Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Abstract

Context encompasses the classification of a certain environment by its key attributes that take the role of semantic markers. It is an abstract representation of a certain data environment. In texts, the context classifies and represents a piece of text in a generalized form. Context can be a recursive construct when summarizing information on a more coarse-grained level. This paper presents identification and standardization of context on different levels of granularity that finally supports faster and more precise information retrieval. The prototypical system presented here applies supervised learning for a semi-automatic approach to extract, distil, and standardize data from text. The approach is based on named-entity recognition and simple ontologies for identification and disambiguation of context. Even though the prototype shown here still represents work in progress, it already demonstrates its potential for mining texts on different levels of context granularity. The paper presents the design of the Contexter system that supports identification and classification of misinformation and fake news around the topic Covid-19.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the KES International.

Keywords: Information Extraction; Named-Entity Recognition; Bag of Words; Information Summarization; Entity Linking; Semantic Markers; Context-Awareness; Economic Information

1. Introduction

Misinformation has many facets, ranging from lies over fake news, half-truth, implausible predictions, or pseudoscientific statements to unconsciously wrong assessments and badly founded statements. Misinformation is producing mistrust among our societies. Trustworthy persons and sources do not appear trustworthy and competent anymore. Mistrust splits societies, hampers reciprocal respect and esteem, and prevents us from creating a socially

* Corresponding author. Tel.: +491788586619.
E-mail address: k.englmeier@hs-sm.de
based concerted defense against Covid-19. Stockpiling toilet paper, flour, yeast, and pasta was emblematic for this social lack of trust. All promises of officials that there is no shortage of all these things went unheard. Even journalists contributed to this climate of mistrust.

The distinguished Berlin newspaper Tagesspiegel, for instance, predicted on March 20, 2020 (see figure 1): “If the rate of infection continues to grow at its present rate, by tomorrow we will have crossed the mark to 1000 infected, and next week we’ll reach around 10,000. Just in Berlin.” [1] This prediction was far from being based on facts. By end of May, Berlin counted around 6,300 infections with a growth rate of 30 to 40 infections per day. The proposition of the authors was rather speculation. Covid-19 provoked and still provokes uncertainty, fear, and anxiety in our societies. Many people had and have problems of what and whom to believe. When people yearn for reliable information, journalists have to be extremely careful when doing their jobs. Their personal opinion and stance may not be part of their reporting. The carelessness of newspapers also adds its piece to the lack of trust many people have in times of Corona.

For the detection of fake news, context information highlights characteristic qualities of a text and represents them in a standardized form. Context information can be considered as the product of iterative summarization of statements and standardization of summary terms. The hierarchy of terms constitute semantic anchors of the text on different levels of granularity, on phrase or paragraph level or addressing the text in its entirety. The standardized representation of information includes also its context representation. In texts, the context classifies and represents a piece of text in a generalized form. Context can be a recursive construct when summarizing text on a more coarse-grained level. The prototype applies supervised learning for a semi-automatic approach to extract, distill, and standardize data from text. The approach is based on named-entity recognition and simple ontologies for identification and disambiguation of content. Even though the prototype shown here still represents work in progress, it demonstrates its potential in the detection of fake news and misinformation.

2. Approach Related Work

Fake news come in different forms [2, 3], mainly as humorous news, hoaxes (sensational news on events that never happened in the described way, just in order to attract readers), or severe fabrications (modified or twisted news to change the meaning of a proposition to a completely different intention). The objective of fake news detection is to classify the latter two as falsified representations of facts. There are different directions in fake news detection in texts:

1. Checking questionable statements against facts retrieved from trustworthy sources. However, there is a caveat: Trustworthy sources may also publish erroneous information.
2. Comparison of information on particular facts from different sources may reveal an outstanding and potentially wrong representation of the same fact addressed in different statements.
3. If the statement in question is part of a discourse (like on social media) and sparks responses with many skeptical ones among them, we can assume that the statement may contain misinformation or fake news. Negations to the original statement also contribute to the pattern of skepticism pointing to potential fake news.

