Capturing themed evidence, a hybrid approach

Enrico Daga
enrico.daga@open.ac.uk
The Open University
Milton Keynes, United Kingdom

Enrico Motta
enrico.motta@open.ac.uk
The Open University
Milton Keynes, United Kingdom

ABSTRACT

The task of identifying *pieces of evidence* in texts is of fundamental importance in supporting qualitative studies in various domains, especially in the humanities. In this paper, we coin the expression *themed evidence*, to refer to (direct or indirect) traces of a *fact or situation* relevant to a *theme of interest* and study the problem of identifying them in texts. We devise a generic framework aimed at capturing themed evidence in texts based on a *hybrid* approach, combining statistical natural language processing, background knowledge, and Semantic Web technologies. The effectiveness of the method is demonstrated on a case study of a digital humanities database aimed at collecting and curating a repository of evidence of experiences of listening to music. Extensive experiments demonstrate that our hybrid approach outperforms alternative solutions. We also evidence its generality by testing it on a different use case in the digital humanities.

CCS CONCEPTS

• **Information systems** → **Information extraction**; *Presentation of retrieval results*; • **Computing methodologies** → **Information extraction**; • **Applied computing** → *Arts and humanities.*

KEYWORDS

themed evidence, information extraction, hybrid method, DBpedia

ACM Reference Format:
Enrico Daga and Enrico Motta. 2019. Capturing themed evidence, a hybrid approach. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Exploring digital resources in search of pieces of evidence relevant to a certain research theme is a difficult and important task for humanities research. The concept of *evidence* is a particularly difficult one, as it relates to a fact being reported in a text, which is relevant to a certain subject of enquiry. However, traversing a book in search of relevant text is a difficult and time consuming task. In this paper, we coin the expression *themed evidence*, to refer to (direct or indirect) traces of a *fact or situation* relevant to a *theme of interest* and study the problem of identifying them in texts.

The task of identifying themed evidence is at the intersection between topical text classification (finding texts relevant to a certain theme) and event extraction (find events mentioned in texts). However, not all topical texts are themed evidence and the event itself is often not reported in the text. As an example, let's consider the following texts: (1) *"The Protestants are but few in number, and their singing is congregational."* (2) *"The best choir-singing, (Roman Catholic) without accompaniment, we have heard, was at Munich."* (3) *"Holland is the country of bells; and the merry chimes are to be heard hourly, from almost every church-tower or steeple."* (4) *"When in Berlin, we had the pleasure of an interview with Professor Dehn, one of the most learned musicians in Germany."* All four sentences are topical of the concept *music* but only the first three are *pieces of evidence* of a listening situation. The second text refers to a specific event while the third reports a recurring one. The first example does reports a judgement that indirectly refers to a typical situation, not a particular event. Ultimately, it is to some extent a subjective decision of the researcher whether a piece of text is relevant or not, in particular when the collection of evidence is part of an exploratory process [21]. However, there is enough consensus among the majority of the cases to make an automatic solution both feasible in principle and very useful in practice.

Specifically, our contribution includes:

- A novel approach that combines corpus-based natural language processing and semantic web technologies to retrieve *themed evidence* in texts.
- An exemplary application of our method to the problem of retrieving listening experiences.
- Two novel gold standards to evaluate competing methods for detecting themed evidence as *listening* [8] and *reading* [9] experiences.
- Extensive experiments supporting the validity of the approach, including evaluation across two domains, aimed at demonstrating the generality of the method.

In the next section we describe a general method for capturing *themed evidence* (Section 2). Section 3 describes the application of the approach to the case of listening experiences. Section 4 reports on our evaluation, including the preparation of the gold standards and the design of baseline methods for the task. Section 5 discusses the results in more details, in particular analysing the relative strength and weaknesses of the various computational components. Related work is illustrated in Section 6. In Section 7 we summarize the main contributions of this work and discuss future developments of this line of research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 A HYBRID APPROACH

We approach the problem of identifying a themed evidence as a binary classification task [32], where a segment of text is evaluated by a classifier that assigns a positive or negative label. In what follows, we provide an outline of the method and its underlying assumptions. We will then instantiate it in our scenario and discuss it in detail in Section 3. Our strategy comprises three major steps: (a) *statistical relatedness analysis*, aimed at evaluating the relevance of the text with respect to the theme making use of a dictionary and background corpora; (b) *detection of relevant entities*, focused on identifying in the text named entities from a Knowledge Graph (KG); and (c) *hybridisation technique*, dedicated to performing an integration of the two.

Statistical relatedness analysis. This activity is performed by relying on a reference corpus for learning a dictionary of words related to the theme, associated with a relevance score. This dictionary is then used to annotate the text so that each word is mapped to a similarity score. Particularly, we use a state-of-the-art algorithm[1] to produce *word embeddings*, a vectorial representation of words that can be interrogated to measure, for example, similarity between words. Pragmatically, the model can be queried with a word to obtain an ordered (scored) list of similar terms. To apply this technique, we select a word as *core concept*, for example "music". Applying the dictionary, we can compute a *relatedness measure* for a text by summing the score of their words, normalised by the length of the text. Formally, let's consider a similarity function

$$Sim : W \times D \rightarrow S \cup \{\epsilon\} \tag{1}$$

where W is the domain of possible words, D is the dictionary, S is the similarity score to the *core concept*, and $\epsilon = 0.00$ is the score for any word not in D, and a text characterized as a sequence of words:

$$T = \{w_1, w_2, \ldots, w_n\} \tag{2}$$

A statistical relatedness is measured as the sum of the scores of terms in the text, normalised by the text length:

$$R = \sum_{w \in T} Sim(w, D) \Big/ |T| \tag{3}$$

A text strongly related to the *core concept* will have a significant overlap of terms with the dictionary, therefore a higher score. However, in order to use this measure to inform a decision, we need to establish a threshold value. To learn the threshold value we can analyse the distribution of the dictionary on a corpus of positive samples. In our case, we calculate both *average* score $\overline{x}$ and *standard deviation* $\sigma_{\overline{x}}$. These values partition the corpus in four segments, from the lower- to the higher-scoring positive samples: (1) $r < (\overline{x} - \sigma_{\overline{x}})$; (2) $\overline{x} < r > (\overline{x} - \sigma_{\overline{x}})$; (3) $(\overline{x} + \sigma_{\overline{x}}) < r > (\overline{x})$; (4) $r > (\overline{x} + \sigma_{\overline{x}})$. From these, we can derive three possible thresholds: $th_1 > (\overline{x} - \sigma_{\overline{x}})$, $th_2 > \overline{x}$, and $th_3 > \overline{x} + \sigma_{\overline{x}}$. A database of negative samples should be used to test the safe threshold to apply. This approach will give a *clue* about the *statistical relatedness* of a text to the *core concept* (theme).

However, this method alone is insufficient at capturing the complexity of a *themed evidence*. Particularly, named entities may not be sufficiently represented in the dictionary, especially when resulting from compound terms such as names of persons or place names (e.g. "the battle of Hastings"), or creative works (e.g. "Prélude à l'après-midi d'un faune"). However, the presence of named entities is a strong account of relatedness to the *themed evidence*.

Detecting entities related to the theme. In this step we make the reasonable assumption that the presence of entities related to the theme would be a suggestion that the text is an account of a listening event. We use a general purpose Knowledge Graph (KG) and, to identify entities, a state-of-art named entity recognition (NER) engine. To map retrieved entities with the theme, we select a *core entity* in the KG as reference node. One way of measuring relatedness between two nodes in a graph is by evaluating their distance. In our work, we use DBpedia Spotlight to identify entities and the taxonomy of DBpedia categories for filtering them. To do that, we apply the following SPARQL query:

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dct: <http://purl.org/dc/terms/>
SELECT  distinct ?sub WHERE { VALUES ?sub { %entities% }
  ?sub dct:subject ?subject .
  ?subject skos:broader{0:%d%} %theme% }
```

where *%entities%* are the entities identified by the NER engine, *%theme%* is the *core entity*, and *%d%* is a threshold distance[2]. Clearly, different entities (DBpedia categories) will have a different distribution of related resources. To learn the value of *%d%* one easy way is to sample related entities at incremental distances and stop when the results starts to include irrelevant entities. This process is manual but can be executed once for configuring the component at design phase. The presence of a relevant entity constitutes an additional *clue* about a potential *evidence*.

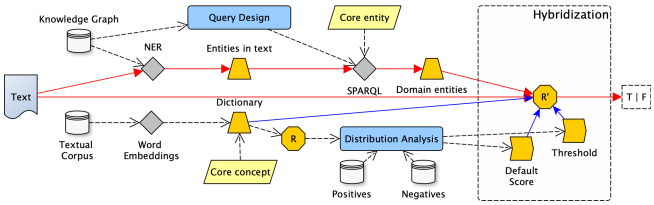

Figure 1: Illustration of the approach. The bottom layer illustrates the statistical analysis. The top layer illustrates the process of detecting entities related to a core entity. In a hybridisation phase, demoting and promoting heuristics are applied.

Hybridization and* noise *correction. Each one of the above *clues* are still insufficient at capturing the complexity and variety of a *themed evidence*. Particularly, we make the following observations:

- The dictionary may inherit noise derived from polysemy, for example, or figurative and metaphorical expressions.
- Related entities may not appear in the linguistic background knowledge or may map to low score terms.

To reduce the noise and benefit from the semantic background knowledge, we perform a hybridisation step in which we *demote* and *promote* words, as follows: (a) we change the scoring formula

[1]The Word2Vec library of Apache Spark, inspired by [27]. This approach has gained a lot of traction in recent years for its efficiency in computing the vector space. See also [13] for an explanation of the method.

[2]The approach could be reproduced with other knowledge graphs. However, this step of the method requires to design a solution for filtering out entities that are not related to the theme.

to only consider terms that occurr as *verbs* or *nouns*; (b) we boost the score of terms mapped to named entities relevant to the theme. Particularly, when a noun or a verb is also mapped to an entity, we *double-up* its similarity score. In case the original score was 0.0 (the mapped term did not appear in the dictionary, for example, an entity such as *Prélude à l'après-midi d'un faune*[3]), a default score is assigned, using the average score learnt during the statistical relatedness analysis. Formally, our model is adapted as follows. First, let's consider the function

$$Ent : W \times G \rightarrow B \tag{4}$$

where $B = \{0, 1\}$, testing whether a word maps to a relevant named entity in the KG G. The similarity function is adapted as follows:

$$\begin{aligned} Sim' : W \times D \times G \rightarrow ((\neg Ent(W, G) \rightarrow Sim(W, D)) \\ \cup (\neg Sim(W, D) \rightarrow \overline{x}) \cup (2 \cdot Sim(W, D))) \end{aligned} \tag{5}$$

The funciton applies the dictionary score when a word is not mapped to an entity. In case it is and the score was zero, a default score is applied using the average score value ($\overline{x}$). Otherwise, if the score was non-zero, the score is doubled (*entity boost*), as previously explained. The resulting relevance function is therefore adapted as follows:

$$R' = \sum_{w \in T, w \in N \cup V} Sim'(w, D, G) \Big/ |T| \tag{6}$$

Where N and V are the sets of all nouns and verbs, respectively. This new function accounts for relevant entities in the text and aims at reducing the negative impact of polysemy and figurative speech.

The overall approach is illustrated in Figure 1. In the following section we describe its application to a real case study.

3 LISTENING EXPERIENCES: A CASE STUDY

The Listening Experience Database Project (LED)[4] is an initiative funded by the UK's Arts and Humanities Research Council (AHRC) aimed at collecting accounts of people's private experiences of listening to music [5]. A listening experience is an exemplary case of a *themed evidence*: *an account of an event involving music and one or more participants*. Since 2012, the LED project has collected over 10,000 unique listening experiences from a variety of textual sources [3]. These are books, for example, published by the Internet Archive[5] or Google Books[6] and explored using either the search facility of the web portal or an application such as a PDF reader. From the curator's perspective, the process starts from a source and moves to selecting an initial set of keywords. For example, terms such as *music**, *sing**, *song*, are consistently used as keywords, and then, elaborating from the retrieved material, expanded with more specific terms, in an iterative and exploratory process. However, the manual approach is clearly limited and prone to errors.

Our method relies significantly on background knowledge. Before illustrating how we applied this method to the detection of listening experiences, we introduce a set of corpora and motivate their adoption.

[3]Wikipedia page: https://en.wikipedia.org/wiki/Prélude_à_l'après-midi_d'un_faune.
[4]The LED Project: http://www.listeningexperience.org/
[5]Internet Archive: https://archive.org/, accessed 22 October 2018.
[6]Google Books: https://books.google.co.uk/, accessed 22 October 2018.

3.1 Resources as background knowledge

The **Project Gutenberg** offers more than fifty-eight thousand books in the public domain [15]. The vast majority of the books are in English, 48790 books, including approximately 1.5 billions of words. We chose Project Gutenberg as representative of the type of sources used by target users, since it was one of the main sources used by the LED curators during the project. We use the repository for generating word embeddings for the statistical relatedness analysis.

The **Listening Experience Database (LED)** includes text excerpts that we used as positive samples and compared with negative samples from other sources.

The **Reuters-21578 (Reu)** corpus is a standard data set adopted extensively for training and evaluating systems for information retrieval, document classification, machine learning and similar corpus-based research [22]. The corpus includes 21.578 news articles of various categories generally related to economics. Crucially, it does not include music within the assigned categories[7].

The **UK Reading Experience Database (RED)** is a research project focused on the development of an open access database [8]. The aim of the project is to investigate the evidence of reading in Britain [10]. This project is interesting as alternative source of themed evidence, providing accounts of experiences of *reading* instead of *listening*.

DBpedia. DBpedia is a large knowledge graph published as Linked Data [4]. Originally generated from Wikipedia info boxes, the project aggregates data from multiple sources and offers a SPARQL endpoint and the NER tool DBpedia Spotlight [25].

3.2 Finding listening experiences

