Digital Studies / Le champ numérique

Hayton-Ruffner, Dylan, Fernando Nascimento and James Broda. 2022. "Concept Detection in Philosophical Corpora." *Digital Studies/Le champ numérique* 12(1): 2, pp. 1–28. DOI: https://doi.org/10.16995/dscn.8106

Open Library of Humanities

Concept Detection in Philosophical Corpora

Dylan Hayton-Ruffner, Bowdoin College, US, hrdylan@gmail.com
Fernando Nascimento, Bowdoin College, US, fnascime@bowdoin.edu
James Broda, Washington and Lee University, US, jbroda@wlu.edu

During the course of research, scholars often search large textual databases for segments of text relevant to their conceptual analyses. This study proposes, develops and evaluates two applications of word embedding algorithms for automated Concept Detection in theoretical corpora: Average Cosine Similarity Retrieval (ACS) and Word Mover's Distance Retrieval (WMDR). Both strategies are evaluated against weighted keyword (KW) search using a test set from the Digital Ricœur corpus tagged by scholarly experts. In our experiments, WMDR outperformed weighted keyword search on the Concept Detection task, which suggests it is a promising strategy for Concept Detection and information retrieval systems focused on theoretical corpora. Besides these initial positive results, has as its major characteristic the ability to use definitions as proxies for concepts; this provides search results that account for the semantic contexts of theoretical concepts.

Au cours de la recherche, les chercheurs tâchent souvent de trouver des segments de texte pertinents dans d'énormes bases de données textuelles pour leurs analyses conceptuelles. Cet article propose, développe et évalue deux applications d'algorithmes de word embedding (plongement lexical) pour la Concept Detection (détection de concept) dans des corpus théoriques : dans l'Average Cosine Similarity Retrieval (ACS – Extraction de similitudes cosinus moyenne) et dans la Word Mover's Distance Retrieval (WMDR – Extraction de distance de déplacement lexique). Les deux stratégies seront évaluées par rapport à une recherche par mot-clés pondérée avec un dispositif de test venant du corpus Digital Ricœur, qui est étiqueté par des experts érudits. Dans nos expériences, la WMDR était plus performante que la recherche par mot-clés pondérée durant la tâche de détection de concept, ce qui suggère que cela soit une stratégie prometteuse pour la détection de concept et pour des systèmes d'extraction d'information axés sur des corpus théoriques. En outre, la WMDR a comme caractéristique majeure la capacité de se servir de définitions en tant que des proxys pour des concepts. Cela fournit des résultats de recherche qui explique les contextes sémantiques de concepts théoriques.

Introduction

Be careful what you wish for. The academic community has always wanted increased access to primary and secondary texts, and the growing and continuous process of digitization has brought about precisely that. Particularly in extensive theoretical corpora, the initial work of defining the most important texts to read may be an overwhelming challenge. The struggle to locate and access rare physical sources that remained concealed in specialized libraries has been replaced by the equally or even more daunting task of finding the most relevant texts for understanding a particular theoretical concept from a myriad of online sources and references, often available in abundance through online institutional subscriptions.

The standard solution to this new problem of the digital-text era has been the use of keyword searches to narrow down results and guide researchers through the labyrinths of JSTOR, Google Scholar, or their specialized corpus. While this strategy can effectively point to relevant references, it has several shortcomings. First, keyword searches often return irrelevant results that do not take into account linguistic phenomena such as polysemy and synonymy that can only be adequately addressed when the context of words is taken into account by the search strategy. Secondly, the large number of results returned by keyword searches is a problem for systems subject to copyright, in which the number of results must be restricted based on legal parameters. Thirdly, keyword searches do not allow an adequate sorting of results and nuanced classifications. Their sorting strategies usually rely on factors extrinsic to the text itself such as metadata or are based only on the number of hits of the keyword. Fourthly, open solutions for integration with academic corpora focus mainly on complete works rather than specific segments within works such as paragraphs. Fifthly, queries are usually formed by keywords combined by logical operators that, in addition to being complicated for users without great technological experience, do not consider the semantic context of the search terms whose related sections within a given corpus are the real purpose of the searching endeavor. Ideally, one should be able to provide the context of the queried term in natural language to guide the search process. Instead of searching for "justice" + "distribution," one could search for "Justice as a form of proportional distribution." The search algorithm should account for the semantic context of the entire search string in its implementation in such a way that segments containing text strings like "a type of allocation that considers proportion" should be ranked very high in the search results.

This article tackles all dimensions of this problem by applying and comparing recent machine-learning algorithms based on semantic spaces such as word2vec and Word Mover's Distance, that allow for searches based on the definition of a concept in

natural language. Such a strategy allows the phenomena of synonymy and polysemy to be appropriately taken into account and supports a precise ranking of finer-grained results, such as paragraphs, based on a semantic similarity between the search text and the paragraphs of the result set. Consequently, these strategies yield a smaller number of more relevant and complete results compared to keyword search to researchers of extensive theoretical corpora. The central objective is to empower the academic community with a tool that can be integrated into specialized corpora to facilitate the identification of fruitful starting points for exploring specific issues or theoretical concepts that will need to be analyzed and questioned through close reading and methodology for each discipline. We call this process of retrieving textual segments relevant to the study of theoretical Concept Detection.

To validate our strategy, we used the corpus of the French philosopher, Paul Ricoeur, who is representative of the different dimensions of the problem we want to address in this article. First, because Ricoeur's corpus is an extensive theoretical corpus, as we will describe in detail below. Ricoeur has published more than fifty books in French and almost a thousand articles over his seven decades of academic production. Secondly, it is a set of relatively recent texts that is still under copyright protection. Thirdly, because the Digital Ricoeur project (Taylor and Nascimento 2016) already makes available a significant portion of Ricoeur's works digitized and TEI formatted to the academic community, which significantly facilitated the corpus gathering and preparation steps of this experiment.

In the first part of the paper, we present the context of our inquiry and discuss some of the most relevant prior work in the area. We then describe our methodology and data gathering process and the test data set used for the experiments. In the following section, we present and discuss the results of the three information retrieval approaches we have applied to retrieve relevant paragraphs associated with theoretical concepts. Finally, we offer a few qualitative observations and conclude with an overall assessment of the results, their limitations, and the next steps to be pursued in our research line.

Information retrieval and concept detection

The core contribution of this study lies in the Information Retrieval research area. Information Retrieval (IR) is a classic problem in natural language processing (NLP) made all the more relevant by the proliferation of digital text data. At its core, text-based IR helps users deal with large-scale text data by locating, analyzing, and retrieving segments that are typically referred to as documents (Salton and Harman 2003). In general, IR involves a corpus of text segments of a large definite size, and

a user interested in retrieving information from that corpus. The user communicates their information needs in the form of a query—typically a set of words—that indicates the segments they are interested in. The task of the IR system is to process the query, search the corpus, and return relevant segments to the user. In practice, perfect recall—returning 100% of relevant segments—is difficult. Thus, IR systems must balance recall, the number of relevant segments returned out of the total set of relevant segments and precision, and the number of relevant segments out of the total number of segments returned in the query. High recall is useless if precision is too low and vice versa (Raghavan and Wong 1986).

Concept Detection, a special case of IR, is the process of finding and retrieving segments that define and expand upon a given concept. Concept Detection is typically conducted using theoretical corpora from fields like philosophy, psychology, and literature which structure texts around sets of concepts. **Figure 1** visually illustrates the location of Concept Detection within the field of IR.

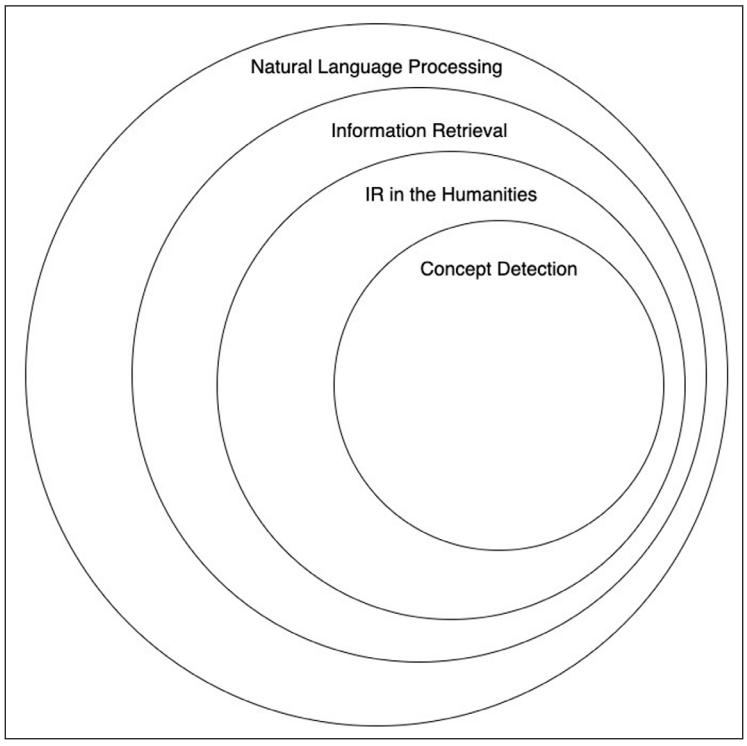

Figure 1: Concept Detection and its research context.