4. Many fake news, even published on different channels, originate from the same author, probably a notorious producer of fake news or a troll. The similar writing style may reveal the authorship of that particular person [4, 5, 6].

5. Fake news is often inclined to conspiracy theories. Their advocates quite often use an offensive, aggressive, and even threatening language in their statements. Writing style and vocabulary are similar to those used in hate speech and, thus, an outstanding characteristic. In the context of propositions related to Covid-19, they point to possible instances of conspiracy theories and/or fake news.

There is huge variety of technical approaches applied in order to detect fake news [7, 8, 9, 10, 11].

The design of Contexter for fake news and misinformation detection bases on a combination of Named Entity Recognition (NER), Bag of Words (BoW), and Word N-Grams [12, 13]. The idea behind the design is that each fact has its individual pattern of words and numeric data. We can assume that a specific combination of keywords and annotated numeric expressions uniquely reflects a particular fact.

The identification of facts starts with information extraction [14] and the annotation of the extracted pieces of text according to their inherent meaning [15]. Annotation has two roles: first, it adds a meaningful term to the extracted text, in particular to numeric data. Such patterns, for example, represent dates, percentages, growth rates, distances, and the like. Second, the annotations (and keywords) from the first annotation process are further annotated yielding an increasingly more abstract representation of the fact(s). Semantic markers [16] are the smallest fraction of a text covering a certain meaning discernable from the other fractions. Together they mark the meaning of a particular piece of text. In other words, iterative annotation constructs semantic markers of increasing complexity.

Fake news detection may also include methods and techniques of event mining [17] when it comes to analyze statements over a discourse (in social media) with events that triggered this discourse. [18] developed a framework for narratives of a therapist-patient discourse that is valuable in our context. His work has been summarized and discussed in [19].

3. Recognition and Representation of Facts

The Contexter system enables the definition blueprints or patterns of facts (including those used in fake news and misinformation). These blueprints are abstract representations of the things described in text. They mainly consist of Named Entities combined with keywords from Bag of Words (BoW) and arranged in close proximity (much like Word N-Grams). On a higher level of the iterative annotation, the N-Grams of annotation terms may span large portions of the text.

The users describe their blueprints by arranging terms from the BoWs and Named Entities. Contexter takes each set of terms and translates it into Regular Expressions. The users produce an initial set of blueprints reflecting facts that are subsequently applied to the text collection in order to find all corresponding text instances. In a next step, Contexter gradually modifies each blueprint, by gradually modifying its elements. This automatic process aims at revealing as many representations of facts as possible. Contexter presents all encountered new instances of them to the users for confirmation and rejection, that, in turn, further directs the learning process.

Representations of facts in text have unique patterns, independently of the source or language. Figure 2, for example, shows two messages covering the same fact from two different news channels in Germany and Spain. Both are very trustworthy sources of information in their respective country.

Among other things both news report a decrease in the number of new fatalities (123) and newly registered cases of Covid-19 infections (373) over the last 24 hours on May 10, 2020. The German news does not qualify for fake news, but for misinformation. It states that number of infections increased by 3046 during the last 24 hours, which is obviously wrong. The rest of the facts that both news address is identical and correct. This example helps to demonstrate how detection of misinformation works using an approach for text analysis as outlined in the next sections. By distilling essential parts from text and annotating and structuring the extracted terms we increase automatic processing and, thus, faster classification of text.
3.1. Generic Named Entities and Bag of Words

Contexter provides a number of theme-specific BoWs (for locations, names, expressions of aggression etc.) and
Named Entities for common patterns in text reflecting time, prices, distances, and the like. This process usually
combines key words and common text and numeric patterns. Finally, each pattern is annotated by one or more
appropriate terms that summarizes the meaning of the pattern. Contexter usually arranges these terms in more
comprehensive hierarchies of terms.

Generic named entities help to standardize factual information and to abstract away the different forms of
expressions for essentially the same thing. However, it does not suffice just to annotate generic patterns. We
represent the numerical values in a standard way, too (see figure 3). Here, we can easily imagine that Named
Entities may relate to ontologies that serve specific interpretation or calculation purposes.

