Universidade Estadual de Campinas
Instituto de Computação

Fagner Leal Pantoja

Semantic Representations based on Language Models

Representações Semânticas baseadas em Modelos de Linguagem

CAMPINAS
2025

Fagner Leal Pantoja

Semantic Representations based on Language Models

Representações Semânticas baseadas em Modelos de Linguagem

Tese apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Supervisor/Orientadora: Profa. Dra. Claudia Maria Bauzer Medeiros
Co-supervisor/Coorientador: Prof. Dr. André Santanchè

Este exemplar corresponde à versão final da Tese defendida por Fagner Leal Pantoja e orientada pela Profa. Dra. Claudia Maria Bauzer Medeiros.

CAMPINAS
2025

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

P196s Pantoja, Fagner Leal, 1987-
Semantic representations based on language models / Fagner Leal Pantoja. – Campinas, SP : [s.n.], 2025.

Orientador: Claudia Maria Bauzer Medeiros.
Coorientador: André Santanchè.
Tese (doutorado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Computação.

1. Processamento de linguagem natural (Computação). 2. Web semântica. 3. Ciência de dados. 4. Mineração de dados (Computação). 5. Grandes modelos de linguagem. I. Medeiros, Claudia Maria Bauzer, 1954-. II. Santanchè, André, 1968-. III. Universidade Estadual de Campinas (UNICAMP). Instituto de Computação. IV. Título.

Informações complementares

Título em outro idioma: Representações semânticas baseadas em modelos de linguagem
Palavras-chave em inglês:
Natural language processing (Computer science)
Semantic Web
Data science
Data mining
Large language models
Área de concentração: Ciência da Computação
Titulação: Doutor em Ciência da Computação
Banca examinadora:
Claudia Maria Bauzer Medeiros [Orientador]
Ronaldo dos Santos Mello
Ricardo Rodrigues Ciferri
Ariadne Maria Brito Rizzoni Carvalho
Hélio Pedrini
Data de defesa: 30-06-2025
Programa de Pós-Graduação: Ciência da Computação

Objetivos de Desenvolvimento Sustentável (ODS)
ODS: 3. Saúde e bem-estar
ODS: 4. Educação de qualidade

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-1784-5512
- Currículo Lattes do autor: http://lattes.cnpq.br/3730346542804597

Claudia Maria Bauzer Medeiros

Ronaldo dos Santos Mello

Ricardo Rodrigues Ciferri

Ariadne Maria Brito Rizzoni Carvalho

Hélio Pedrini

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Dedicatória

Aos meus parentes cabanos, à Jessica e aos que se dedicam ou se interessam por estes estudos da computação.

A linguagem é polissêmica, requer interpretação em fatores linguísticos e extralinguísticos.
Para entender o que o outro diz, não basta entender suas palavras, mas também seu pensamento e suas motivações.
(atribuído a Vygotsky)

A system trained only on form has a priori no way to learn meaning.
(Bender and Koller)

Agradecimentos

Agradeço à minha orientadora professora Claudia por todos ensinamentos, pelas orientações, pela oportunidade, pela disposição de seu tempo e por todo cuidado durante este percurso. Uma pessoa que admiro muito e que é uma fonte de inspiração para os alunos do laboratório LIS. Agradeço ao meu co-orientador professor André Santanchè pelos ensinamentos acadêmicos e pela oportunidade de ingressar no doutorado.

À minha mãe e meu pai, Francisca e Amarildo, por todo amor e esforço depositado em minha criação e na garantia de uma educação formal de qualidade. Aos meus irmãos, Linda e Albert, que foram os primeiros em que busquei inspiração na vida e também pela existência dos meus queridos sobrinhos Sophia, Maitê e Arthur. Agradeço a minha família como um todo, meus avós, a todos os meus tios que também sempre foram grandes incentivadores e a todos meus queridos primos.

À minha amada Jessica por todo amor, carinho, parceria, paciência e disposição. Tenho a sorte de partilhar a vida com alguém tão especial. Agradeço por estar comigo durante este percurso que tanto nos custou. Nossas conversas me abrem horizontes e também me ajudaram a amadurecer algumas ideias presentes na tese.

Aos membros do laboratório LIS que sempre estiveram disponíveis para fazer contribuições, sugestões e críticas ao trabalho. Em especial a Ivelize Bernardo, Jacqueline Midlej, Luana Loubet, Lucas Batista, Heitor Mattosinho, Matheus Mota, Marcos Felipe, Kleber Oliveira, Victor Roth, Márcio Saraiva, Leandro Taccioli, Lucas Carvalho, Juan Beleño, Felipe Cardoso, Luiz Celso e Francisco Nardi. Para mim, é uma honra integrar um laboratório que reúne tantos pesquisadores competentes.

Agradeço aos professores do Instituto de Computação (IC) que através de suas disciplinas expandiram minha visão acerca da computação. Agradeço também aos funcionários do IC e em especial aos integrantes da Secretaria pelo apoio durante este período. Agradeço a Universidade Estadual de Campinas.