Concepts and the task of analyzing concepts is one of the central characteristics of philosophical thought. It is perceived by many as the hallmark of philosophical systems since Plato through contemporary works both in the continental philosophy, particularly within a phenomenological tradition, and also in analytic philosophy and

logico-linguistic works of philosophers like Russel, Frege, Carnap, Pierce, and others. More recently, concepts and conceptual analysis became one of the key problems of cognitive science. Jesse Prinz, for instance, said that concepts are "the basic timber of our mental lives" (Prinz 2004, 1). Solomon, Medin, and Lynch (1999) and Margolis and Laurence (2015) highlighted the central place of concepts in both philosophy and cognitive science by offering the broad definition of concepts as the building blocks of thought being the basis for classification and categorization, inference, and conceptual combination. But if we take a step beyond this general definition of concepts, we will find ourselves facing philosophical, cognitive, and linguistic dimensions full of ongoing discussions about the origin of concepts (empirical, social, or innate), their structural elements, their ontological status, and the relationship between concepts as cognitive and linguistic entities.

Given the applied context of our work, we will focus on this last linguistic dimension and start from the assumption that, regardless of the precedence and details of the functioning of the human cognitive apparatus, there is a direct relationship between the conceptual cognitive system and the linguistic system. Bergen (2012) and Evans (2009, 2015) described the relationship between language and the conceptual apparatus. They argued that linguistic constructions function as schemes that activate cognitive regions linked to concepts derived from ground experiences and compose these basic concepts in complex contextual structures. In particular, the theory proposed by Evans (2009, 2015) appeals to us because of its flexibility and comprehensiveness in relation to the ways in which basic (analogical) and linguistic (parametric) concepts complement each other for the formation of cognitive models. Evans' theory holds "that analog (body-based) content is supplemented by propositional information derived from linguistically mediated content, thus fleshing out the representations in the conceptual system"(Evans 2015, 279).

For practical purposes in this work, and without delving into the logical intricacies and critique of the intensional logic, we will explore the predications of concepts in terms of Carnap's distinction between extension and intension that is broadly comparable to the Fregean distinction (respectively) between reference and sense (Parsons 2016, 58). Hence, to identify a concept, we will be looking for intensional characteristics that constitute its formal definition, and extensional characteristics that point to the scope of applicability of the concept by naming the particular objects or instances that it refers to. Therefore, if we consider the concept "narrative" in Paul Ricoeur's philosophy, one of its intensional content is "a structure imposed on events, grouping some of them together with others'' (Ricoeur 1984, 58). And some examples of its extensional content would be "fictional," "historic," "tragedy," "novel."

In order to systematize and formally specify the characterization of concepts in a way that can be communicated to experts participating in the creation of training data, and to reviewers of the computational experiment, we adopt the International Organization for Standardization (ISO) standard nomenclature document 1087 that provides an operative definition of a concept that is suitable for the practical objectives of this research. An *object*, as defined by the ISO, is "anything perceivable or conceivable" (ISO 2019). Objects have *characteristics*, which are abstractions of the properties of an object (ISO 2019). *Concepts* combine these characteristics into units of knowledge (ISO 2019). For example, the concept "planet" combines all the characteristics of a planet—round, massive, stellar, etc.—into a single identifiable entity. As we discussed, the set of characteristics a concept combines is called its *intension*. Concepts may also be abstract. The concept "justice" combines the terms "truth," "right," and "law" into an idea of judicial equality.

The standard also uses the definition of the *extension* of a concept as the totality of objects to which a concept corresponds (ISO 2019). Extensions of "planet" might be Saturn, Jupiter, or Earth, but also might include generic objects like "heavenly body" or "astronomical body." A concept can be visualized as a cloud of these extensions semantically related by the characteristics the concepts contain. The concept relates and describes each object, conveying their characteristics in a single unit.

In concept detection, as in IR, a user expresses their information need through a *query*, a word or set of words that refers to a specific concept within a corpus of theoretical texts. The goal of the system is to return text segments to the user that are relevant to the definition of the concept. The quality rather than the quantity of the results returned by the system is paramount. The complexities of theoretical corpora prevent users from processing large volumes of information quickly. Additionally, recent theoretical corpora are often restricted by fair use copyright law. Databases are allowed to display only a small subset of the entire corpus in response to a user's query for a period of 80 years after the author's death. Thus, search algorithms have to make a tradeoff between the number of results shown and the amount of context displayed around each result. Since context is indispensable for a researcher, as it frames the meaning and content of each search result, Concept Detection requires that results be in context. To fulfill this constraint and comply with copyright law (Taylor and Nascimento 2016), Concept Detection queries must return small numbers of high-quality paragraphs, which provide both the content and context required by the researcher.

Concept Detection is vital when conducting a conceptual analysis, in which a researcher explores the expression of a concept in the works of one or more writers (Chartrand et al. 2016). Such studies can cover decades' worth of material and require

large amounts of time to complete. Concept Detection expedites this process by locating all the areas of target works—for instance, paragraphs—that are relevant to and useful for the researcher. Such capacity may empower research communities with quick and easy access to high-quality materials, freeing up scholars to focus on the interpretative and creative effort of expanding and applying a certain theoretical corpus.

In the next section, we provide a quick overview of the prior work in information retrieval with a special focus on developments related to concept search.

Prior work

Keyword search

Early IR research produced primarily term-matching driven algorithms (e.g., Boolean Retrieval and P-Norm), which rely on lexically matching the terms in the query with the terms in the document (Salton, Fox, and Wu 1983; Greengrass 2000). There are two fundamental issues with these keyword-driven approaches: *synonymy*, many words can mean the same thing, and *polysemy*, a single word can mean many things (Deerwester et al. 1990). Consider the case when a query term *q* matches a document term *d*. Because words have many meanings, *q* may have an entirely different meaning from *d* even though they are lexically identical. Consider the case when *q* does not match *d*. Because many words mean the same thing, *q* may mean the same thing as *d* despite their lexical differences.

For example, a user is querying a database looking for content related to presidential speeches. They formulate a query with the words [PRESIDENT, SPEECH]. Consider the following segments:

> The president's mode of speech was difficult to understand.
> Obama gave a good lecture at the beginning of the summit.

Because of synonymy and polysemy, the lexical matches between the two segments and the query are misleading. Segment 1 has multiple terms that match the query: "president" and "speech." In the case of "president," the terms both lexically and semantically match. Both the query and the document refer to a president of the United States. The term "speech" in the document and query lexically matches but differs semantically. The user references a formal address while the document refers to the act of speaking. Although Segment 2 has no lexical matches, it has multiple semantic matches. "President" and "Obama" both refer to US Presidents. "Speech" and "lecture" both reference a formal address. Thus, despite the lexical similarities between the query and Segment 1, Segment 2 is more relevant.

Beyond keyword search

Given the drawbacks of term-matching-driven approaches, further research has focused on computing latent semantic structures in natural language and exploiting those structures to return relevant segments. Vector space models (VSMs) express text segments as vectors in a vector space spanned by the vocabulary of the corpus, computing document relevance as the distance between the document vector and the query vector (Raghavan and Wong 1986). Latent Semantic Indexing (LSI) builds upon VSMs utilizing Singular Vector Decomposition to simplify the vector space and leverage latent semantic structures in the text (Deerwester et al. 1990). Recent advancements, Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA), compute latent semantic structure with Bayesian statistics (Hofmann 1999; Blei, Ng, and Jordan 2003; Wei and Croft 2006). Language modelling has also been applied to the task of information retrieval in approaches like Query-Likelihood Retrieval, KL-Divergence Retrieval, and the Relevance Model (Zhai 2008; Zhai and Lafferty 2001; Lavrenko and Croft 2017).