In what follows we apply our method to the case of discovering listening experiences. We refer to the following texts as guide examples (taken from the gold standard [8]):

RECMUS-619, positive: *Introduced to the Anacreontic Society, consisting of amateurs who perform admirably the best orchestral works. The usual supper followed. After propitiating me with a trio from 'Cosi Fan Tutte', they drew me to the piano.*

MASONB-31, positive: *In the evening we went to Rev. Baptist Noel's chapel, where one is always sure of edification from the sermon if not from the psalms.*

MASONB-88, negative: *Flags and pendants were suspended from the windows, [...] the colors of the German States were waving harmoniously together, and the banners of the Fine Arts, with appropriate inscriptions, particularly those of music, poetry and painting, were especially honored, and floated triumphant amidst the standards of electorates, dukedoms, and kingdoms.*

Initially, we selected as core concept the noun `music[n]` and as core entity the DBpedia category http://dbpedia.org/resource/Category:Music. Next, we learnt a dictionary of words related to *music* (*core concept*), associated with a relevance score (see Formula 1), to perform the statistical relatedness analysis. This activity

[7]Reuters-21578 was downloaded from: http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html, accessed 2 April 2019.
[8]Reading Experience Database, 1450–1945: http://www.open.ac.uk/Arts/RED/index.html, accessed 2 April 2019.

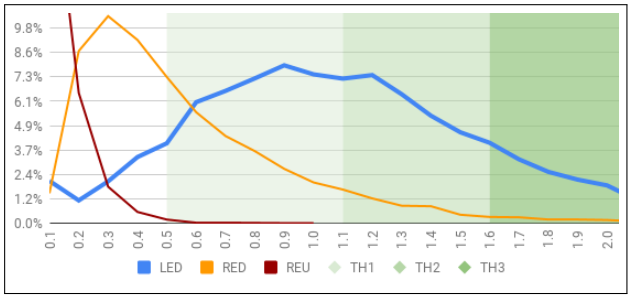

Figure 2: Distribution of the dictionary in the corpora. The horizontal axis represents discrete score values and the vertical the percentage of texts. The three possible thresholds are shown as green shades.

was performed by using Project Gutenberg as input *textual corpus* for the Word2Vec algorithm (see Figure 1).

However, before applying the algorithm, we performed a preparation step[9]. The reasons are (1) to treat different variants as the same word, for example the words "sing", "singing", and "sung" to all be considered `sing[v]`; (2) to reduce the noise of some polisemic words (for example, distinguish the verb `play[v]` from the noun `play[n]`); and (3) to reduce the computational space by (a) ignoring clutter such as stopwords, codes, formulas, or numbers, and (b) reducing multiple variants to a single core meaning, as explained. In addition, terms shorter than 3 letters have been ignored as well as standard english stopwords. We ignored numbers and terms including non-alphanumeric characters (except the single dash). After that, we translated the text into a clean list of of POS-tagged *lemmas*. For example, the first sentence of `RECMUS-619` would become: `introduce[v] Anacreontic[n] Society[n] consist[v] amateur[n] perform[v] admirably[r] best[j] orchestral[j] work[n]`. The process was executed on an Hadoop Cluster with the algorithm as implemented by Apache Spark[10]. We kept the default values for all the parameters (*n-gram* window size: 100, step 10), we tuned the number of partitions to 10.

The produced Word2Vec model (*word embeddings* in Figure 1) was queried to obtain a sorted list of 10.000 words similar to `music[n]` (*dictionary*). Some examples: `melody[n]` (7.8010), `guitar[n]` (6.8451), `inspiriting[j]` (6.3402), `heartful[j]` (4.2634), `psalm[n]` (4.0559)[11].

Clearly, a text about music will have a significant overlap of terms with the dictionary, therefore a higher relevance score (see Formula 3). In order to learn the value of the threshold to apply we analysed the distribution of the dictionary on three reference corpora: LED, as a source of positive examples, and Reu and RED as sources of negative examples. (The LED corpus used for this analysis has been properly purged from the samples we will use in the evaluation.) Figure 2 displays the result of our analysis. The score of LEDs is generally higher then the ones of other corpora. This is not surprising. Particularly, we observe that the items in the negative corpora largely fall in the lower spectrum of the diagram. Following the guidelines of our approach, we decided to derive the threshold as the *average* $\overline{x}$ = 1.078270082 relevance score of a text in LED minus the *standard deviation* $\sigma_{\overline{x}}$ = 0.5711202598: $th_1 = overlinex - \sigma_{\overline{x}}$ = (1.078270082 − 0.5711202598) = 0.5071498222 ($th1$ in Figure 2). At this stage we can have a *clue* about the *musicality* of a text. For example, in `RECMUS-619` many terms have good statistical relation to `music[n]`, such as: `piano[n]` (6.3695), `orchestral[j]` (7.0926), `amateurs[n]` (4.6013), and `trio[n]` (5.6045). However, this is also true for `MASONB-88`, for example the adverb `harmoniously[r]` (4.9675) and the adjective `triumphant[j]` (3.8086). In addition, there are paragraphs where terms related to music may be not appear or be statistically prominent, like in `MASONB-31`.

This method alone is insufficient at capturing the complexity of a themed evidence. Particularly, musical terms can be used in texts that are not reporting a listening event (like the example `MASONB-88`) and musical entities may not appear in the trained embeddings. In addition, literary passages with a pompous style may include terms and figurative expressions that recall music, while not alluding to a *situation* involving the *theme* music. The second phase of our hybrid approach relates to the detection of entities categorised as `dbr:Music` in DBpedia. We adopted the query reported in our method setting distance *%d%* to 5 (higher distances seemed to include many wrong scarsely related entities). These are used in our hybridization phase.

The hybridization step demotes some terms and promotes others, as explained in Section 2. For example, in the case of `MASONB-31`, the NER entity identified the resource http://dbpedia.org/resource/Evening_Prayer_(Anglican) and assigned the score $\overline{x}$ = 1.078270082 (*default score* in Figure 1) to the term `evening[n]`, neglected by the previous step as not related to the core concept. In `MASONB-88`, several terms with purely rethorical role have a high statistically relatedness to `music[n]`, for example, `harmoniously[r]`, `triumphant[j]`, and `amidst[i]`. By demoting those, this text will not be classified as listening experience.

4 EVALUATION

We evaluated our approach with extensive experiments based on a gold standard, comparing with (a) four variants from components of our method and (b) two baseline methods based on alternative hypotheses. In addition, we test our method on a different domain, to demonstrate generality. In summary, we perform the following experiments:

[Hy] Our **Hybrid** method, as described.

[Em] We evaluated our approach against the hypothesis that identifying themed evidence using statistical methods at the linguistic level (e.g. musical-like discourse) would be sufficient. To do that, we implemented a similar pipeline by only considering the method based on our statistical relatedness analysis, without filtering or entity boost (**Embeddings**).