Agradeço aos professores que compuseram a banca examinadora pelas sugestões e críticas ao trabalho e por seus esforços para participar da apresentação. Em especial, agradeço ao professor Hélio Pedrini pelas contribuições ao longo da elaboração do trabalho. Aos doutores Marco Antonio Filho e Tiago Grangeia pela colaboração no projeto Jacinto.

Aos amigos que apoiaram e compreenderam minha ausência durante a pós-graduação. Aos profissionais Marcia Polaro, Flávia Câmara e Bruno Brabo que me acompanharam terapeuticamente durante este período de intenso desgaste emocional e mental que é a pós-graduação.

Agradeço à minha empregadora Universidade Federal do Pará (UFPA), e ao CTIC, pelo apoio e pela concessão da licença que possibilitou a realização da defesa, bem como pela compreensão diante da minha dupla jornada como servidor e estudante. Por fim, manifesto minha gratidão ao sistema público de educação brasileiro. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Os modernos Modelos de Linguagem (e.g., GPT e BERT) fornecem novas abordagens para Representação de Semântica, encapsulando, em espaços vetoriais, os padrões estatísticos subjacentes aos textos. Entretanto, apesar das pesquisas associadas, ainda existem lacunas quanto ao uso de modelos de linguagem aplicados a criação, ao gerenciamento e à análise de textos médicos. Com o objetivo de preencher algumas destas lacunas, neste trabalho projetamos e analisamos dois tipos de representações semânticas: Anotações Semânticas e Modelagem de Tópicos. Tais representações viabilizam a incorporação de significado semântico em formatos de dados interpretáveis por máquinas. Nossas contribuições utilizam, como pano de fundo, dados públicos do domínio clínico.

Três de nossas contribuições estão relacionadas às Anotações Semânticas. A primeira é um modelo de linguagem – que chamamos de *Envoy* – especializado em Reconhecimento de Entidades Nomeadas biomédicas. Este modelo serve como base para construir duas contribuições adicionais associadas com anotações semânticas. O Envoy é acionado pelo método *Harena Semantics* (nossa segunda contribuição) para realizar a anotação semântica de conceitos relevantes contidos no texto de entrada. Uma característica distintiva de nossa abordagem é a superposição de anotações realizadas por humanos com aquelas inferidas automaticamente pelo Envoy. Em um estudo de caso, aplicamos o Harena Semantics para produzir nossa terceira contribuição: o *Paciente Virtual Semântico*, uma representação semântica que modela casos clínicos como uma rede de conceitos conectada à Web Semântica. Nossos resultados preliminares sugerem um potencial promissor para o engajamento de criadores de recursos semânticos.

A quarta contribuição diz respeito à Modelagem de Tópicos. Aqui, o modelo Envoy é utilizado para elicitar tópicos semânticos a fim de representar uma coleção de Casos Clínicos extraídos do *corpus* CliCR. Com esse próposito, desenvolvemos uma nova abordagem chamada ABT (*Attention-based Topics*), uma representação estatística baseada em modelos de tópicos. Nesta linha, o ABT produz tópicos por meio de uma Agregação Hierárquica aplicada às sentenças de entrada, representada em um espaço vetorial inferido pelo modelo de linguagem BERT. Os resultados da validação indicam que os tópicos produzidos exibem: (1) bons valores na métrica Coerência de Tópicos; e (2) diferentes graus de especialização/generalização, de acordo com o modelo de linguagem utilizado como base.

Abstract

Modern Language Models (e.g., GPT and BERT) provide new approaches to Semantic Representation by embedding in vector spaces the statistical patterns underlying texts. Despite advances in related research, there are gaps concerning the application of language models to the creation, management and analysis of medical texts. In order to fill some of these gaps, we designed and analyzed two types of semantic representation, namely, Semantic Annotations and Topic Modeling. These representations enable the integration of semantic meaning into machine-interpretable data formats. Our contributions use public open data from the clinical domain.

Three of our contributions are associated with Semantic Annotations. The first one is a language model – called *Envoy* – specialized in biomedical Named Entity Recognition. The model is used as a basis to construct two additional contributions concerning Annotations. Envoy is invoked by our *Harena Semantics* method (which corresponds to the second contribution) to perform the semantic annotation of relevant concepts inside medical texts. A distinctive feature of our approach is the superimposition of annotations added by humans with annotations inferred by the Envoy model. As a case study of Harena Semantics, we applied it to produce our third contribution: the *Semantic Virtual Patient*, a semantic representation that models clinical cases as a network of concepts linked to the Semantic Web. Our preliminary results suggest a potential for engaging semantic resource creators.