Although the approaches described above have proved effective, recent developments in natural language processing provide even more powerful tools for IR research efforts like Concept Detection, which rely on effectively capturing semantic relationships in natural language. One such algorithm is word embedding, which computes vector representations of words and has been shown to provide a powerful computational representation of word relationships. While LSI, VSM, and even a simple co-occurrence matrix derives vector representations of words, modern word embeddings utilize neural networks. The original network design, proposed by Mikolov and colleagues in 2013, is simple yet powerful. It has three layers: an input layer, a hidden "embedding" layer, and a softmax output layer (Mikolov et al. 2013). The "embedding" layer is a set of hidden neurons of size d, a hyperparameter indicating the number of embedding dimensions. The dense connections between the input layer and the hidden layer form a matrix of weights, of size n by d, where n is the size of the vocabulary and d is the number of dimensions in the embedding space. Each row of this matrix is a word embedding, or word vector, for a word in the corpus. During training, these weights are altered computing high-quality embeddings for each word (Mikolov et al. 2013). Word embeddings have been shown to improve existing information retrieval systems, like language modelling—by providing an enhanced representation of words (Ganguly et al. 2015; Ye et al. 2016).

Concept Detection in philosophy

While the majority of IR research takes place outside the humanities, there has been a concerted effort within the field to apply the techniques developed for generalized

information retrieval to philosophical corpora. Pulizzotto et al. (2018) evaluate the use of vectorization, linear algebra, K-means, and positive and unlabelled data classification techniques (to extrapolate a greater set of unlabelled data from a smaller set of labelled data) in analyzing the concept of "mind" in the *Collected Papers* of C. S. Peirce. They perform a qualitative analysis and conclude that these techniques helped them overcome the problems of polysemy, synonymy, and ellipsis (instances in which a concept is discussed with no obvious linguistic markers) (Pulizzotto et al. 2018).

Chartrand and colleagues also investigate clustering and conduct a quantitative analysis of their novel clustering-based technique for concept detection: COFIH (Concept-Finding Heuristic) (Chartrand et al. 2016). In COFIH, the corpus is converted into a co-occurrence matrix. The queried concept is expressed as a document consisting of signifiers or concept words and added to the matrix. All segments containing at least one signifying term are extracted. This set is clustered and for each cluster, a prototype or typical vector is built. The entire corpus is then checked for similar vectors to this prototype. The most similar vectors from all the clusters explored become the extension of the concept or the set of text segments relevant to the concept.

COFIH is evaluated on the Concept Detection task using a corpus of the collected papers of C. S. Peirce and compared against expert judgement. COFIH achieves 69% recall. An in-depth analysis of the results for the concept "law" found the retrieved text segments to be of a very high quality. However, COFIH returns more than 10 times the number of segments compared to keyword. According to its authors, COFIH shows promise, but further validation is needed before significant conclusions can be drawn.

In addition to clustering-based IR, neural network-based classification has also been applied to the problem of concept detection. De Pasquale and Meunier explore the use of perceptrons in what they call the "categorization of small segments of text into a set of thematic categories" (De Pasquale and Meunier 2003). De Pasquale and Meunier approach Concept Detection as a supervised learning problem with labelled training data and test sets. They achieve success with some thematic categories reaching nearly 80% recall and 50% precision for the category "knowledge." However, their model is not universally successful, and does not reach much above 50% recall on the rest of their categories. Their limited success with a simple perceptron suggests that more complicated neural networks may be more successful at the categorization process.

Forest and Meunier apply one such neural network, ART1 (Grossberg 1988), to the task of thematic analysis, which involves the "the discovery and identification of the multiples relations between different themes that make a textual corpus consistent and intelligible" (Forest and Meunier 2005). The classifier is used to group segments of

text from Descartes' *Discours de la méthode* into classes. The lexicon of each class is computed, and the resulting set of word groups is used as the baseline for thematic analysis. Like Pulizzotto et al. (2018), Forest and Meunier focus on qualitative analysis and frame the tool as a guide for each reader's journey through the text.

In "Detecting Large Concept Extensions for Conceptual Analysis," Chartrand applies LDA to Concept Detection (Chartrand, Cheung, and Bouguessa 2017). Chartrand uses topics to infer the presence of concepts. The LDA-based algorithm searches for a concept's "signifier" or "concept word" in the topic-word distribution of a topic and matches the topic with that concept based on the presence or absence of such signifying words within the topic. Textual segments—typically referred to as documents—highly associated with the identified topic are determined to be relevant to the query. The algorithm is evaluated on a corpus of law-related segments. Labelled test data is obtained through crowd-sourced tagging and expert judgement. While LDA achieves some success, it fails to score above 18% recall and 65% precision.

Although IR strategies based on word embedding have yet to be thoroughly evaluated in philosophy or the humanities, recent work has yielded promising results. Chartrand and colleagues improved upon their initial model (Chartrand, Cheung, and Bouguessa 2017) by applying LCTM, a variant of the classic LDA topic model that represents topics as distributions of latent concepts using word embeddings (Chartrand and Bouguessa 2019; Hu and Tsujii 2016). The studies find that using a topic model enhanced with word embeddings yields significantly better results on the task of Concept Detection and provides a better representation of the queried concept.

In the following section, we describe our alternative approach to Concept Detection that, like LDA and other strategies, relies on the latest machine learning techniques. Specifically, our model is based on word embeddings that are created using the word2vec algorithm (Mikolov et al. 2013). We start by explaining the corpus used for our application and how we created the test set. We then describe the main methods applied and discuss some characteristics of their application to the specific problem of concept retrieval.

Method and materials

Corpus and test set

A common obstacle in Concept Detection research in theoretical corpora is access to an organized large-scale textual database. In order to overcome this difficulty, we leveraged the Digital Ricoeur project repository which digitizes the works of philosopher Paul Ricoeur into an online database and has amassed a large digital corpus

of philosophical writings providing a unique opportunity for text analysis projects. The corpus used in this experiment consists of 232 French works, 5,980,285 tokens, and 91,621 unique words. While digitized collections exist in both French and English, the French corpus was selected to minimize the effect of human translation, allowing the project to leverage the writer's exact wording.

In order to evaluate a Concept Detection tool, a test set of segments tagged by scholarly experts for their relevance to a set of concepts is required. With this set, the ability of the tool to accurately find and retrieve segments associated with a concept can be assessed against ground truth values backed by scholarly consensus.

Our test set contains paragraphs from *The Symbolism of Evil* (SM), a monographic work that explores several well-defined concepts: "mythe," "symbole," and "evil." Each paragraph in the set is tagged with one of four categorical variables—"Defines," "Relates to," "Sub-concept," "Not related"—indicating its relevance to a set of concepts—"mythe" (myth), "homme" (man), and "symbole" (symbol). These three concepts were selected because they are among the most frequent terms ("homme" is the most mentioned term with 693 occurrences, "mythe" is the second-most frequent term with 612 occurrences, and "symbole" is mentioned 563 times) in the book, *The Symbolism of Evil*, and they express essential aspects of the conceptual core of this work. We also wanted a set of concepts that was both representative and relatively small to facilitate tagging work by specialists. We also find it relevant for our analysis that the concepts "mythe" and "symbole" are more directly linked to this particular work, while the concept "homme" is more comprehensive and appears with great frequency in several other documents in the Ricoeurian corpus. Regarding the criteria shown to experts for each tag see Appendix A. These tags indicate categorical judgements of the segment's relationship to the concept. The tags were provided by four scholarly experts.

To evaluate a Concept Detection algorithm, each segment must be determined to be either "Relevant" or "Irrelevant" to the Concept Detection query. To obtain this set of binary relevance tags, each categorical tag was mapped to one of two relevance tags: "Relevant" and "Irrelevant." The tags "Defines" and "Relates to" were mapped to "Relevant" because they indicate that a segment is either defining or expanding upon the concept. The tags "Sub concept" and "Not Related" were mapped to "Irrelevant."

Because unanimous agreement was uncommon in the mapped test set, the expert's binary relevance tags had to be aggregated into a final relevance judgement for each segment. A majority-rule strategy was used to determine this consensus among the

expert's tags. The tag reaching a simple majority was adopted as the final relevance judgement. In the event of a tie, the segment was labelled "Irrelevant" to the concept. Eighteen percent of consensus determinations ended in a tie. Of the 106 segments in the test set, 39 were determined relevant for "mythe," 37 were determined relevant for "symbole," and 8 were determined relevant for "homme."

Concept Detection methods

This section describes the three Concept Detection strategies evaluated in this paper. All three retrieval strategies are search ranking tools, which, given a concept query and corpus, rank segments according to their relevance to the queried concept.

Weighted keyword retrieval (Weighted KW)

Weighted keyword retrieval, a commonly used search tool in academic research, is the baseline for this study. In keyword retrieval, the user provides a keyword corresponding to a concept (i.e., "mythe"). Keyword retrieval then iterates through the corpus and assigns each paragraph a score based on the number of occurrences of the keyword. The paragraphs are then ranked, and the top *n* paragraphs are returned as "relevant" textual expressions of the concept in the corpus. **Figure 2** provides a pseudocode description of weighted KW Retrieval.