[Ent] We performed tests using the entity detection pipeline alone, assuming that the presence of a musical entity is a sufficient indication of a *music-themed* evidence (**Entities**).

[Em+F] To evaluate the impact of the noise correction part related to entity boosting, we performed tests by only using *part-of-speech* demotion (**Embeddings - filtered**).

[Hy-F] To demonstrate the impact of the noise correction part related to part-of-speech demotion, we include results for a method

[9] We used StanfordNLP for these tasks, particularly the POS-tagging [33]. See also: https://stanfordnlp.github.io/CoreNLP/

[10] Apache Spark: https://spark.apache.org/.

[11] Scores are round up to the fourth digit, for readability.

equivalent to our hybrid approach but without part-of-speach filtering (**Hybrid - unfiltered**).
[Fo] As alternative hypothesis, we trained a Machine Learning classifier abstracting features from three learning corpora (LEDs, RED, and Reu). This first baseline method was implemented using a Random Forest Classifier (**Forest**).
[St] As second baseline method, we developed a pipeline equivalent to *Em* but where the dictionary is generated using a purely statistical approach (**Statistical**).
[Hy/R] Finally, to demonstrate the portability of our approach we tested our hybrid method for the detection of themed evidence on a different domain, the one of *reading experiences*, from the Reading Experience Database Project. The experiment labelled **Hybrid (RED)** reproduced our case study using the core concept `book[n]` and the core entity `dbr:Literature`, keeping all the other elements unchanged.

Before discussing the results, we now describe the gold standard and give details on the two baseline methods **Forest** and **Statistical**.

4.1 Gold standards

To evaluate our approach we developed a gold standard of listening experiences. We selected 500 positive samples from the LED database, sourced from 17 books, and added 500 negative samples. Negative samples were selected from the same sources by identifying one or more paragraphs with a self-consistent meaning and a similar length to the respective positive. The average length of the samples is 125 words. Notably, negative samples include cases where terms are clearly related to music but the text itself does not report a listening event. We evaluated the gold standard by inter-rater agreement. Ten users annotated 200 items each, in addition to the experts setting up the gold standard and selecting the samples. Finally, each item in the gold standard was checked by three raters. We determined the overall agreement between the raters, subtracting out agreement due to chance, using Fleiss' kappa [12]. The resulting value was $k = 0.669$, that is interpreted as *substantial agreement* between the raters. However, in addition to accuracy, we also want to measure how much the gold standard is *pessimistic*, i.e. how difficult it is to distinguish positive from negative samples. For this reason, we performed a term frequency analysis comparing four sets of texts of same length: (1) LED_{gs}: the 500 positive samples; (2) $Reuters_{gs}$: 372 negative samples; (3) RED_{gs}: 1404 negative samples; and (4) NEG_{gs}: the 500 negative samples manually selected. Samples from Reuters and Reu where selected randomly until reaching the amount of words of LE. We compared the corpora with measures used for document similarity, considering each corpus a single document obtained concatenating the samples. We computed the term frequency (TF) measure[13]. Therefore, the similarity between the corpora (documents) can be measured as the number of distinct shared words. $Reuters_{gs}$ and RED_{gs} share with LED_{gs} 6% and 10% of the words respectively. Instead, the manually selected samples have 12% words in common with the positive set (LED_{gs}). In addition, we inspected the top most frequent words in the positives' corpus (LED_{gs}) and checked their frequencies in the other corpora. These are almost absent in $Reuters_{gs}$, have some occurrences in RED_{gs} but are even more present in NEG_{gs}. Ultimately, we computed the *cosine similarity* [18] between the four vectors of terms frequencies: $LED_{gs}/Reuters_{gs}$: 0.267, LED_{gs}/RED_{gs} : 0.376, and LED_{gs}/NEG_{gs} : 0.718 (values are cut to the third digit). We can conclude that the resulting gold standard is both accurate and pessimistic. In order to experiment on the generality of our method, we also developed an equivalent gold standard of *reading experiences* following the same methodology illustrated so far. The related Fleiss' kappa was $k = 0.6813307483$, that is interpreted as *substantial agreement* between the raters. Crucially, the inter-rater agreement score achieved in both gold standards demonstrates that there is enough consenus on what a themed evidence is, despite the variety of forms they can have in a text.

4.2 Baseline methods

Before discussing the results, it is worth reporting details about how the preparation of the baseline methods *Forest* and *Statistical* was conducted.
Baseline method: Random Forest Classifier (Forest). We used a Random Forest Classifier [16] as implemented by Apache Spark[14]. The training set had the following characteristics[15]: (a) 9059 LEs from the LED database as positives (excluding the samples part of the gold standard). (b) A combination of the RED and Reu corpora as negative examples (30770 texts). (c) The feature set was prepared as a Bag of Words including the first 10000 most frequent terms in the LE set. (d) Features were represented as the term frequency of each word in the sample. The resulting collection is then split in a training set (70%), used to train the classifier with labelled samples and a test set (30%), to evaluate its capacity of discerning positive from negatives. The tests where promising and reporting good values for standard measures: F1 = 0.814 and Accuracy = 0.846. Ultimately, we expect this method to be able to distinguish texts that are similar to listening experiences by learning their features from the examples of the LE database.
Baseline method: a dictionary computed using TF/IDF (Statistical). A dictionary equivalent to the one used in the proposed approach could be produced by relying on statistical NLP techniques such as TF/IDF. The Project Gutenberg collection also includes a Music *shelf*. Therefore, we built a dictionary of words occurring in documents classified as music in the Project Gutenberg collection. First, we computed the TF/IDF score with respect to the whole corpus. Next, we selected the books classified in the Music shelf and computed the average TF/IDF value of their words, resulting in an list of unique terms associated with a score. The assumption is that the term will be relevant for the category "Music" if occurring with a high score in more documents of the shelf (79 documents). We applied this dictionary in the same way as the one developed using word embeddings, and conducted a threshold analysis, as described in Section 3.

[12] Fleiss' kappa: https://en.wikipedia.org/wiki/Fleiss'_kappa.
[13] A term here is a POS-tagged word in its abstract form. E.g. *played* will be `play[v]`.