The fourth contribution concerns Topic Modeling. Here, the Envoy model is used to elicit semantic topics to represent a collection of Clinical Cases extracted from the CliCR corpus. To this purpose, we developed a new approach, called ABT (*Attention-based Topics*), a statistical representation based on topic models. In this line, ABT produces topics through a Hierarchical Aggregation applied to the sentences contained in a given reference corpus, which is represented in a vector space inferred by the BERT language model. The results of our evaluation showed that the topics produced exhibit: (1) good values according to the Topics Coherence metric; and (2) different degrees of generalization/specificity according to the language model used as a basis.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Motivation

The arrival of Language Models [18, 17, 38, 146, 98, 13, 115, 23, 116, 120, 46, 3] has propelled the term “Artificial Intelligence” to international attention. In fact, what everyone (including the population in general, the press, and countless scientists and politicians) calls “artificial intelligence” is just a subfield thereof – namely, machine learning using Large Language Models (LLM) [3, 88, 144].

One of the main reasons for such a technological leap was the establishment of the Transformer architecture [147]. The Transformer architecture is a Deep Neural Network capable of holding and processing massive amounts of information more efficiently than previous approaches. There are many language models based on the Transformer architecture, for instance, BERT [46] and GPT [120].

In this work, we experiment with BERT-derived models in two Natural Language Processing (NLP) applications to produce Semantic Representations [150, 2] for case studies in the clinical domain: (1) Semantic Annotations [48, 73]; and (2) Topic Modeling [22, 62, 7]. These two distinct applications require processing and analysis of large volumes of textual information.

This work focuses on technical, practical and theoretical aspects of language models working with textual constructions for health-related applications. We are specifically interested in investigating how language models can support the creation, management and analysis of medical textual data.

1.2 Terminology

Throughout this text, we use specific terms according to the following definitions. From time to time this will be restated, to help the reader.

- Semantic Representations: The term *semantic representation* has different meanings in different contexts, although closely related, in linguistics, philosophy, cognition sciences, computational linguistics and computer science as a whole. In this work, we use it to refer to computational techniques to represent human language [150]. A semantic representation “reflects the meaning of the text as it is understood by a language speaker. It is produced by information extraction methods that are reliable and computationally efficient" [2].

- Semantic Annotations: A semantic annotation describes a resource (usually a textual resource), by means of formal concepts [48]. It is the process to connect some digital content to the resources of the Semantic Web, towards linking such contents (in particular, text) to networks of ontologies [73]. In our work, annotations are marked directly on the free-text content of a (clinical) textual construction, through the usage of markdown-derived tags `{ }` and `( )`, following the pattern: `{text}(annotation`$_1$`, ..., annotation`$_n$`)`. The markdown tag curly braces surrounds the text, marking the resource, and is followed by the parenthesis tag indicating the referred concept. The annotation can refer to an ontological concept – i.e., a concept formally defined within an ontology – in the form of `(ontology:concept)` or refer to a named entity as `(entity)`.

- Topic Model: A topic model is a technique used in information retrieval to extract hidden semantics from a collection of textual documents by identifying the main topics referred to within that collection – e.g., see [1]. It is a semantic representation that gives an overview of a textual collection under analysis. Topic models are explicit representations [22] that probabilistically associate documents with topics and topics with words. It explicitly represents the latent semantic structure [7] – or gist [62] – of a textual collection. Our topic model comprises a set of topics T. Each topic $t \in T$ is characterized by a set of sentences S and a set of words W.

- Language Model: This is a computational abstraction that represents a unit of textual information (such as a word, a sentence, a paragraph, or a document) through a numerical vector that encodes the most relevant characteristics of this textual unit that are typically induced by a deep neural network [146]. This set of relevant characteristics alludes to a type of semantics grounded on the Distributional Hypothesis [93], which states that words that occur in the same context tend to have similar meanings. Language models are also known as feature vectors, word representations, word embeddings, vector representations, or word vectors [18, 146, 23].

1.3 Context of the Research and Research Questions

The immense amount of textual information available on the internet has the potential to be used as sources of knowledge ready to be consumed by humans and machines. As has been repeatedly seen in recent years, in particular with the appearance of deep learning, such sources of knowledge can leverage human knowledge to new levels. Nevertheless, despite a large volume of research to process such information, there is still much to be done, in particular to extract and process such content in order to use it in a reliable and ethical way. This claim applies to any data type; this thesis is, however, concerned with textual data only. Hence, from now on, our claims should be considered under this context, though some can be generalized to other data types – e.g., videos, or sound.

One of the main endeavors towards processing of textual big data concerns the extraction of the embodied semantics and its incorporation in machine-interpretable data formats. Our work is particularly concerned with uncovering the statistical semantics latent in textual information – this also being an important aspect of tools such as ChatGPT and other language models. In order to accomplish such a goal, it is necessary to approach many long-standing problems within the information processing research agenda. In this work, we examine two among such research questions:

(RQ1): How to encode structured semantics directly into textual constructions?

(RQ2): How to reveal and extract the latent semantics which is kept hidden underlying the statistical patterns occurring in textual constructions?

We have four main contributions to answer these questions, detailed further on in section 1.5:

1. *Envoy*, a neural method to recognize biomedical concepts.
2. *Harena Semantics*, a hybrid method (mixing neural and ontology-based approaches) to annotate extra semantics into biomedical texts.
3. *Semantic Virtual Patient*, a semantic representation which encodes clinical cases in a network of biomedical concepts. This representation is the product of applying the semantic annotation method that uses the Envoy model as one of its components.
4. *Attention-based Topics* (ABT), a neural method to elicit topics to summarize a collection of texts, which can then be used to create semantic representations.