Algorithm 1: Weighted Key Word Retrieval (KW)

```
Input  : Key_word
Input  : Corpus − [paragraph1, paragraph2, paragraph3...]
Input  : N
Output: Returns N paragraphs determined relevant to the concept
results = [];
foreach paragraph p_i ∈ Corpus do
    matches = 0;
    foreach word w_i ∈ p_i do
        if w_i Key_word then
            matches += 1;
        end
    end
    results.add(p_i, matches);
end
results.sortbymatches();
return top N results
```

Figure 2: Key Word Retrieval pseudocode.

Word embeddings

While weighted KW is a ranking-based implementation of term-matching approaches to information retrieval, the following algorithms and applications, ACS and WMD

Retrieval, make use of word embeddings. While many high-quality word embedding models are now open source and readily available, our research required domain specific embeddings which closely reflect the semantic relationships in Ricoeur's corpus. Achieving this specificity without sacrificing model quality is difficult as word embeddings models often require large amounts of training data (Mikolov et al. 2013). We solved this problem by initializing our model with embeddings for NLPL's Model 43, a generalized, open-source, French-language model trained on web documents (Kutuzov et al. 2017). The model was then trained on the Digital Ricoeur corpus, described in the Methods and Materials, for 200 epochs with an embedding size of 100 using the skip-gram training strategy. We used the word2vec implementation available in the open-source package Gensim (Rehurek and Sojka 2010).

Average Cosine Similarity Retrieval (ACS)

This section proposes Average Cosine Similarity Retrieval (ACS), a simple, word embedding-driven retrieval algorithm for concept detection. The user provides a conceptual query comprising a single keyword (e.g., [FREEDOM], [EVIL], [JUSTICE]) corresponding to a specific concept. Segment relevance to the concept is calculated as the average cosine similarity between the embeddings of the component words of the segment and the embedding of the query's keyword. Given a queried keyword represented by the word vector q and given a document d consisting of words represented by the word vectors $[w_1,w_2,...,w_n]$,, the relevance of the document to the query is calculated as: $\frac{1}{n}\sum_{i=1}^{n}\frac{q\cdot w_n}{\|q\|\|w_n\|}$

Each segment in the corpus is scored and the top segments are returned to the user. In practice, small segments are ignored as their length makes the ACS calculation volatile. Our version of ACS labels all segments under 30 words in length "Irrelevant" to the concept.

ACS can be conceptualized as a nuanced keyword search, leveraging not only the keyword, but also the most similar words to the keyword in the word2vec model. Because these words co-occur frequently with the keyword in the corpus, it is likely that they represent either characteristics or extensions of the concept. Therefore, the presence of these words could indicate the presence of the concept.

If this assumption is correct, ACS has the potential to improve upon keyword search in two areas. First, keyword retrieval search struggles with cases of synonymy. Because many words can mean the same thing, there may be segments that are relevant to the concept but do not contain the keyword. However, segments like these will contain many words related to the concept and the keyword. If these words are close to the keyword in the embedding space, ACS will give the segment a high score even though the keyword is absent. Second, keyword search also struggles with cases of polysemy.

Because a word can mean many things, some segments will contain the keyword but not be relevant to the concept. However, it is likely that these segments will contain many words unrelated to the concept and the keyword. If these unrelated words are far from the keyword in the embedding space, then ACS will assign the segment a low score despite the presence of the keyword. **Figure 3** provides pseudocode for ACS Retrieval.

Algorithm 2: Average Cosine Similarity Retrieval (ACS)

Input : Key_word
Input : $Corpus - [paragraph1, paragraph2, paragraph3...]$
Input : Min_size
Input : N
Output: Returns N paragraphs determined relevant to the concept
results = [];
remove_small_segments(Min_size, $Corpus$);
foreach *paragraph* $p_i \in Corpus$ **do**
    sum = 0;
    foreach *word* $w_i \in p_i$ **do**
        sum += cosine_similarity(w_i, Key_word);
    end
    score = sum/len(p_i);
    results.add(p_i, score);
end
results.sortbyscore();
return top N results

Figure 3: ACS Retrieval pseudocode.

Word Mover's Distance Retrieval (WMDR) for Concept Detection

Although ACS leverages words similar to the keyword, there is no guarantee that these words represent the characteristics or extensions of a concept. Thus, algorithms that take groups of words as queries provide a promising evolution from the one-word approach of ACS and keyword. Moreover, while in past approaches like Boolean retrieval multiword queries were restricted to Boolean expressions (e.g., "mythe" or "symbole"), a strategy capable of processing and leveraging a query expressed in natural language would be both easier for users to work with and better able to manage the semantic complexities of theoretical corpora.

Word Mover's Distance, as proposed by Kusner, Sun, Kolkin, and Weinberger in 2015, is a word embedding-driven metric for measuring the semantic dissimilarity between segments (Kusner et al. 2015). Word Mover's Distance is a special case of the earth mover's distance (EMD) or Wasserstein metric, which is used to measure the distance between probability distributions in statistics. Intuitively, the earth mover's distance can be thought of as the work required to "cover" one distribution with the other.

WMD is a measure of the amount of work it would take to change the word embeddings of one segment into the word embeddings of another segment. Each segment is represented as a distribution of words, and EMD is used to calculate the distance between these two distributions using a word embedding model as a metric space (Kusner et al. 2015).

In WMD retrieval (WMDR), our application of Word Mover's Distance (WMD) for Concept Detection tasks, paragraphs are ranked by their similarity to a canonical definition of the concept. This definition works as a proxy for the concept itself. The user provides a query consisting of a textual segment (the concept definition) of natural language that defines the concept of interest. Each paragraph in the corpus is given a score based on its Word Mover's Distance to the query segment. Stop words are removed from both the query and the segments in the corpus before calculating similarity. The top segments are returned to the user.

WMD retrieval has the potential to resolve several issues that plague keyword search. First, its use of word embeddings provides a powerful representation of natural language, helping with issues of polysemy and synonymy. More importantly, it allows for the formation of queries in natural language, which is not only easier for users, but can also provide more semantic information than sets of keywords. Unlike ACS, WMDR explicitly requires the user to provide a definition: a set of words in context that represent characteristics and extensions of the concept. Thus, paragraphs that are similar to the query text likely define or relate to the concept. The definition paragraph also contains the fingerprint of the author's definition style—words like "defines" and "relates" that structure the author's explanations. Other segments that express definitions will contain this fingerprint and have a high similarity to the definition paragraph. **Figure 4** provides pseudocode for WMDR.

Algorithm 3: Word Mover's Distance Retrieval (WMDR)

Input : $Definition$
Input : $Corpus - [paragraph1, paragraph2, paragraph3...]$
Input : N
Output: Returns N paragraphs determined relevant to the concept
results = [];
foreach *paragraph* $p_i \in Corpus$ **do**
    distance = WMD($Definition$, p_i);
    results.add(p_i, distance);
end
results.sortbydistance();
return top N results

Figure 4: WMD Retrieval pseudocode.

Experiment setup

The performance of weighted KW, ACS, and WMDR on the task of Concept Detection was evaluated using the Digital Ricoeur test set described in Methods and Materials. Each algorithm was used to rank all paragraphs in the test set for three concepts: "mythe," "symbole," and "homme." Stop words were removed from all paragraphs before ranking. The stop word list from Spacy's fr_core_news_sm model was used (Honnibal and Montani 2017). ACS and WMDR used embeddings from the specialized model trained on the Ricoeur corpus. We utilized the Word Mover's Distance implementation available in Gensim (Pele and Werman 2008; Pele and Werman 2009; Rehurek and Sojka 2010). First, the Concept Detection strategy was provided with the paragraphs from the test set and a query referring to that concept. For ACS and weighted KW, the keyword associated with the concept was provided (e.g., "mythe"). For WMDR, a canonical definition from both inside and outside the test set was provided. Definitions from inside the test set were removed when ranking that concept. For the full definitions used, see Appendix B. With these two inputs, the strategy returned the paragraphs ranked by their relevance to the concept. From this ranking we computed precision at K ($P@K$), a measurement of the precision of the search results up to and including the K-ranked paragraph (Manning, Raghavan and Schutze 2008). We collected $P@K$ for values of K from 1 to 20. While calculating precision, the expert judgements were used as the ground truth. **Figure 5** outlines the structure of the experiment.

