthology/P15-1150 .

TERRANA, D., PILATO, G., 2013, "Detection, Clustering and Tracking of Life Cycle Events on Twitter Using Electric Fields Analogy". In: *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pp. 220–227. https://doi.org/10.1109/ICSC.2013.46 .

TIPPING, M. E., BISHOP, C. M., 1999, "Mixtures of Probabilistic Principal Component Analyzers", *Neural Computation*, v. 11, n. 2 (feb), pp. 443–482. ISSN: 0899-7667. https://doi.org/10.1162/089976699300016728 .

TONKIN, E. L., PFEIFFER, H. D., 2013, "Zombies Walk Among Us: Cross-Platform Data Mining for Event Monitoring". In: *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pp. 452–459. https://doi.org/10.1109/ICDMW.2013.52 .

TUMASJAN, A., SPRENGER, T., SANDNER, P., et al., 2010, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". In: *Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178–185, Washington, DC, USA. AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441 .

TZELEPIS, C., MA, Z., MEZARIS, V., et al., 2016, "Event-based media processing and analysis: A survey of the literature", *Image and Vision Computing*, v. 53, pp. 3–19. ISSN: 02628856. https://doi.org/10.1016/j.imavis.2016.05.005 .

UMAMAHESWARI, E., GEETHA, T., 2015, "Learning event patterns from news text using bootstrapping", *International Journal of Information and Communication Technology*, v. 7, n. 1, pp. 1. ISSN: 1466-6642. https://doi.org/10.1504/IJICT.2015.065992 .

VAN DER MAATEN, L., POSTMA, E., VAN DEN HERIK, J., 2009, *Dimensionality Reduction : A Comparative Review*. Relatório técnico, Tilburg centre for Creative Computing, Tilburg, Netherlands, oct. https://www.tilburguniversity.edu/upload/59afb3b8-21a5-4c78-8eb3-6510597382db{_}TR2009005.pdf .

VAN HAGE, W. R., MALAISÉ, V., DE VRIES, G. K. D., et al., 2012, "Abstracting and reasoning over ship trajectories and web data with the Simple Event Model (SEM)", *Multimedia Tools and Applications*, v. 57, n. 1, pp. 175–197. ISSN: 13807501. https://doi.org/10.1007/s11042-010-0680-2 .

VON LUXBURG, U., 2007, "A tutorial on spectral clustering", *Statistics and Computing*, v. 17, n. 4 (dec), pp. 395–416. ISSN: 0960-3174. https://doi.org/10.1007/s11222-007-9033-z .

WALTHER, M., KAISSER, M., 2013, "Geo-spatial Event Detection in the Twitter Stream". In: Serdyukov, P., Braslavski, P., Kuznetsov, S. O., et al. (Eds.), *Advances in Information Retrieval*, Springer Berlin Heidelberg, pp. 356–367, Moscow, Russia. ISBN: 978-3-642-36973-5. https://doi.org/10.1007/978-3-642-36973-5_30 .

WANG, W., 2012, "Chinese News Event 5W1H Semantic Elements Extraction for Event Ontology Population". In: *Proceedings of the 21st International Conference Companion on World Wide Web - WWW'12 Companion*, p. 197, Lyon, France. ACM. ISBN: 9781450312301. https://doi.org/10.1145/2187980.2188008 .

WANG, W., ZHAO, D., 2012, "Ontology-Based Event Modeling for Semantic Understanding of Chinese News Story". In: Zhou, M., Zhou, G., Zhao, D., et al. (Eds.), *Natural Language Processing and Chinese Computing - NLPCC, Proceedings in First CCF Conference*, v. 333, *Communications in Computer and Information Science*, Springer Berlin Heidelberg, pp. 58–68, Beijing, China, oct. ISBN: 9783642344558. https://doi.org/10.1007/978-3-642-34456-5_6 .