1.3.1 Encoding Semantic Representations

The issues raised by our research questions concern the devising of Semantic Representations [62] to efficiently encode semantic information in computational abstractions. The related literature mostly concerns methods grounded on mainly two distinct traditions to deal with knowledge representation: structured representations and statistical representations [62]. More generally, the distinction between structured and statistical representations has roots grounded on the debate between Rationalism and Empiricism perspectives of epistemology [12, 92].

In this thesis, we propose to address the problem of semantic representation making use of both types of approaches – namely, structured and statistical – drawn from two distinct (but increasingly related) research areas:

- Semantic Annotations [48, 73]: a process to link content (in our case, text fragments) to the network of ontologies hosted on the Semantic Web [19], embedding extra semantics into content. This line of research provides the means to answer our first question, since it deals with so-called “external semantics”.

- Topic Modeling [22, 62, 7]: a method to identify the main topics covered by textual collections. This line of research involves the second question, since it deals with inner semantics (aka latent semantics or statistical semantics)

We claim our semantic representations grounded on concepts from these two research areas can be used in a complementary way to provide computers with means to deal with semantic meaning. Inner semantics is captured by the statistics in language models, and outer semantics is added through the use of ontologies.

With this work, we hope to narrow the gap between these two types of semantics: (1) external semantics to be embedded in textual constructions to enrich their content; and (2) the inner semantics inherently held in statistical patterns underlying written texts. More specifically, the main contributions involve two kinds of representation:

- A structured representation: an ontology-based method which adds semantics directly onto the textual body of free-text constructions.

- A statistical representation: a neural method to extract statistical semantics from free-text collections.

1.3.2 Case studies applied to the health domain

Our proposal is applied to case studies in the clinical domain, using a variety of biomedical data sources. Among them, the main data sources used are Clinical Cases [135] (or Case Reports), which are textual narratives centered on the description of health conditions and complaints patients. A clinical case can be, among others, associated with depicting rare diseases, unusual presentation of common conditions, differential diagnosis, decision making or novel treatments. Typically, a clinical case contains many words referring to diseases, anatomic features, medicines, symptoms, treatments adopted, etc. Case reports are moreover a valuable source of health knowledge that may be leveraged as computable artifacts to feed a wide-range spectrum of applications.

We have used three biomedical data sources along this research:

- The clinical case repository of Harena [45]. The Harena system, designed and developed by the research group in which this thesis was developed, is an e-learning environment based on case resolution that is used as a supporting pedagogical tool for medical students.

- CliCR [135]: a large collection of 10.538 real clinical case reports collected from BMJ Case Reports. The data span the years 2005–2016.

- ACD: a dataset we have created in order to train/test our method for Biomedical Named Entity Recognition. We created it from the concatenation of two other existing datasets: BC5CDR [86] (whose sentences contain chemicals and diseases) and AnatEM [118] (containing anatomies). Thus, the ACD sentences contain words referring to the entities: anatomy, chemical (compounds) and diseases.

1.4 Related Work - a brief overview

This research is concerned with semantic representations, and also involves contributions in topic elicitation. Thus, related work concerns both issues. Chapter 2 covers additional related work on Language Models with applications in topic modeling and clinical studies. This brief overview of related work also contrasts some of the results with our research.

1.4.1 Semantic Annotations and Language Models

In the context of semantic annotations based on Language Models, the work [11] studies the use of LLMs for annotation of input data to improve accuracy of downstream tasks, like semantic similarity and semantic search. There are also works comparing the performance of LLMs and human annotators [104, 24]. Our work proposes to combine both kinds of annotations (LLM-based and human-performed), fostering human-system interaction and the potential for community engagement in the construction of semantic resources.

Examples of research that is concerned with semantic annotation of medical data include [31, 158, 70]. The work of Irwin and coleagues [70] describes a methodology for creating a semantic representation for information extracted from dental exams. Their semantic representations are created combining Concept Maps (to represent concepts identified via a Bayesian method) and semantic networks (to represent relationships among concepts). Differently, our work is based on a neural network which assigns label probabilities to the words of clinical cases and link it to ontology concepts.

Campillos-Llanos et al., [31] present a tool for annotations of Electronic Health Records (EHR) based on a hybrid approach combining BERT-based models, ontologies and lexicon-rules. Unlike this proposal, our work does not rely on rules, thus alleviating the need of intensive labour by domain experts to specify the rules. Our annotated data is produced as a by-product of the authoring process by using lightweight tags provided by the Harena Semantics.

Typically, the semantic annotation process involves some kind of NER (Named Entity Recognition) procedure. Lee et al., [84] introduce BioBERT, a BERT-based model specialized in the biomedical language. BioBERT trained several models to recognize different named entities (e.g., one model to handle diseases, another one to deal with proteins, and so on). We based our NER implementation on the BioBERT model, extending it to recognize multiple entities (anatomy, chemicals, and disease) in a single model.