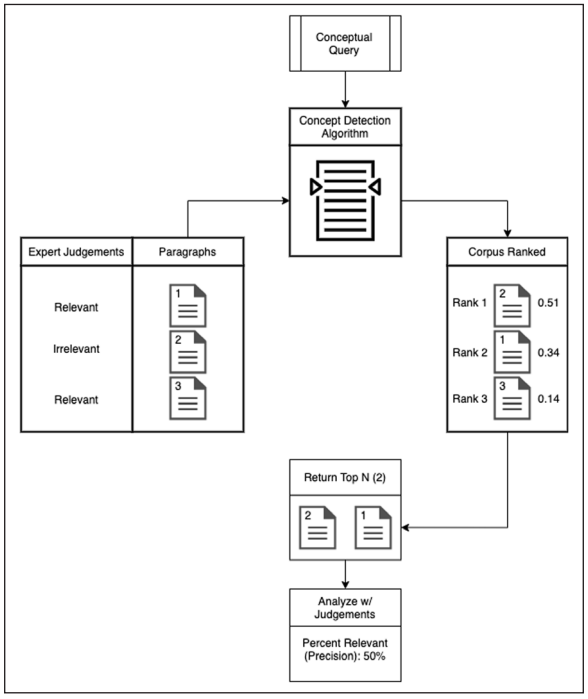

Figure 5: Experiment setup.

Results

"Mythe"

Figure 6 graphs precision at K (P@K) for all three Concept Detection methods for the concept "mythe." P@K values for the concept "mythe" were noticeably high with all three strategies reaching precision values above 0.6 for almost all values of K. In addition, all three strategies achieve perfect precision at low values of K (1–3). However, WMDR continues to show higher precision values as K increases (through K = 20), although the difference in precision between WMDR and the other two strategies narrows. While ACS has higher precision values than weighted KW at lower values of K (1–6), it equals or has lower precision at higher values of K (6–20).

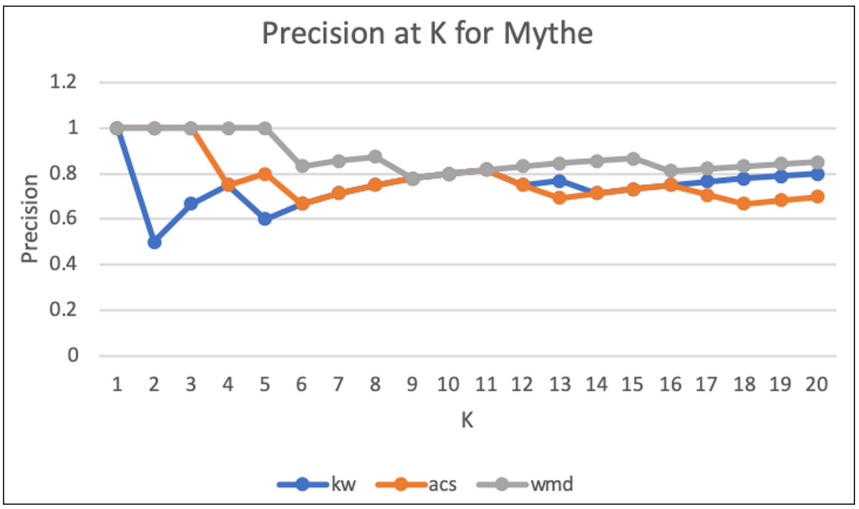

Figure 6: P@K for "mythe".

"Symbole"

Figure 7 graphs P@K for all three Concept Detection methods for the concept "symbole." Just as with the concept "mythe," in the concept "symbole," WMDR has higher precision than both ACS and KW at low values of K (1–10), with a noticeably larger gap in precision at these values of K compared to the results from "mythe." However, at higher values of K (10–20), WMDR achieves lower precision than the benchmark, weighted KW, by roughly 10%. ACS has lower precision than weighted KW at almost all values of K (4–6, 8–20).

"Homme"

Figure 8 graphs P@K for all three Concept Detection methods for the concept "homme." For the concept "homme," all three strategies have very similar values of precision from K = 4 through K = 20 with WMDR reaching a higher precision that both ACS and

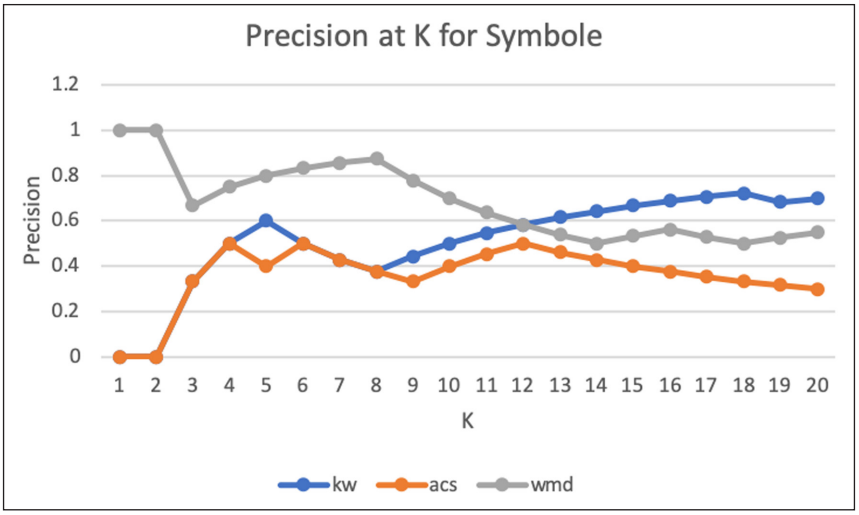

Figure 7: P@K for "symbole".

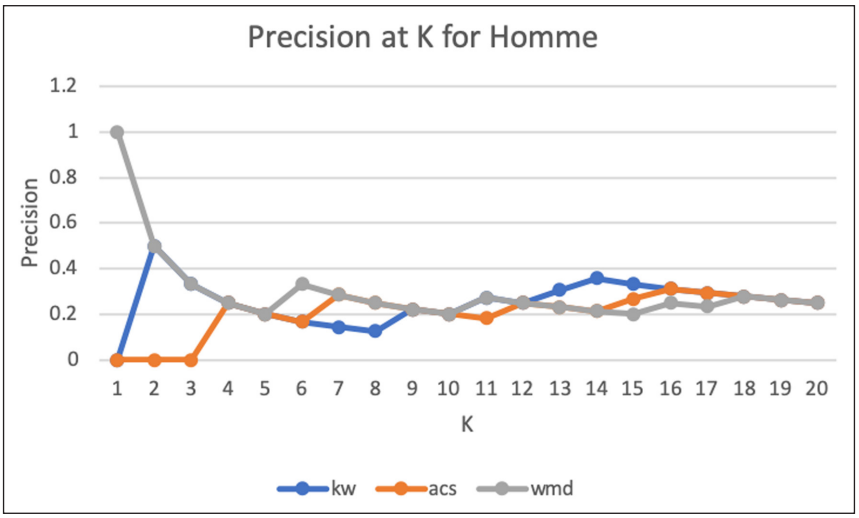

Figure 8: P@K for "homme".

KW at K = 1, and ACS exhibiting a lower precision than both WMDR and KW from K = 1 to K = 4. Overall, the precision values reached by the three strategies are noticeably lower (~40%) compared to the results from "mythe" and "symbole."

Discussion

As shown in Results, WMDR has higher precision than both weighted KW and ACS at values of K from 1 to 10 in two out of three concepts. However, at higher values of K, WMDR scores higher in one concept, lower in one concept, and roughly equal in one concept. Thus, WMDR outperforms the benchmark, weighted KW, at low values of K

(1–10), a promising indication that it is well-suited to the task of concept detection. While WMDR does have a mixed performance at values of K (10–20), this finding is less relevant when the constraints of Concept Detection are considered. As mentioned in the introduction, returning many full segments is often impossible because of fair use copyright law, which restricts the percentage of the corpus that can be returned from a search. Moreover, a high number of search results runs the risk of overloading the user. Thus, WMDR's success in identifying and ranking highly small numbers of good quality paragraphs indicates that it may be a powerful solution for concept detection. Despite these promising initial findings, the size of the Digital Ricoeur test set limits the extent of this finding. To make a stronger claim about the efficacy of WMD Retrieval in philosophical and theoretical concept detection, a larger test set is required.

ACS fares worse than WMDR, performing less than or equal to both WMDR and weighted KW at all values of K across the three concepts. There are several possible reasons for this. First, a simple average is not nuanced enough of a metric and heavily affected by both segment size and keyword density. Second, a single word may not be enough to capture the intension of the concept. The word "symbole" may have much to do with the concept symbol, but the cloud of words that surround it within the word2vec model do not fully define the concept.

All three strategies perform noticeably worse on the concept "homme." This is due to the concept's sparsity in the text. Only 8 segments from the test set were deemed relevant by the expert consensus strategy.