WANG, W., ZHAO, D., ZOU, L., et al., 2010, "Extracting 5W1H Event Semantic Elements from Chinese Online News". In: Chen, L., Tang, C., Yang, J., et al. (Eds.), *Web-Age Information Management, Proceedings in 11th International Conference, WAIM 2010*, v. 6184, *Lecture Notes in Computer Science*, Springer-Verlag, pp. 644–655, Jiuzhaigou, China, jul. ISBN: 3642142451. https://doi.org/10.1007/978-3-642-14246-8_62 .

WANG, X., GERBER, M. S., BROWN, D. E., 2012a, "Automatic crime prediction using events extracted from twitter posts", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 7227 LNCS, pp. 231–238. ISSN: 03029743. https://doi.org/10.1007/978-3-642-29047-3_28 .

WANG, X. H., ZHANG, D. Q., GU, T., et al., 2004, "Ontology based context modeling and reasoning using OWL". In: *IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second*, pp. 18–22, Orlando, FL, mar. IEEE. ISBN: 0-7695-2106-1. https://doi.org/10.1109/PERCOMW.2004.1276898 .

WANG, X. J., MAMADGI, S., THEKDI, A., et al., 2007, "Eventory - An event based media repository". In: *ICSC 2007 International Conference on Semantic Computing*, pp. 95–102, Irvine, CA. IEEE Computer Society. ISBN: 0769529976. https://doi.org/10.1109/ICSC.2007.70 .

WANG, X., FU, H., XU, C., et al., 2014, "Provenance Logic: Enabling Multi-Event Based Trust in Mobile Sensing". In: *2014 IEEE 33rd International Performance Computing and Communications Conference (IPCCC)*, pp. 1–8, Austin, TX. IEEE. ISBN: 978-1-4799-7575-4. https://doi.org/10.1109/PCCC.2014.7017107 .

WANG, Z., CUI, P., XIE, L., et al., 2012b, "Analyzing Social Media via Event Facets". In: *Proceedings of the 20th ACM International Conference on Multimedia - MM '12*, pp. 1359–1360, Nara, Japan, b. ACM. ISBN: 9781450310895. https://doi.org/10.1145/2393347.2396484 .

WATANABE, K., OCHI, M., OKABE, M., et al., 2011, "Jasmine: A Real-time Local-event Detection System Based on Geolocation Information Propagated to Microblogs". In: *International Conference on Information and Knowledge Management, Proceedings*, CIKM'11, pp. 2541–2544, New York, NY, USA. ACM. ISBN: 978-1-4503-0717-8. https://doi.org/10.1145/2063576.2064014 .

WEI, Y., SINGH, L., GALLAGHER, B., et al., 2016, "Overlapping Target Event and Story Line Detection of Online Newspaper Articles". In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 222–232. IEEE, oct. ISBN: 978-1-5090-5206-6. https://doi.org/10.1109/DSAA.2016.30 .

WEILER, A., GROSSNIKLAUS, M., SCHOLL, M. H., 2015a, "Run-Time and Task-Based Performance of Event Detection Techniques for Twitter". In: Zdravkovic, Jelena and Kirikova, Marite and Johannesson, P. (Ed.), *Advanced Information Systems Engineering, Proceedings of CAiSE 2015, 27th International Conference on*, v. 9097, *Lecture Notes in Computer Science*, Springer International Publishing, pp. 35–49, Stockholm, Swe-

den, jun.a. ISBN: 978-3-319-19068-6. https://doi.org/10.1007/978-3-319-19069-3_3 .

WEILER, A., GROSSNIKLAUS, M., SCHOLL, M. H., 2015b, "Evaluation Measures for Event Detection Techniques on Twitter Data Streams". In: Maneth, S. (Ed.), *Data Science - Proceedings of 30th British International Conference on Databases, BICOD 2015*, v. 9147, *Lecture Notes in Computer Science*, Springer International Publishing, pp. 108–119, Edinburgh, Scotland, jul.b. ISBN: 978-3-319-20424-6. https://doi.org/10.1007/978-3-319-20424-6_11 .