Other initiatives also address NER by extending language models, refining word vectors, or fine-tuning the model to specialized datasets [83, 14, 91, 97]. BERN [77] is a neural biomedical multi-type NER tool based on the BioBERT model. It uses a separate NER model for each entity type, and then combines the results using decision rules to handle overlapping entities (i.e., when the models tag a term with different tags, for instance, two models tagging the term “androgen” as gene and chemical). In contrast, our NER model does not need decision rules, since the language model itself contains the statistical context to decide which entity is the most likely to occur. This distinction is due to our model being trained on a dataset containing all the entities of interest, thus avoiding training one model for each entity.

1.4.2 Topic Modeling and Language Models

Many methodologies address the Topic Modeling problem, such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) [22]. The first one is based on linear algebra and the latter is a probabilistic model. The LDA uses two complementary distributions: a topic over words distribution that describes the relationship between topics and words; and a second distribution that allocates topics to documents [27]. LDA is one of the most robust and efficient methods for topic modeling. The reader is directed to [37, 1] for surveys on reviewing and classifying methods for topic models.

In the health context, Song et al., [133] "jointly infer specialist-disease and ICD-coded diagnoses as the latent topics" from electronic health records. In a subsequent work, Song et al., [132] "observed distinct temporal evolution patterns of diseases" by analyzing disease prevalence across different age groups. Wang et al., [151] present a hierarchical topic model that infers topics at different levels of granularity of phenotypes and subphenotypes.

In another branch of research, Kulkarni et al., [82] investigate the presence of Post Traumatic Stress Disorder in textual reports on Reddit[1] using topic models. Our work, instead, is applied to Clinical Cases.

Neural Topic Modeling (NTM) [7, 47, 142, 114, 157, 68, 8, 82] is a research trend that combines topic modeling with Neural Language Models. A neural language model is a specific type of language model that relies on neural networks. NTM-based works claim to produce more interpretable topics than prior methods, yielding improvements in the state-of-the-art concerning topic coherence measure [68].

For instance, Top2vec [7] infers topic vectors by applying vector algebra over the neural vectors of words and documents embedded in the same vector space. BERTopic [63] is a topic model based on clusters of documents grouping vectors inferred through the Sentence-BERT framework [121].

Top2Vec and BERTopic perform best when applied to low-dimensional data, as both rely on HDBSCAN as their clustering technique. Consequently, they require dimensionality reduction on sentence vectors prior to clustering. In contrast, our approach employs a hierarchical clustering algorithm, which dispenses with dimensionality reduction and thus works directly with the original information.

Wu et al., [152] proposes a method that uses dependency matrices to model relationships between topics, enabling the construction of topic hierarchies. In contrast, our approach generates topic hierarchies through a simpler process, relying on the successive steps of the clustering algorithm.

Partially similar to our work, Qiu et al., [119] also present a method that combines SentenceBERT, a clustering algorithm and TF-IDF, but further prompt engineering to automatically label the topics.

[1] `https://www.reddit.com/`

1.5 Overview of the methodology and Research Contributions

Our research methodology addresses different kinds of semantics associated with textual constructions. Our approach is characterized by combining language models with a Topic Modeling method and Semantic Annotations to produce semantic representations to deal with inner and outer semantics – thereby contemplating our two Research Questions. We focus on technical, practical and theoretical aspects of language models working with clinical case reports to produce and/or to extract semantic meaning.

Our methodology consists of a suite of methods designed to extract and specify semantic meaning in machine-interpretable data models, which serve as a basis for distinct kinds of semantic representations.

Figure 1.1 depicts an overview of the four main contributions of this thesis, connecting the methods and models mentioned in section 1.3. The first three contributions (*Envoy*, *Harena Semantics* and the *Semantic Virtual Patient*) are presented in chapter 3, whereas the fourth – ABT– appears in chapter 4.