[14] Apache Spark: https://spark.apache.org/.
[15] We tested with many variants, changing number and nature of negative samples or using a different feature set, for example, keeping all the words occurring in the training set, only the ones occurring in the positive examples, or keeping the first n most frequent words. We omit these details, for space reasons, and only report on the best performing variant.

Table 1: Experiments results. The table reports the approach in the first column, the number of items classified as positives (P) among the total 1000. The number of correct positive samples is reported in column C. Measure values are cut to the third digit.

	P	C	Prec.	Rec.	F1	Acc.	Err.
Fo	202	189	0.935	0.378	0.538	0.676	0.324
St	657	474	0.721	0.948	0.819	0.791	0.209
Em	559	457	0.817	0.914	0.863	0.855	0.145
En	808	462	0.571	0.924	0.706	0.616	0.384
Em+F	482	424	0.879	0.848	0.863	0.866	0.134
Hy-F	609	472	0.775	0.944	0.851	0.835	0.165
Hy	551	460	0.834	0.920	0.875	0.869	0.131
Hy/R	534	425	0.795	0.850	0.822	0.816	0.184

4.3 Results

Our experiments employed a number of annotators developed on top of the Stanford NLP library[16] and applied to the gold standard described in Section 4.1. Results are summarized in Table 1. The *Forest* classifier achieves the highest precision but with a very low recall and accuracy slightly above random. This result reinforces the idea that *surface* features (e.g. bag of words) alone are not sufficient to capture complex concepts such as *themed evidence*, even when adopting a large training set. The difference in performance with the testing in the tuning phase (F1 80%) reinforces our argument that our gold standard is actually pessimistic. The *Statistical* classifier achieves a good recall but a low precision, meaning that users looking for *themed evidence* will be asked to review many false positives. The best performing annotator, from the baselines, is the one using a statistical analysis based on word embeddings (Embeddings). However, without applying noise correction, precision is generally lower (as expected). In contrast, adopting the part-of-speech filter improves precision (Embeddings, filtered), reinforcing our assumptions that *facts* are better represented by *verbs* and *nouns* and that a methaphorical use of terms related to the *core concept* is more prominent in other parts of speech such as adjectives or adverbs, which are ignored in this variant. The *Entities* approach alone has a performance slightly above random. This is not surprising and confirms the idea of a robust gold standard involving theme-related entities but not necessarily themed *evidence*. The *Hybrid (Unflitered)* approach boosts the score of any term linked to the DBpedia category without POS-filtering and makes visible how the NER step allows us to uncover underrepresented elements in the dictionary, improving recall. Our *Hybrid* approach incorporates the best of both worlds by focusing on factual components of the discourse (verbs, nouns, and entities) that indicate the presence of a *themed evidence*. Considering the result of the Hybrid (RED) experiment, we can conclude that our method is generally applicable to other domains with minimal configuration. In addition, we calculated Cohen's kappa comparing Hybrid with both annotators used to evaluate the gold standard of listening experiences. The two values, 0.806 and 0.702, demonstrate a *substantial agreement* between our system and human annotators.

[16]Stanford NLP: https://stanfordnlp.github.io/CoreNLP/

5 DISCUSSION

In this section we focus on the errors and try to characterise open challenges for the identification of *themed evidence*. We discuss the performance of our approach (Hybrid) and of the three good-performing methods: Embeddings, Embeddings (filtered), and Hybrid (unfiltered). Comparing the results of all four methods, we can classify the texts in three categories: (1) the texts correctly classified by all four methods - *easy* (787); (2) the ones classified incorrectly by all four methods - *difficult* (86); and (3) the ones identified by some of the four methods - *challenging* (127). First (A), we look for a correlation between amount of errors, the type of error (false positives or false negatives), and the size of the texts. Second (B), we focus on samples correctly classified by some of the four methods (but not all) to discuss the role of each component of our method. Third (C), we select some paradigmatic examples from the set of difficult texts to highlight open issues.

(A) We grouped the samples in three categories, depending on the size: small (<240 characters), medium (240<>1000 characters), and large (>1000). The 787 easy samples included 70% of the small, 75% of the medium, and 88% of the large sized samples: larger texts tend to be easier to classify. In addition, we observed that 20% of the small samples are also difficult, while less than 1% of the medium and large texts. We can conclude from this analysis that generally small texts are challenging to classify. Also, 67% of the difficult texts were negative samples, bringing the conclusion that it is generally harder to avoid false positives then false negatives.

(B) Our hybrid method failed to classify 131 samples of which only 45 were correctly classified by any of the other methods (0.045% of the total set). We now analyse the impact of the *POS-filtering* and the *Entity boost*. 46 items failed by the Hybrid (Unfiltered) method were correctly classified by the Hybrid method with part of speech filter. It is notable how all of the samples in this group are negatives. It is clear how focusing on verbs and nouns contributes to improving precision and accuracy. However, the Embeddings (Filtered) missed to identify 48 positives (and 0 negatives) that were correctly identified by Embeddings, impacting recall negatively. With regard to the Entity boost, 36 items failed by both Embeddings and Embeddings (filtered) methods have been correctly classified by the two hybrid methods. This needs to be attributed to the role of the entity boost. It is notable how all the samples in this group are *positives*. However, the application of the entity boost resulted in 33 false positives that where correctly classified by the statistical method alone as negatives. This shows how the entity boost has a positive role in improving recall while reducing precision.

The length of the texts does not seem to have a significant impact on how the various parts of the method cooperate. Also, POS-filtering improves precision while entity boost contributes positively to recall. We argue that both POS-filtering and entity boost has the effect of compensating some limitations of Embeddings. However, there is still room for improvement, especially to reduce the impact that entity boost can have on precision (mainly related to false positives).

(C) Finally, we present some of the *difficult* samples, the ones incorrectly classified by all the methods considered in this section (the 0.086% of the gold standard, 28 positives and 58 negatives).