Named Entity Recognition (NER) in the context described here operates with bags of words (BoW) addressing
locations, persons, organizations, or institutions (Wall Street, Dow Jones, White House, Bangladesh, for instance).
Furthermore, we use key words (such as “Mr.” or “Health Senator” like in the text of figure 1) that hint to names of
persons. The system takes these names and feeds them into the respective bag of words. There are further interesting
key terms pointing to names. For example, the term “by” following the title of an article leads the list of names

† Both texts have been retrieved on May 11, 2020, on https://www.tagesschau.de/newsticker/liveblog-coronavirus-
montag-103.html and https://elpais.com/sociedad/2020-05-11/los-nuevos-diagnosticos-y-fallecidos-caen-al-minimo-
desde-el-pico-de-la-pandemia.html. The German broadcast service tagesschau later corrected its error.
authoring that article. The identification of proper names benefits from the analysis of sequential dependencies when bags of words can be produced automatically instead of manually. There are promising approaches to automatically identify names (and other important key expressions) in texts using conditional random fields (CFR) [20] or hidden Markov Models (HMMs) [21]. Inclined to CFR we integrated a feature that proposes, for example, all names starting with capital letters and followed by an abbreviation as organization names, such as “National Institute of Health (NIH)” or “Health and Human Services (HHS)”.

We can easily imagine domain specific BoWs for business, energy, cooking, travel, and the like. These BoWs help to thematically discriminate pieces of text. The most important BoWs, however, address locations, persons, organizations, and the like. In most cases, these collections of names are produced manually or extracted semi-manually from registers. Nevertheless, we can harness conditional relationships between specific terms and proper names. There are always typical patterns in texts that undoubtedly point to names of persons. Analyzing conditional relationships leads to the identification of quite a number of proper names that otherwise need to be handled manually. Contexter feeds identified proper names back to the respective BoWs. These domain-specific collections of terms are indispensable for the correct identification of content.

The meaning of a piece of text is more than just the meaning of one or more specific terms surrounded by further terms. Annotation helps to identify the overarching theme manifested by these terms. In only a few cases, the meaning of a piece of text can be inferred just by the individual meaning of its terms.

3.2. Representation of Content in Distilled and Structured Form

In many situations, the overarching meaning is not expressed by even a single word in the phrase. For the correct interpretation of text, we need representations on a higher level, that is, on a more abstract level. This does not mean that we abstract away details. On the contrary, we add essential details.

```
<fatality>
  <duration>
    en un día
  </duration>
  <increase>
    <delta>123</delta>
  </increase>
</fatality>

<infection>
  <increase>
    <delta>373</delta>
  </increase>
</infection>
```

Fig. 3.: Structure of Named Entities indicating the overarching themes “fatality” and “infection”.

The next level of abstraction is achieved again by operating on the named entities of the previous phases. Named entities on this level may indicate an increase or decrease in infections, casualties, prices, cases, or the like. It may also reflect a current situation in a particular town, country, or region.

Figure 3 shows fragments of the distilled and annotated information from the texts shown in figure 2. The Named Entities reflect the overarching theme of the pieces of text addressing the of infections and fatalities. Distilled and structured data are easier to handle by mining processes in order to detect discrepancies in text, in particular if they are manifested in numerical data (figures 373 vs. 3046 in the example above).

Named Entity Recognition (NER) presented so far serves the standardization of data, in terms of assigning standardized annotations to the underlying data. It also helps to detect term patterns and keywords in text fragments that constitute conditional dependencies among the terms. Furthermore, it reveals the correct and unambiguous meaning of text fragments and their components. By the iterative and incremental application of NER we can produce theme-specific hierarchies of named entities that represent the meaning of content on different levels of granularity. The hierarchies are the building blocks of content representation of text. We may call them content clusters of Named Entities or content schemas.