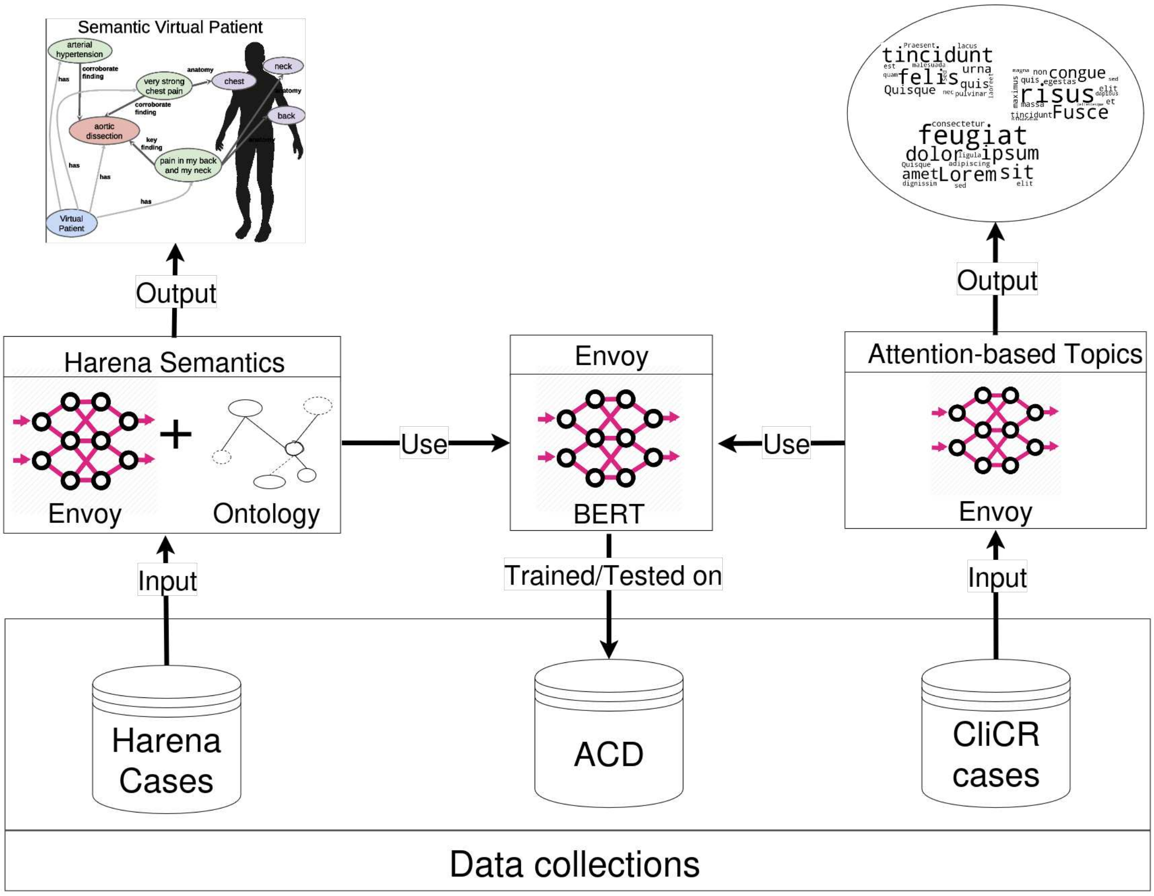


Figure 1.1: A high-level overview of the main thesis contributions.

The first contribution shown in the figure – our Envoy model – is independently used to construct the subsequent contributions. Envoy is a BERT-based neural network that trains a language model to identify biomedical concepts inside input texts. We trained

Envoy using our ACD data collection, which contains sentences pre-annotated with three types of entities: Anatomy, Chemical and Disease. After the training phase, the model can recognize these entities in any arbitrary text (i.e., sentences not observed by the model during its training).

The second contribution – our Harena Semantics – invokes the Envoy model to annotate the clinical cases drawn from the Harena clinical cases data repository. The annotation method combines the recognized biomedical entities with ontology concepts to produce the third contribution: the Semantic Virtual Patient, a semantic representation that models the clinical cases as a network of concepts linked to the Semantic Web.

The figure also portrays our fourth contribution – Attention Based Topics – a method for topic modeling using the Envoy model as a component to highlight the biomedical concepts inside collections of clinical cases. As shown in the Figure, the Envoy model is used to construct the two final representations: the Semantic Virtual Patient and the Attention-based Topics.

The subsections that follow give more details on the main contributions of this thesis.

1.5.1 Envoy: a language model for Named Entity Recognition

Our first contribution is a Named Entity Recognition (NER) method – which we call Envoy – to recognize and classify biomedical concepts inside the input textual constructions. The Envoy model uses a BERT-based language model as a base, enhanced by an extra fine-tuning layer to perform the NER to identify three types of entities: anatomy, chemical and disease. Our fine-tuning layer adapts the BERT model to perform a classification of tokens, instead of language modeling, which is the BERT objective in its original setting.

In order to perform the fine-tuning, we created a dataset called ACD from the concatenation of two other existing datasets: BC5CDR (whose sentences contain chemicals and diseases) and AnatEM (containing anatomies). As such, the ACD sentences contain words referring to the entities: anatomy, chemical and disease. During the fine-tuning phase, the model is fed by a total 21.223 ACD sentences, each of which containing up to 128 words. The model achieves a suitable performance after 2 up to 5 complete passes through the training dataset (i.e., training epochs).

The test set comprises 10.867 sentences from ACD. To identify the optimal setting, a validation protocol was employed to evaluate the model on the test set across a range of hyper-parameter configurations. In the best setting, the model achieved a F1 score around 85%, which is a result comparable to the state-of-the-art NER models at the time (2020).

1.5.2 Harena Semantics: a framework to embed semantic meaning into textual construction

This second contribution concerns the Harena Semantics framework, which leverages language models to help clinical studies. More specifically, it highlights the use of automatic semantic annotations to enhance knowledge about clinical case reports.