This has the objective of illustrating some of the issues discussed and to sketch an agenda for future work.

MASONB.txt-59 (positive): "We did not learn much in relation to church music this day, either in the Moravian or Baptist Chapel.". The term music was the only flagged word in the statistical part of the method. Entities such as `dbr:Chapel` and `dbr:Baptists` have a distance from the category Music that is higher then the value used (5). One research direction is therefore to find a better way of discrimanting the relevance of entities, beyond the distance between the two nodes.

*RECMUS.txt-616 (negative): "Besides this, special mention is made of **Spohr**, who frequently met **Moscheles** at the house of Baron Poifere de Cere.".* Both the highlighted words had high score in the dictionary (the book is taken from Project Gutenberg, source of the embeddings - see Section 3). In addition, they both mapped to DBpedia music-related entities (`dbr:Louis_Spohr` and `dbr:Ignaz_Moscheles`), resulting in a further boost of their score! As a result, all methods wrongly classified this short text as a listening experience. A semantic analysis of the sentence may contribute to uncover issues where relevant entities are mentioned in a situation that does not actually include a listening event.

MASONB.txt-185 (positive): "The churches are large, and filled with altars, monuments, statues, interesting to the eye, and often with music not less attractive to the ear". Music was the only flagged word. No relevant entities mentioned and a figure of speech (*attractive to the ear*) is used. Here, we see how understanding metaphorical language is key to the detection of some themed evidence.

*REOPER.txt-848 (negative): "Madame **Grisi**, at that time in the zenith of her **beauty** and her artistic fame, held in allegiance beneath her fair, soft sceptre, a whole **string** of young fashionables of the day, over whom she had acquired an extraordinary influence".* This is a case where few very high-scored nouns (highlighted) misguided the decision of the system. Also, the term *string* is polysemic and it does not refer to an instrument string in this sentence.

Ultimately, the detection of themed evidence raise many important issues typical of complex retrieval tasks. A reasonable solution needs to incorporate other methods and heuristics within the current approach. This is the main line of research we intend to follow to further improve the performance on the task.

6 RELATED WORK

We consider work related to the retrieval of information for humanities research, information (event) extraction, distributional semantics, and applications of Semantic Web technologies in the humanities.

Retrieval for the humanities and qualitative research. The problem of building a corpus of *pieces of evidence* from literary databases is common in research relying on typical historiographic methods. More in general, the task of classifying texts as relevant to a research *theme* relates to most research areas in the humanities, where the development of new concepts is an intrinsic part of the scholarly enquiry [28]. In information retrieval (IR), *complex concepts* are conceived as short phrases combining two or more words into novel meanings. These are merely considered as noun phrases, *compounds*, or short phrases with compositional meanings ("he was trying to keep his temper") [24]. However, these definitions are insufficient to capture *themed evidence*. Research on text classification and topic modelling tries to tackle the problem of quantifying the relevance of a given piece of text with respect to one or more categories [1]. Complex concepts are also invoked as multi-keyword specifications of a difficult IR task, where users are supported by generating sets of keywords from the neighborhood of a result of interest in an iterative and exploratory process [11]. The problem of keyword expansion and concept-based search is common in digital humanities research [28] and the overcoming of keyword-based approaches a recurring theme in Semantic Web research [12]. Building the right set of keywords for a certain task is a problem. One way to tackle it is to learn these keywords automatically.

Distributional semantics. Distributional approaches to semantics are based on the assumption that the meaning of a word is a function of the contexts in which it occurs [14, 23]. This strand of research is essentially corpus-based and constitutes a rich family of approaches that share a bottom-up perspective on meaning [20]. Approaches to the generation of vectorial representations of words (embeddings) achieve good results in several tasks, particularly in measuring word or document similarity. Hence, we employ word embeddings to measure how much a text is generally *close* to a core concept, *music* in our case study. Context-predicting models are considered superior to purely statistical ones [6]. For this reason, we generate our dictionary using the approach described in [27] in the implementation of Apache Spark Word2Vec. In addition, we compare with a dictionary developed using the classical TF/IDF model, in our evaluation. However, embeddings suffer from a low control on word ambiguity and polisemy. For example, in our case study, the role of musical and sound metaphors in language is broad (*"Sounds good?"*) and can easily be reflected as noise in the distributed model.

Semantic Web technologies for the Humanities. Approaches combining linguistic and semantic technologies for supporting humanities research are a recent, important trend [2, 26]. Hybrid approaches combining information retrieval techniques and semantic technologies have a long tradition (a seminal work on this topic is [36]). Linked Data and NER together have been extensively employed recently in a number of knowledge extraction and data mining tasks including classification (e.g. [29]). In our work, we use DBpedia Spotlight [25] to identify entities in texts and DBpedia as a background knowledge to compensate for the absence of complex, compound entities in the dictionary and to give relevance to words with low statistical significance but high conceptual relevance, as discussed in the paper.

Event Extraction. The task of retrieving *themed evidence* to support qualitative studies can be related to research in *event extraction* - where the goal is to extract event information from texts in domains such as Biomedicine [7], Finance and Politics [19], and Science [34]. The notion of *event* varies across domains but it is generally considered as something happening at a specific time, which constitutes an incident of substantial relevance [17]. The task is therefore to identify the action triggering the event (e.g. the verb *attack*) and then the associated roles. Data-driven approaches usually involve statistical reasoning or probabilistic methods like Machine Learning techniques. In contrast, knowledge-based methods are generally top-down and based on pre-defined templates, for example, lexico-semantic patterns [19]. The two approaches can be