WEILER, A., GROSSNIKLAUS, M., SCHOLL, M. H., 2016, "An evaluation of the run-time and task-based performance of event detection techniques for Twitter", *Information Systems*, (dec), pp. 207–219. ISSN: 03064379. https://doi.org/10.1016/j.is.2016.01.003 .

WENG, J., LEE, B.-S., 2011, "Event Detection in Twitter". In: *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 401–408, Barcelona, Spain. AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2767 .

WESTERMANN, U., JAIN, R., 2006, "{\rm E} - A Generic Event Model for Event-Centric Multimedia Data Management in eChronicle Applications". In: *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pp. x106–x106, Atlanta, GA, USA. IEEE. ISBN: 0-7695-2571-7. https://doi.org/10.1109/ICDEW.2006.1 .

WESTERMANN, U., JAIN, R., 2007, "Toward a common event model for multimedia applications", *IEEE Multimedia*, v. 14, n. 1, pp. 19–29. ISSN: 1070986X. https://doi.org/10.1109/MMUL.2007.23 .

WOLFE, T., 2010, *The New Journalism*. Reino Unido, Pan MacMillan. ISBN: 1581130155.

WU, Y., SUN, H., YAN, C., 2017, "An event timeline extraction method based on news corpus". In: *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pp. 697–702, Beijing, China, mar. IEEE. ISBN: 978-1-5090-3618-9. https://doi.org/10.1109/ICBDA.2017.8078725 .

XIE, L., SUNDARAM, H., CAMPBELL, M., 2008, "Event mining in multimedia streams", *Proceedings of the IEEE*, v. 96, n. 4, pp. 623–647. ISSN: 00189219. https://doi.org/10.1109/JPROC.2008.916362 .

XU, W., SUN, H., DENG, C., et al., 2017, "Variational Autoencoder for Semi-Supervised Text Classification". In: *AAAI Conference on Artificial Intelligence*. https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14299https://arxiv.org/abs/1603.02514 .

XU, Z., WEI, X., LUO, X., et al., 2015, "Knowle: A semantic link network based system for organizing large scale online news events", *Future Generation Computer Systems*, v. 43-44 (feb), pp. 40–50. ISSN: 0167739X. https://doi.org/10.1016/j.future.2014.04.002 .

YANG, X., ZHANG, T., XU, C., 2015, "Cross-Domain Feature Learning in Multimedia", *IEEE Transactions on Multimedia*, v. 17, n. 1 (jan), pp. 64–78. ISSN: 1520-9210. https://doi.org/10.1109/TMM.2014.2375793 .

YANG, Y., PIERCE, T., CARBONELL, J., 1998, "A Study of Retrospective and On-line Event Detection". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '98*, pp. 28–36, Melbourne, Australia. ACM Press. ISBN: 1581130155. https://doi.org/10.1145/290941.290953 .

YANG, Z., HU, Z., SALAKHUTDINOV, R., et al., 2017, "Improved Variational Autoencoders for Text Modeling using Dilated Convolutions". In: Precup, D., Teh, Y. W. (Eds.), *Proceedings of the 34th International Conference on Machine Learning*, v. 70, *Proceedings of Machine Learning Research*, pp. 3881–3890, Sydney, Australi, aug. PMLR. http://proceedings.mlr.press/v70/yang17d.htmlhttp://arxiv.org/abs/1702.08139 .

YAU, S. S., LIU, J., 2006, "Hierarchical Situation Modeling and Reasoning for Pervasive Computing". In: *Proceedings of the Fourth IEEE Workshop on Software Technologies for Future Embedded and Ubiquitous Systems and Second International Workshop on Collaborative Computing, Integration, and Assurance (SEUS-WCCIA'06)*, pp. 5–10, Gyeongju, Korea. IEEE Computer Society. ISBN: 0769525601. https://doi.org/10.1109/SEUS-WCCIA.2006.25 .