The Harena Semantics framework is geared towards clinical studies and is aimed to enable one to annotate semantic concepts directly onto the textual body of clinical case reports. The framework has two main components: (i) Versum, a markdown-based language which provides a regular schema for semantic annotations; and (ii) a semantic annotation assistant that automatically recognizes and annotates potential biomedical concepts. The assistant implements a Named Entity Recognition (NER) task combining concepts recognized both by the language model Envoy and by matches with ontology concepts. This hybrid approach benefits both from information formally structured within ontologies and from statistical information underlying the linguistic structure of the clinical case report. Besides, it enables one to superimpose annotations manually added by humans and the automatic annotations made by our NER algorithms.

Through this framework, the Harena system aligns with the principles of the Semantic Web, enabling domain experts to intuitively author clinical cases and annotate them with semantic concepts, thereby adding a higher level of abstraction to the digital content. The framework enables the system's users to cooperate on the gradual building of an interconnected network of Knowledge Bases.

1.5.3 Semantic Virtual Patient, a semantic representation for clinical cases

To get a pilot evaluation of the Harena Framework, we conducted a case study in the context of Medical Education, applying the framework to produce a specific semantic representation – which we call the Semantic Virtual Patient – that connects clinical case reports to knowledge bases maintained in the Semantic Web. Ultimately, the knowledge contained in the clinical case reports is part of an inference network (i.e., the Semantic Web), which connects digital objects of a multitude of types, such as symptoms and diseases.

The Semantic Virtual Patient has structured information (in an RDF graph format) that can be interpreted by machines, expanding the range of potential applications for clinical cases. For instance: (i) it facilitates the retrieval, reuse, and aggregation of cases or parts of cases – e.g., enabling queries such as "cases in which the patient experienced shortness of breath" or "cases where the ECG was fundamental to diagnose a heart disease"; (ii) data from cases can be used beyond the scope of medical education.

In a usability evaluation, the authors of [110] annotated relevant symptoms and indicated whether they are directly related to the clinical case (e.g., arterial hypertension and acute onset of chest pain), or key to the diagnosis (e.g., pain radiating to neck and back), or just distractors to the learner, misleading learners to a wrong direction (e.g., symmetric radial pulses is a specific sign but present in only one third of the patients). Results of this pilot evaluation suggest the framework's potential to engage users in the collaborative creation of semantic resources.

1.5.4 Attention-based Topics, a semantic representation for collections of text

Our fourth contribution, presented in chapter 4, is an approach to the topic modeling problem called Attention-based Topics (ABT). It relies on neural language models as computational abstractions to encode (i.e., computationally represent) and to reveal semantic aspects underlying the structure of a given set of textual data.

Although there is extensive work on topic modeling and neural language models, there is a lack of research to measure, analyze and validate the topics learned. To this end, ABT enables one to explore the themes covered by an input textual collection and analyze them as a hierarchical structure of topics.

We have conducted an evaluation using the CliCR corpus, a collection of clinical case reports described in natural language text. Our validation of ABT relies on observing the effects, in the output topics, of changing the language model – including our Envoy model in this validation. Preliminary results show that, although each language model sometimes produces completely different topics, none of them outperforms the others regarding the Topic Coherence (TC) metric. Nevertheless, the tree structure of the topic collection evidences some interesting patterns revealed by each language model.

1.6 Organization of the thesis

This thesis is structured as a collection of papers and their discussion, organized in 6 chapters, following the structure defined by the Institute of Computing, Unicamp.

Chapter 2 is a technical report that provides a bibliographic overview of Neural Networks, Deep Neural Networks, Transformers, as well as the history and development of the so-called Neural Language Models. It concludes with a section that critically analyzes language models addressing some interdisciplinary aspects [111]. It serves as background to understand and contextualize the rest of the thesis.

Chapter 3 corresponds to our paper "Harena Semantics: a framework to support Semantic Annotation in Citizen Science systems" published at the 15th International Conference on Health Informatics (HealthInf 2022) [110]. The paper describes a practical application of neural language models assisting semantic annotation of clinical case reports. It briefly presents the Envoy model and Harena Semantics, also introducing the concept of Virtual Patients - and thus our first three contributions.

Chapter 4 corresponds to our paper "Semantic Representations based on Topic Models" submitted for publication (under review) to a journal. The paper analyzes some characteristics of the ABT method and reports quantitative and qualitative evaluations. It describes our fourth main contributions, the ABT model. Our evaluation of ABT uses a series of BERT-related models, including our Envoy model.

Chapter 5 contains a brief discussion of some of our main findings. Additionally, we examine the difficulties of standardizing validation protocols to evaluate language models due to their holistic and interdisciplinary nature. Finally, we critically analyze neural language models as a technological tool that enables vast possibilities of applications on a diversity of domains.

Chapter 6 concludes the thesis, highlighting the contributions and some directions for future work.

Besides the papers in Chapters 2, 3, and 4, others were also published in the course of this thesis, directly related to this research. There follows a list of publications, including the ones that compose the thesis.