Qualitative considerations

Before evaluating the segments returned by the three strategies, we first discuss the expected distribution of concepts in the test set. We chose two concepts, "mythe" and "symbole," that are widely developed by Ricoeur in the book *La Symbolique du Mal.* The third concept "homme" is not directly addressed in this work and the use of the concept word is polysemic, offering significant difficulty for the consensus in the interpretation of specialists.

Figure 9 graphs the combined relevance score of the top 20 segments for each of the three concepts. Relevance score was calculated by assigning each tag a numeric value based on its indication of concept presence: Defines – 4, Relates to – 3, Sub concept – 2, Not related – 1. The segments tags were summed resulting in an overall relevance score between 4 and 16. While "mythe" and "symbole" are comparable in terms of ratings for the top 20 high-ranked paragraphs, "homme" is clearly perceived as less prevalent in the test set, which coincides with our expectations, provides another validation point for the experts rating, and informs some of the observations of this section.

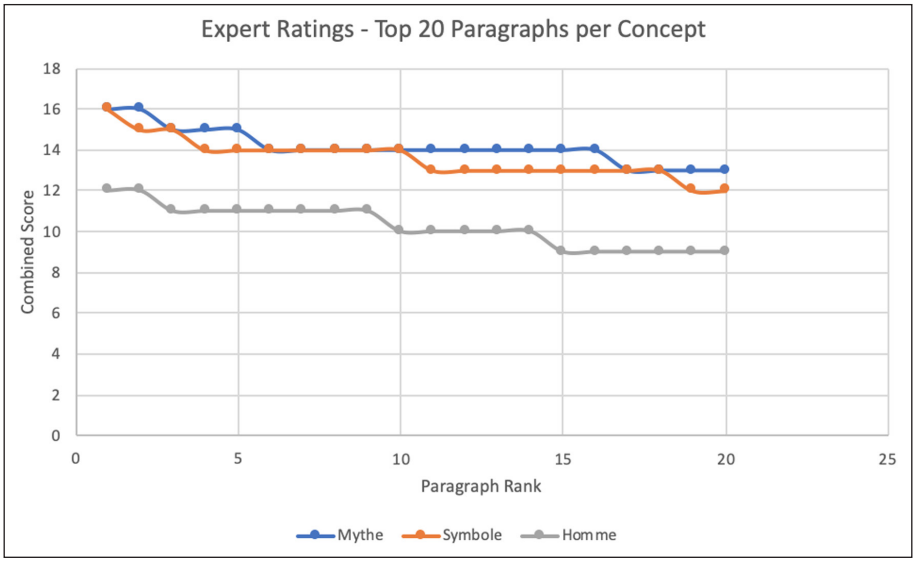

Figure 9: Expert scores for top 20 segments per concept.

As expected, the three strategies performed significantly worse in the results for "homme" than for "mythe" and "symbole" and this result is linked to the scarcity of significant segments for the "homme" concept. The top segment of weighted KW[1] for "homme" (man) is an interesting example of how multiple occurrences of the concept word may be misleading to this type of algorithm. Although it contains the concept word "homme" five times, no expert considered it a "defines" for the concept "homme," one rated as "relates to," two rated it as "sub-concept" and one as "not related." WMDR ranked this same paragraph 427 in fourth showing slightly less sensitivity to multiple occurrences of the keyword.

It is also relevant that WMDR ranked segment 405[2] as its top segment for the concept "homme" (man) in agreement with experts (three ranked it as "relates to"),

[1] Paragraph 427 – "c'est cette structure mythique elle même qui ramène à la diversité des 'mythes' quelle est en effet la signification ultime de cette structure mythique elle désignerait nous dit on l'accord intime de l'"homme' du culte et du 'mythe' avec la totalité de l' être elle signifierait une plénitude indivisible où le surnaturel le naturel et le psychologique ne seraient pas encore scindés mais comment le 'mythe' signifie t'il cette plénitude le fait essentiel est que cette intuition d'un englobant cosmique dont l'"homme' ne serait pas séparé et cette plénitude indivise antérieure à la scission du surnaturel du naturel et de l'humain ne sont pas donné mais simplement visées c'est en intention seulement que le 'mythe' restitue quelque intégrité c'est parce qu'il a lui même perdu cette intégrité que l'"homme' la répète et la mime par le 'mythe' et le rite l'"homme' primitif est déjà l'"homme' de la scission dès lors le 'mythe' ne peut être qu'une restauration ou une restitution intentionnelle et en ce sens déjà symbolique."

[2] Paragraph 405 – "enfin platon lui même malgré le 'mythe' orphique de l'âme exilée dans un corps qui l'ensevelit malgré par conséquent la tentation de durcir le 'symbole' de la captivité corporelle dans une gnose du corps méchant malgré même les gages qu'il donne pour l'avenir à cette gnose sait parfaitement que la captivité corporelle ne doit pas être prise à la lettre mais comme signe du serf arbitre la clôture du corps n'est finalement que l'œuvre du désir et celui qui concourt le plus à charger l'enchaîné de ses chaînes c'est peut être lui même phédon de ainsi la captivité du corps et

despite the fact that the segment does not contain the concept word itself. This could be an indication of a sensitivity to semantic relationships and an ability to deal with polysemy and synonymy.

WMDR performed quite well with regards to the concept "mythe" which is a key concept in the book *La Symbolique du Mal*. Of the 20 top paragraphs identified by WMDR, 17 were ranked as relevant by the experts' consensus. Moreover, if we consider the three misclassified segments according to the experts' consensus, all of them were semantically close to the concept. Paragraph 433 was rated as "defines" by two experts and as "sub-concept" by two experts. Paragraph 8 was rated as "sub-concept" by three experts and as "relates to" by the fourth expert. Paragraph 426 was rated as "defines" by two experts and as "sub-concept" by two experts. Overall, out of 80 experts' ratings for the top 20 paragraphs identified by WMDR, 47 were rated "defines" by experts, 16 as "relates to," 17 as "sub-concept," and none as "not related." The top-ranked paragraph by WMDR (431),[3] was rated as "defines" by three experts and as "sub-concept" by the fourth. The same expert evaluation applied to the second-highest ranked paragraph (7). Weighted KW also performed well for "mythe." It misclassified only 4 out of the top 20 paragraphs, according to the experts' consensus, performing only slightly worse when compared to WMDR.

The third concept analyzed, "symbole," is also another core concept of Ricoeur's book. Both weighted KW and WMDR performed significantly worse for this concept in comparison to "mythe." In the top 20 paragraphs for weighted KW, 7 were rated as non-relevant by the experts' consensus while 9 of the top 20 paragraphs for WMDR were non-relevant according to experts. Even having more non-relevant paragraphs, WMDR outperformed weighted KW in the average precision@K metric as discussed above. On a fine-grained analysis, it is interesting that the two highest-ranked paragraphs by weighted KW were rated as non-relevant by the experts' consensus. Both of them were rated as "sub-concepts" by two experts, "relates to" by one expert and "defines" by the fourth. This result may be related to the difficulty of distinguishing between segments related to or defining a given concept and its sub-concepts when only searching for

même la captivité de l'âme dans le corps sont le 'symbole' du mal que l'âme s'inflige à elle même le 'symbole' de l'affection de la liberté par elle même le déliement de l'âme assure rétrospectivement que son liement était liement par le désir fascination active passive auto-captivité se perdre ne signifie pas autre chose."

[3] Paragraph 431 – "mais pourquoi le 'mythe' en se scindant prend il la forme du récit ce qu'il faut comprendre maintenant c'est pourquoi le modèle exemplaire auquel le 'mythe' et le rite font participer affecte lui même l'allure d'un drame c'est en effet parce que le signifié ultime de tout 'mythe' est lui même en forme de drame que les récits dans lesquels la conscience mythique se fragmente sont eux mêmes tissus d'événements et de personnages parce que son paradigme est dramatique le 'mythe' est lui même événementiel et ne se donne nulle part ailleurs que dans la forme plastique du récit mais pourquoi le 'mythe' récit renvoie t il symboliquement à un drame."

keywords. Paragraph 404,[4] ranked 4th by weighted KW, is an interesting example as it is defining a sub-concept (the symbol of the serv-arbitre) and not the concept of symbolism itself.

The top 2 results returned by WMDR were both rated as relevant by the experts' consensus, and the top result (paragraph 420)[5] is particularly meaningful as it has been rated as "defines" by the four experts and captures critical aspects of Ricoeur's definition of symbol in this work. This paragraph has also been captured by weighted KW, but it has been ranked in the 12th position.