BoWs play also an important role in the definition of these schemas. A keyword embedded in a pattern of named entities is important when it comes to correctly identify the overarching theme of the piece of text. In the example of figure 3, we have a Covid-19 scene of a particular day in Spain. The meaning of each of these elements and the act in the scene can only be correctly identified by the Named Entities in conjunction with keywords. Terms like “asciende” (increases) or “neuen” (new) together with their adjacent terms are an essential ingredient for correctly classifying the phenomenon “increase”. This example shows that keywords and Named Entities take the role of semantic markers for a particular content. Semantic markers and their relationships among them constitute content schemas.

The identification of semantic markers is a process covering a series of phases:

1. At first, we independently identify clusters of Named Entities and keywords. BoWs provide the keywords while Basic Named Entities contribute standardized versions of small and generic fractions of text. Basic Named Entities may even be aggregated to enhance the content representation (like the Named Entity “increase” and “infection” in figure 3). Incrementally, we obtain more complex representations (much like in ontologies) that include semantic relationships among the elements. Even if automatic NER can identify a wide range of elementary entities, this process starts with humans defining the patterns (or blueprints) for these entities and controlling the results of the automatic identification.

2. The next step is the manual definition of more complex content clusters like the ones shown in figure 3. This is usually an iterative process of defining and testing prototypes for these clusters. If prototypes match on more than one occasion over the whole text collection, in our case Covid-19 reports, we consider them as candidates for further analysis.

3. Phase 2 is the starting point for the learning process to identify term hierarchies. The Contexter system takes the confirmed prototypes of blueprints, and continuously applies them to the text collection. Whenever the system identifies a similar pattern of semantic markers it tries to find further patterns that match this pattern. Each instance found is marked as potential candidate for a new content cluster. If we take the example in figure 3 as seed, we may find similar text pattern comprising the same Named Entities but differ in keywords or Basic Named Entities applied.

4. The results of the automatic NER are controlled and evaluated. In this situation, we need again human intervention. They manage theme-specific BoWs and content schemas and classify the output of the system. A candidate proposed by the system can be confirmed or rejected. Rejected content schemas serve as negative example, in order to avoid that the system proposes a wrong schema again.

NER on a higher level of abstraction is similar to identifying meaningful N-Grams in text. The only difference here is that the sequence of Named Entities is not completely fix, much like in natural language when phrases with the same meaning differ in their word order. The more solid content schemas Contexter can find the better performance it gets.
4. Conclusion

This paper presented the state of work of Contexter, a prototypical system that operates on Named Entity Recognition and uses theme-specific Bag of Words to identify semantic markers in text that point to the specific meaning of texts or text fragments.

The application areas of the content schemas are manifold. The main purpose is identifying facts in texts and represent them in a distilled and standardized way along their respective context. This facilitates the comparison of facts in different sources and, thus, supports the detection of fake news and misinformation.

Named entities and terms from BoWs identify the meaning of words as they appear in a phrase or fragment of text. However, they explicitly include numerical data that are very important for correct reflection of meaning in text, in particular those that address Covid-19. Iteratively applying standardization to already extracted and annotated pieces of text adds semantic hierarchies that enable the identification of meaning along generalization or specification aspects. This, in turn, makes text comparisons more precise and versatile.

We can also use content schemas to classify small text fragments, like microposts in social media for instance. By analyzing news or prominent statements on social media channels and microposts we can identify the contextual anchors [22] of propositions along a discourse line. This, in turn, helps to classify microposts (as hate speech or conspiracy theory, for instance) and to identify the root of the discourse.

Contexter is still work in progress, but we already noticed that our content schemas have a certain proximity to ontologies. We use the schemas for text interpretation on a basic level and gradually produce concept hierarchies. However, we clearly see the necessity to add more functionality to schemas, in particular, when parts of the schema address factual (i.e. numerical) information. Quite often calculations can be helpful to check the plausibility of statements based on numerical information. Let us take the example text of figure 1 which predicts a ten-fold increase of infections in Berlin in one week. By comparing the increase rate obtained from trustworthy sources with the predicted one we may notice an extraordinary difference leading to the conclusion that the predicted rate in the text is not plausible.

A further objective of Contexter is a stronger involvement of humans in the development and management of text mining tools, in general, to enhance the adoption of this technology on a broader scale. This involvement results in a more active role of the users in designing, controlling, and adapting of the learning process that feeds, in this case here, the automatic detection of misinformation and fake news.