combined and machine learning methods used to learn such patterns [30]. Event extraction is considered usually as a closed domain procedure and one of the open issues relates to the portability of the approaches [17]. Research in *open domain event extraction* focuses essentially on social media data [31] where the task is the extraction of statements, similar to the one of *key-phrases* extraction [35].

However, themed evidence is a broader notion, which incorporates the one of event, where the nature of the event itself is often assumed, implicit, and left to the reader. In our work, we focus on the domain of humanities scholarship and on the task of supporting qualitative research and aim at a general solution capable of supporting different domains of enquiry in cultural studies. In particular, because pieces of themed evidence are a peculiar type of texts strongly characterized by vagueness and indirectness, we opted for a *low-committment* hybridization between two background knowledge bases, a source of *linguistic* competence (word embeddings), and a source of *conceptual* competence (a knowledge graph).

7 CONCLUSIONS

In this paper we tackled a novel problem that relates to the classification of texts as *themed evidence* and contributed a hybrid approach combining statistical reasoning and a knowledge graph. Extensive experiments demonstrate the effectiveness of our method as well as its potential portability to other *themes* beyond the case study of listening experiences. Ultimately, our aim is to set the basis of a novel research line focused on intelligent methods for supporting users in collecting *themed evidence* from texts in a generic, principled, and explainable way.

As next step, we are going to deploy our approach as a end-user service on the LED portal[17] (indeed, an earlier version is already available, implementing the Embeddings method). However, to apply the classifier to the scan of books we also need to consider how best to segment the text. Our approach normalises scores by length, therefore a moving window of flexible sizes (between 5 and 10 sentences) should be able to capture all relevant paragraphs. We intend to expand our experimentation on *reading experiences* and perform a systematic and comparative analysis of the performance and related errors on both gold standards. An interesting research direction is about integrating the current approach to support secondary themes, where the objective would be to identify, for example, musical experiences related to other topics of interest such as childhood, religion, or war.

[17] http://led.kmi.open.ac.uk

REFERENCES

[1] Kjersti Aas and Line Eikvil. 1999. Text categorisation: A survey.

[2] Alessandro Adamou, Enrico Daga, and Leif Isaksen. 2016. Proceedings of the 1st Workshop on Humanities in the Semantic Web co-located with 13th ESWC Conference 2016 (ESWC 2016). CEUR Workshop Proceedings.

[3] Alessandro Adamou, Mathieu d'Aquin, Helen Barlow, and Simon Brown. 2014. LED: curated and crowdsourced linked data on music listening experiences. *Proceedings of the ISWC 2014 Posters & Demonstrations Track* (2014).

[4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer.

[5] Helen Barlow and et al. 2017. Listening to music: people, practices and experiences.

[6] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 238–247.

[7] Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Complex event extraction at PubMed scale. *Bioinformatics* 26, 12 (2010).

[8] Enrico Daga. 2019. Themed Evidence: Listening Experiences - Gold Standard. https://doi.org/10.5281/zenodo.3250645

[9] Enrico Daga. 2019. Themed Evidence: Reading Experiences - Gold Standard. https://doi.org/10.5281/zenodo.3250679

[10] Simon Eliot. 2012. The Reading Experience Database; or, what are we to do about the history of reading? *Dostopno na: http://www. open. ac. uk/Arts/RED/redback. htm (23. marec 2011)* (2012).

[11] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems* 20, 1 (2002), 116–131.

[12] Fausto Giunchiglia, Uladzimir Kharkevich, and Ilya Zaihrayeu. 2009. Concept search. In *European Semantic Web Conference*. Springer, 429–444.

[13] Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* (2014).

[14] Zellig S Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[15] Michael Hart. 1992. The history and philosophy of Project Gutenberg. *Project Gutenberg* 3 (1992), 1–11.

[16] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. IEEE, 278–282.

[17] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska De Jong, and Emiel Caron. 2016. A survey of event extraction methods from text for decision support systems. *Decision Support Systems* 85 (2016), 12–22.

[18] Anna Huang. 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. 49–56.

[19] Wouter IJntema, Jordy Sangers, Frederik Hogenboom, and Flavius Frasincar. 2012. A lexico-semantic pattern language for learning ontology instances from text. *Web Semantics: Science, Services and Agents on the World Wide Web* 15 (2012).

[20] Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics* 20, 1 (2008), 1–31.

[21] Peter Leonard. 2014. Mining large datasets for the humanities. In *80th IFLA General Conference and Assembly*.

[22] David Lewis. 1997. Reuters-21578 text categorization test collection. *Distribution 1.0, AT&T Labs-Research* (1997).

[23] Will Lowe. 2001. Towards a theory of semantic space. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 23.

[24] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

[25] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*. ACM, 1–8.

[26] Albert Meroño-Peñuela, Ashkan Ashkpour, Marieke Van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank Van Harmelen. 2015. Semantic technologies for historical research: A survey. *Semantic Web* (2015).

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[28] Stephen Osadetz, Kyle Courtney, Claire DeMarco, Cole Crawford, and Christine Fernsebner Eslao. 2018. Searching for Concepts in Large Text Corpora: The Case of Principles in the Enlightenment.. In *DH*. 254–256.

[29] Heiko Paulheim. 2013. Exploiting Linked Open Data as Background Knowledge in Data Mining. *DMoLD* 1082 (2013).

[30] Jakub Piskorski, Hristo Tanev, and Pinar Oezden Wennerberg. 2007. Extracting violent events from on-line news for ontology population. In *International Conference on Business Information Systems*. Springer, 287–300.

[31] Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1104–1112.

[32] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427–437.

[33] Kristina Toutanova and Christopher D Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Empirical methods in natural language processing. 38th Meeting of the ACL*.

[34] Maria Vargas-Vera and David Celjuska. 2004. Event recognition on news stories and semi-automatic population of an ontology. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*. IEEE, 615–618.

[35] Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, and Craig G Nevill-Manning. 2005. Kea: Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*. IGI Global, 129–152.

[36] Qi Zhou, Chong Wang, Miao Xiong, Haofen Wang, and Yong Yu. 2007. SPARK: adapting keyword query to semantic search. In *The Semantic Web*. Springer.