Conclusion

In this article, we applied and compared the results of three information retrieval strategies to search for semantically relevant paragraphs for concepts in a theoretical corpus. Our test data set consists of segments from the book *La Symbolique du Mal* written by the French philosopher Paul Ricoeur, which is part of the collection made available by the Digital Ricoeur portal. The paragraphs in this test data set were analyzed by four specialists in Ricoeur's philosophy and classified according to their semantic relationship with three concepts of Ricoeurian philosophy: "homme" (man), "symbole" (symbol), and "mythe" (myth).

As the benchmark retrieval algorithm for this project, we used a more robust version of the widely applied keyword search strategy that considers the number of times the concept word appears in each paragraph to classify the relevance of the returned paragraphs.

---

[4] Paragraph 404 – "ce qui nous confirme que le symbolisme du serf arbitre bien qu'enfoncé encore et comme noyé dans la lettre des représentations démoniaques est déjà à l'œuvre dans l'aveu du suppliant de babylone c'est que le même symbolisme de l''homme' aux membres liés se retrouve chez les écrivains qui ont usé de ce 'symbole' avec une claire conscience que c'était un 'symbole' ainsi saint paul sait il que l''homme' est inexcusable bien que le péché soit dit régner dans les membres dans le corps mortel rom 6 12 et que le corps lui même soit dit corps de péché rom 6 6 et l''homme' entier asservi au péché si Saint Paul ne parlait pas symboliquement du corps de péché comme d' une figure du serf arbitre comment pourrait il s'écrier si vous avez jadis offert vos membres comme esclaves à l'impureté et au désordre de manière à vous désordonner offrez les de même aujourd'hui à la justice pour vous sanctifier rom 6 19 le 'symbole' du corps asservi c'est le 'symbol' d' un être pécheur qui est à la fois acte et état c'est à dire d'un être pécheur dans lequel l'acte même de s'asservir s'abolit comme acte et retombe en état le corps est le 'symbole' de cette liberté oblitérée d'un constitué dont le constituant s'est évacué dans le langage de saint paul pacte c'est l'offrende du corps à la servitude si jadis vous avez offert vos membres comme esclaves l'état c'est le règne que le péché ne règne donc plus dans vos corps mortels une offrande de moi même qui est en même temps un règne sur moi même voilà l'énigme du serf arbitre de l'arbitre qui se rend serf."

[5] Paragraph 420 – "qu'est donc le 'mythe' en deçà de sa prétention étiologique qu'est donc le 'mythe' s'il n'est pas gnose c'est encore une fois à la fonction du 'symbole' que nous sommes renvoyés le 'symbole' avons nous dit ouvre et découvre une dimension d'expérience qui sans lui resterait fermée et dissimulée il faut donc montrer en quel sens le 'mythe' est une fonction de second degré des 'symbole's primaires que nous avons explorés jusqu'à présent."

We then compared the benchmark results with two algorithms that we applied to the problem in question. Both use machine learning to create a vector space that seeks to represent the semantic relationships between words in the corpus. While ACS averages the distances between the embedding of the concept word and the embeddings of each word in a paragraph, WMDR uses word embeddings to estimate the similarity (distance) between a textual segment written in natural language (in our case a concept definition) and a paragraph.

As described in the results section, WMDR outperformed both weighted KW and ACS in the three concepts considered in the experiment for values of precision at K up to 10. WMDR's performance on the Digital Ricoeur test set indicates that WMDR retrieval is well suited to the task of Concept Detection and capable of outperforming common techniques like keyword search. Since a small number of highly relevant paragraphs is the ultimate goal of this type of application, WMDR retrieval is a promising algorithm that could improve the performance of Concept Detection tools in theoretical corpora.

Although promising, these results need to be expanded and verified using a larger test set consisting of theoretical corpora tagged by experts for concept presence. The great difficulty for this next step is the process of semantic analysis (tagging) performed by experts in the discipline related to the corpus. To that end, we are working to develop an effective tool for evaluation and calculation of consensus among specialists. Such a tool should pave the next steps of the project. In addition, we are working on other information retrieval strategies that use new word embeddings implementations and transformers algorithms which are important future lines of research in concept detection.

We believe that the results of this experiment and research approach can be used in digital libraries of theoretical corpora to improve research results and offer more relevant results to users. Such an improvement can have a significant positive effect on the quality of research around a given theoretical corpus, especially considering the growing volume of works available in digital format, by freeing up research time previously devoted to scouring books and articles, for writing, analysis, and other creative intellectual activities.

Appendix A: Types of semantic relationships

** DEFINES – The segment describes characteristics (3.2.4) of the concept. These characteristics are essential to a proper understanding of a concept.

Example: The segment "The fragile offshoot issuing from the union of history and fiction is the assignment to an individual or a community of a specific identity that we can call their narrative identity." DEFINES the concept of narrative identity.

** SUB-CONCEPT – The segment describes characteristics of a subordinate concept (3.2.14) of the concept.

Example: The symbolism of evil is a sub-concept of the concept symbolism; narrated time is a sub-concept of the concept of time; configuration is a sub-concept of threefold mimesis.

** RELATES TO – The segment describes an associative relation (3.2.23) between the concept and another concept. Differs from sub-concepts in that the concepts do not have a hierarchical relationship.

Example: The segment "This connection between self-constancy and narrative identity confirms one of my oldest convictions, namely, that the self of self-knowledge is not the egotistical and narcissistic ego whose hypocrisy and naivetée the hermeneutics of suspicion have denounced, along with its aspects of an ideological superstructure and infantile and neurotic archaism." RELATES TO the concept of narrative identity.

** NOT RELATED – The segment is not related at all with the concept.

Example: The segment "Our comparison between analytic working-through and the work of the historian facilitates the transition from our first to our second example. This is borrowed from the history of a particular community, biblical Israel. This example is especially applicable because no other people has been so overwhelmingly impassioned by the narratives it has told about itself." is NOT RELATED to the concept of narrative identity.

Appendix B: WMDR definitions

Mythe

Le mythe est une fonction de second degré des symboles primaires. Le signifié ultime de tout mythe est lui-même en forme de drame que les récits dans lesquels la conscience mythique se fragmente sont eux-mêmes tissus « d'événements » et de « personnages »; parce que son paradigme est dramatique, le mythe est lui-même événementiel et ne se donne nulle part ailleurs que dans la forme plastique du récit. Mais en perdant ses prétentions explicatives le mythe révèle sa portée exploratoire et compréhensive, ce que nous appellerons plus loin sa fonction symbolique, c'est-à-dire son pouvoir de découvrir, de dévoiler le lien de l'homme a son sacré.

Homme

l'homme est capable de poursuivre le plaisir pour lui-même et d'en faire un motif autonome. L'homme est organiquement et psychologiquement fragile. l'homme est étonnant pour

l'homme ; l'union de l'âme et du corps ne peut manquer de scandaliser l'idéalisme naturel à l'entendement diviseur. L'homme est homme par son pouvoir d'affronter ses besoins et parfois de les sacrifier.

Symbole

Le symbole est le mouvement du sens primaire qui nous fait participer au sens latent et ainsi nous assimile au symbolisé sans que nous puissions dominer intellectuellement la similitude. C'est en ce sens que le symbole est donnant; il est donnant parce qu'il est une intentionnalité primaire qui donne analogiquement le sens second.

Competing interests

The authors have no competing interests to declare.

Contributions

Authorial contributors

Authorship is alphabetical after the drafting author and principal technical lead. Author contributions, described using the CASRAI CredIT typology, are as follows:

Author name and initials:

James Broda – jb

Dylan Hayton-Ruffner – dhr

Fernando Nascimento – fn

Authors are listed in alphabetic order. The corresponding author is fn.

Conceptualization: fn;

Methodology: jb, dhr, fn;

Software: dhr;

Formal Analysis: jb, dhr, fn;

Investigation: jb, dhr, fn;

Data Curation: dhr;

Writing – Original Draft Preparation: dhr, fn, jb;

Writing – Review & Editing: jb, dhr, fn;

Editorial contributors

Section editor:
Iftekhar Khalid, The Journal Incubator, University of Lethbridge

Copy editor:
Shahina Parvin, The Journal Incubator, University of Lethbridge

Layout editor:
Christa Avram, The Journal Incubator, University of Lethbridge

References

Bergen, Benjamin K. 2012. *Louder Than Words: The New Science of How the Mind Makes Meaning*. New York: Basic Books.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Chartrand, Louis, and Mohamed Bouguessa. 2019. "Mixing Syntagmatic and Paradigmatic Information for Concept Detection." *arXiv [cs.CL]. arXiv*. Accessed August 05, 2021. http://arxiv.org/abs/1904.04461.