References

[1] Jacobs, S., Wächter, K., Coates, L. (2020) “First Major German City Under Curfew, will Berlin follow soon?”. Der Tagesspiegel, March 20, 2020. Retrieved at https://www.tagesspiegel.de/berlin/coronavirus-outbreak-first-major-german-city-under-curfew-will-berlin-follow-soon/25664744.html on May 11, 2020.
[2] Rubin, V. L., Chen, Y., Conroy, N. J., (2015) “Deception Detection for News: Three Types of Fakes”. Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community: 1—4.
[3] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihailea, R. (2019) “Automatic Detection of Fake News”. Proceedings of the 27th International Conference on Computational Linguistics, 27: 3391–3401.
[4] Juola, P. (2012) “Detecting stylistic deception”. Proceedings of the Workshop on Computational Approaches to Deception Detection: 91—96.
[5] Levitan, S.I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., Rosenberg, A. (2015) “Cross-Cultural Production and Detection of Deception from Speech”. Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection: 1—8
[6] Gupta, M., Han, J. (2011) “Heterogeneous Network-Based Trust Analysis: A Survey”. ACM SIGKDD Explorations Newsletter 13 (1): 54–71
[7] Zhou, X., Zafarani, R. (2019) “Fake News Detection: An Interdisciplinary Research”. Companion Proceedings of the 2019 World Wide Web Conference: 1292.
[8] Zafarani, R., Zhou, X., Shu, K., Liu, H. (2019) “Fake News Research: Theories, Detection Strategies, and Open Problems”. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining: 3207—3208.
[9] Cha, M., Gao, W., Li, C.-T. (2020) “Detecting Fake News in Social Media: An Asia-Pacific Perspective”. Communications of the ACM 63 (4): 68—71.
[10] Katsaros, D., Stavropoulos, G., Papakostas, D. (2019) “Which machine learning paradigm for fake news detection?”. IEEE/WIC/ACM International Conference on Web Intelligence: 383—387.
[11] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. (2017) “Fake News Detection on Social Media: A Data Mining Perspective”. ACM SIGKDD Explorations Newsletter 19 (1): 22—36

[12] Wynne, H.E., Wint, Z.Z. (2019) “Content Based Fake News Detection Using N-Gram Models”. Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS2019): 669—673.

[13] Woods, W. A. (1970) “Context-Sensitive Parsing”. Communications of the ACM 13(7): 413—445.

[14] Cowie, J., Lehner, W. (1996) “Information Extraction”. Communications of the ACM 39(1): 80—91.

[15] Salton, G., Allan, J., Buckely, Ch., Singhal, A. (1997) “Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts”, in Karen Sparck Jones and Peter Willett, Readings in Information Retrieval, San Francisco: 478—483.

[16] Jancsary, J., Neubarth, F., Schreitter, S., Trost, H. (2011) “Towards a context-sensitive online newspaper”. Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation: 2—9.

[17] Calvo Martinez, J. (2018) “Event Mining over Distributed Text Streams”. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining: 745—746.

[18] Schneider, P. (2013) “Language usage and social action in the psychoanalytic encounter: discourse analysis of a therapy session fragment”. Language and Psychoanalysis 2 (1): 4—19.

[19] Murtagh, F. (2014) “Mathematical Representations of Matte Blanco’s Bi-Logic, based on Metric Space and Ultrametric or Hierarchical Topology: Towards Practical Application.” Language and Psychoanalysis 3(2): 40—63.

[20] Sha, F., Pereira, F. (2003) “Shallow Parsing with Conditional Random Fields”. Proceedings of the HLT-NAACL conference: 134-141.

[21] Freitag, D., McCallum, A. (2000) “Information Extraction with HMM Structures Learned by Stochastic Optimization”. Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence: 584—589.

[22] Dang, V., Croft, B.W. (2010) “Query Reformulation Using Anchor Text”. Proceedings of the Third ACM International Conference on Web Search and Data Mining: 41—50.