YE, J., DASIOPOULOU, S., STEVENSON, G., et al., 2015, "Semantic web technologies in pervasive computing: A survey and research roadmap", *Pervasive and Mobile Computing*, pp. 543–549. ISSN: 15741192. https://doi.org/10.1016/j.pmcj.2014.12.009 .

YUAN, Q., CONG, G., MA, Z., et al., 2013, "Who, Where, When and What: Discover Spatio-temporal Topics for Twitter Users". In: *Proceedings of the*

19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'13, pp. 605–613, New York, NY, USA. ACM. ISBN: 978-1-4503-2174-7. https://doi.org/10.1145/2487575.2487576 .

ZARRINKALAM, F., BAGHERI, E., 2016, "Event Identification in Social Networks", *Encyclopedia with Semantic Computing*, v. 1, n. 1 (jun), pp. 1–8. ISSN: 0033-5177. http://arxiv.org/abs/1606.08521 .

ZHAO, J., WANG, X., MA, Z., 2014, "Towards Events Detection from Microblog Messages", *International Journal of Hybrid Information Technology*, v. 7, n. 1, pp. 201–210. ISSN: 17389968. https://doi.org/10.14257/ijhit.2014.7.1.16 .

ZHAO, L., SUN, Q., YE, J., et al., 2015, "Multi-Task Learning for Spatio-Temporal Event Forecasting". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'15*, Sydney, NSW, Australia, aug. ACM. ISBN: 9781450336642. https://doi.org/10.1145/2783258.2783377 .

ZHAO, M., ZHANG, C., LU, S., et al., 2016a, "STeller: An approach for context-aware story detection using different similarity metrics and dense subgraph mining". In: *2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 152–157. IEEE, maya. ISBN: 978-1-5090-1915-1. https://doi.org/10.1109/CSCWD.2016.7565980 .

ZHAO, S., LIU, T., ZHAO, S., et al., 2016b, "Event causality extraction based on connectives analysis", *Neurocomputing*, v. 173, n. P3 (jan), pp. 1943–1950. ISSN: 09252312. https://doi.org/10.1016/j.neucom.2015.09.066 .

ZHENG, L., JIN, P., ZHAO, J., et al., 2014, "A Fine-Grained Approach for Extracting Events on Microblogs". In: Decker, H., Lhotská, L., Link, S., et al. (Eds.), *25th International Conference Database and Expert Systems Applications, DEXA 2014*, v. 8644, pp. 275–283, Munich, Germany. Springer International Publishing. ISBN: 978-3-319-10073-9. https://doi.org/10.1007/978-3-319-10073-9_22 .

ZHENG, S., XU, J., ZHOU, P., et al., 2016, "A neural network framework for relation extraction: Learning entity semantic and relation pattern", *Knowledge-Based Systems*, v. 114 (dec), pp. 12–23. ISSN: 09507051. https://doi.org/10.1016/j.knosys.2016.09.019 .

ZHENG, S., HAO, Y., LU, D., et al., 2017, "Joint entity and relation extraction based on a hybrid neural network", *Neurocomputing*, v. 257, n. 2016 (sep), pp. 59–66. ISSN: 09252312. https://doi.org/10.1016/j.neucom.2016.12.075 .

ZHONG, G., WANG, L.-N., LING, X., et al., 2016, "An overview on data representation learning: From traditional feature learning to recent deep learning", *The Journal of Finance and Data Science*, v. 2, n. 4 (dec), pp. 265–278. ISSN: 2405-9188. https://doi.org/10.1016/j.jfds.2017.05.001 .

ZHU, X., OATES, T., 2013, "Finding News Story Chains Based on Multidimensional Event Profiles". In: *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pp. 157–164, Paris, France, France. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. ISBN: 978-2-905450-09-8. http://dl.acm.org/citation.cfm?id=2491748.2491782 .