Chartrand, Louis, Jackie C. K. Cheung, and Mohamed Bouguessa. 2017. "Detecting Large Concept Extensions for Conceptual Analysis." *Machine Learning and Data Mining in Pattern Recognition*: 78–90. Accessed August 05, 2021. https://link.springer.com/chapter/10.1007/978-3-319-62416-7_6. DOI: https://doi.org/10.1007/978-3-319-62416-7_6

Chartrand, Louis, Jean-Guy Meunier, Davide Pulizzotto, José López González, Jean-François Chartier, Ngoc Tan Le, Francis Lareau, and Julian Trujillo Amaya. 2016. "CoFiH: A Heuristic for Concept Discovery in Computer-Assisted Conceptual Analysis." In *Proceedings of the 13th International Conference on Statistical Analysis of Textual Data* 1:85–95.

De Pasquale, Jean-Frédéric, and Jean-Guy Meunier. 2003. "Categorisation Techniques in Computer-Assisted Reading and Analysis of Texts (CARAT) in the Humanities." *Computers and the Humanities* 37 (1): 111–118. DOI: https://doi.org/10.1023/A:1021855607270

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science*. DOI: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Evans, Vyvyan. 2009. *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199234660.003.0007

———. 2015. "What's in a Concept? Analog versus Parametric Concepts in LCCM Theory." In *The Conceptual Mind: New Directions in the Study of Concepts*, edited by Stephen Laurence and Eric Margolis, 251–284. London: MIT Press. Accessed August 5, 2021. http://www.jstor.org/stable/j.ctt17kk9nr.

Forest, Dominic, and Jean-Guy Meunier. 2005. "NUMEXCO: A Text Mining Approach to Thematic Analysis of a Philosophical Corpus." *Text Technology* 14 (1): 33–45. DOI: https://doi.org/10.16995/dscn.247

Ganguly, Debasis, Dwaipayan Roy, Mandar Mitra, and Gareth J. F. Jones. 2015. "Word Embedding Based Generalized Language Model for Information Retrieval." In *SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 795–798. New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/2766462.2767780

Greengrass, Ed. 2000. *Information Retrieval: A Survey*. Accessed August 05, 2021. https://archive.org/details/Ed_Greengrass___Information_Retrieval_A_survey/mode/2up.

Grossberg, Stephen, ed. 1988. *Neural Networks and Natural Intelligence*. Vol. 637. Cambridge, MA: The MIT Press.

Hofmann, Thomas. 1999. “Probabilistic Latent Semantic Indexing.” *Proceedings of the 22nd Annual International ACM*. DOI: https://doi.org/10.1145/312624.312649

Honnibal, Matthew, and Ines Montani. “spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing,” 2017.

Hu, Weihua, and Jun’ichi Tsujii. 2016. “A Latent Concept Topic Model for Robust Topic Inference Using Word Embeddings.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics 2:* 380–386. Berlin: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/P16-2062

ISO (International Organization for Standardization). 2019. “ISO 1087:2019(en) Terminology Work and Terminology Science — Vocabulary.” *Online Browsing Platform (OBP). 1087. ISO*. Accessed September 1, 2021. https://www.iso.org/obp/ui/#iso:std:iso:1087:ed-2:v1:en.

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. “From Word Embeddings To Document Distances.” In *International Conference on Machine Learning*: 957–966. Accessed September 1, 2021. http://proceedings.mlr.press/v37/kusnerb15.html.

Kutuzov, Andrei, Murhaf Fares, Stephan Oepen, and Erik Velldal. 2017. “Word Vectors, Reuse, and Replicability: Towards a Community Repository of Large-Text Resources.” In *Proceedings of the 58th Conference on Simulation and Modelling*: 271–276. Linköping University Electronic Press.

Lavrenko, Victor, and William B. Croft. 2017. “Relevance-Based Language Models.” *ACM SIGIR Forum* 51(2): 260–267. DOI: https://doi.org/10.1145/3130348.3130376

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze. 2008. “Evaluation in Information Retrieval.” *Introduction to Information Retrieval*: 139–161. Cambridge: Cambridge University Press. Accessed August 06, 2021. DOI: https://doi.org/10.1017/CBO9780511809071.009

Margolis, Eric, and Stephen Laurence, eds. 2015. *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, MA: The MIT Press. Accessed August 06, 2021. http://www.jstor.org/stable/j.ctt17kk9nr.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *arXiv [cs.CL]*. arXiv. Accessed August 06, 2021. http://arxiv.org/abs/1301.3781.

Parsons, David. 2016. *Theories of Intensionality: A Critical Survey*. Berlin: Springer. DOI: https://doi.org/10.1007/978-981-10-2484-9

Pele, Ofir, and Michael Werman. 2008. “A Linear Time Histogram Metric for Improved SIFT Matching.” In *Computer Vision – ECCV 2008*: 495–508. Berlin: Springer. DOI: https://doi.org/10.1007/978-3-540-88690-7_37

---. 2009. “Fast and Robust Earth Mover’s Distances.” In *2009 IEEE 12th International Conference on Computer Vision*: 460–467.

Prinz, Jesse J. 2004. *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.

Pulizzotto, Davide, Jean-François Chartier, Francis Lareau, Jean-Guy Meunier, and Louis Chartrand. 2018. “Conceptual Analysis in a Computer-Assisted Framework: Mind in Peirce.” *Umanistica Digitale* 2. Accessed August 06, 2021. https://doaj.org/article/bac7cf05059149d5ad4f132c0ba43d5a.

Raghavan, Vijay V., and S. K. M. Wong. 1986. "A Critical Analysis of Vector Space Model for Information Retrieval." *Journal of the American Society for Information Science*. Accessed August 06, 2021. DOI: https://doi.org/10.1002/(SICI)1097-4571(198609)37:5<279::AID-ASI1>3.0.CO;2-Q

Rehurek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New*. Citeseer. Accessed August 06, 2021. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.695.4595.

Ricoeur, Paul. 1984. *Time and Narrative, Volume 1*. Chicago: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226713519.001.0001

Salton, Gerard, Edward A. Fox, and Harry Wu. 1983. "Extended Boolean Information Retrieval." *Communications of the ACM*. dl.acm.org. Accessed August 06, 2021. https://dl.acm.org/doi/pdf/10.1145/182.358466?casa_token=2wqGCKm4Mr0AAAAA:NxJ0KFf_32O-i9g6SZsY-G0ouR_HO0QsY7DrQb7_IsaHsdewrhKVbw7LVPmH6FwibgZmtxEzTt4DW2M.

Salton, Gerard, and Donna Harman. 2003. "Information Retrieval." In *Encyclopedia of Computer Science*: 858–863. John Wiley and Sons Ltd.

Solomon, Karen O., Douglas L. Medin, and Elizabeth Lynch. 1999. "Concepts Do More Than Categorize." *Trends in Cognitive Sciences* 3 (3): 99–105. DOI: https://doi.org/10.1016/S1364-6613(99)01288-7

Taylor, George H., and Fernando Nascimento. "Digital Ricœur." *Ricoeur Studies/Etudes Ricoeuriennes* 7, no. 2 (2016). DOI: https://doi.org/10.5195/ERRS.2016.383

Wei, Xing, and William B. Croft. 2006. "LDA-Based Document Models for Ad-Hoc Retrieval." *Proceedings of the 29th Annual International ACM*. dl.acm.org. Accessed August 06, 2021. https://dl.acm.org/doi/abs/10.1145/1148170.1148204?casa_token=eDA-1idhXjsAAAAA:tvuLa3UlBmAPQW51Ay7GruqTlvJFr7G9rinHA4P3_dQoqQ2ALFfaKfMGOQqxRFC5KMld54L6C562DAk.

Ye, Xin, Hui Shen, Xiao Ma, Razvan Bunescu, and Chang Liu. 2016. "From Word Embeddings to Document Similarities for Improved Information Retrieval in Software Engineering." In *Proceedings of the 38th International Conference on Software Engineering*, 404–415. ICSE '16. New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/2884781.2884862

Zhai, Chengxiang. 2008. "Statistical Language Models for Information Retrieval." *Synthesis Lectures on Human Language Technologies* 1 (1): 1–141. Accessed November 18, 2021. https://www.morganclaypool.com/doi/pdf/10.2200/S00158ED1V01Y200811HLT001. DOI: https://doi.org/10.2200/S00158ED1V01Y200811HLT001

Zhai, Chengxiang, and John Lafferty. 2001. "Model-Based Feedback in the Language Modeling Approach to Information Retrieval." In *Proceedings of the Tenth International Conference on Information and Knowledge Management*. Accessed August 06, 2021. https://dl.acm.org/doi/abs/10.1145/502585.502654?casa_token=crNN3eJZaycAAAAA:UpG9PZ_R3SfMU6zsGnkDExfVFYejDGHILPn98yTCcl02VsiWTeQj3FeJS1qWSbTJXPjgGAv-6vt5W3E.