1. Fagner Leal Pantoja, André Santanchè, and Claudia Bauzer Medeiros. A bibliographic survey of Neural Language Models with applications in topic modeling and clinical studies. [111] *Technical report, Institute of Computing, Unicamp*, 2024. pages 1–26. (Chapter 2)

2. Fagner Leal Pantoja, Marco Antonio de Carvalho Filho, and André Santanchè. Harena Semantics: a framework to support Semantic Annotation in Citizen Science systems. [110] *Proceedings of the 15th International Conference on Health Informatics (HEALTHINF)*, 2022. pages 336–343. (Chapter 3)

3. Fagner Leal Pantoja, André Santanchè, and Claudia Bauzer Medeiros. Semantic Representations based on Topic Models. Submitted to *Journal of Universal Computer Science (J.UCS)*, 2025 (under review). pages 1–25. (Chapter 4)

4. Marcos Felipe de Menezes Mota, Fagner Leal Pantoja, Matheus Silva Mota, Tiago de Araujo Guerra Grangeia, Marco Antonio de Carvalho Filho, and André Santanchè. Analytical Design of Clinical Cases for Educational Games. [45] *Joint International Conference on Entertainment Computing and Serious Games (ICEC-JCSG)*, 2019. pages 353–365.

5. André Santanchè, Heitor Soares Mattosinho, Marcos Felipe De Menezes Mota, Fagner Leal Pantoja, Gabriel De Freitas Leite, Ana Claudia Tonelli, Fernando Salvetti Valente, Juliana De Castro Solano Martins, Sandro Queirós, Tiago De Araujo Guerra Grangeia, and Marco Antonio de Carvalho Filho. Virtual Patient Platform and Data Space for Sharing, Learning, Discussing, and Researching [125]. *IEEE International Conference on e-Science and Grid Computing (e-Science).* 2023. pages 1–10

Finally, we point out that the Harena system and semantics are being used in a series of research efforts and graduate studies conducted at the Institute of Computing, Unicamp, jointly with researchers from the Faculty of Medicine and the Faculty of Nursing, as well as researchers from the Netherlands and Portugal. This shows the potential of our contribution to other groups, giving margin to additional future perspectives.

Chapter 2

A bibliographic survey of Neural Language Models with applications in topic modeling and clinical studies

This text presents a literature review of Neural Language Models, which are deep neural networks to encode a given language. The scope of this review covers two main topics: (i) Transformers-based Neural Networks, established as state-of-the-art in addressing Natural Language Processing (NLP) problems and a suitable approach to train language models; and (ii) Neural Language Models that compress the statistical semantics of textual data into word vectors. These word vectors computationally represent the basic units of the language at hand.

In fact, obtaining a computational representation for textual constructs is a long-standing problem that has challenged diverse NLP approaches. We analyzed the usage of language models for Topic Modeling and for Semantic Annotation of Virtual Patients. The establishment of transformers-based language models opens up vast possibilities and perspectives on interdisciplinary topics. This text concludes with a critical analysis addressing issues regarding applications based on language models.

2.1 Introduction

Natural Language Processing (NLP) is a field of Computer Science whose goal is to convert human language into a representation that is interpretable by computers. It is an interdisciplinary research area that incorporates concepts from various other fields, such as Statistics and Linguistics.

Manning and Schütze [92] classify NLP methods into statistical and non-statistical approaches. Statistical approaches rely on patterns that commonly occur in a language, while non-statistical approaches focus on mapping and computationally implementing the rules that structure the language. The distinction between statistical and non-statistical approaches has roots grounded in the philosophical debate surrounding the perspectives of Rationalism and Empiricism [92].

In the epistemological realm [12], Rationalism claims the ideas of deductive reasoning are possible because they are innate, prior to all experience. In turn, Empiricism states that none of our ideas are innate, and the mind would be a blank tablet when we are born. Subsequently, Kant considered both the concept of active mind (from rationalism) and the role of sensations (from empiricism) as essentials in knowledge acquisition. In turn, Bertrand Russell "explicitly rejected the existence of innate ideas" [12]. The debate remains open and has led to the development of several philosophical schools.

In the field of Linguistics, the rationalist perspective is characterized by the belief in the existence of an innate language fixed in the human brain through genetic inheritance. Advocated by Noam Chomsky [92], rationalism has been crucial to the development of the theory of Formal Languages, which serves as the foundation for current programming languages. Formal languages constitute a special class of language that lacks ambiguity and, therefore, can be interpreted/compiled by computers. The ability to interpret a language in a non-ambiguous manner is essential for a computer to execute commands instructed by humans through a programming code [92].

In contrast to programming languages, natural languages are inherently ambiguous, since a word or phrase can have more than one meaning [61]. In natural language cases, the empiricist perspective assumes that, instead of pre-constructed linguistic structures, the human mind possesses generic operations of association, generalization, and pattern recognition. These cognitive abilities, combined with a rich sensory system, enable humans to learn detailed language structures. This hypothesis forms the basis of Machine Learning methods that use statistical models to recognize patterns and complex structures in a dataset. This statistical approach is grounded in the Information Theory developed by Claude Shannon [92].

Manning and Schütze [92] point out that "the difference between the approaches is not absolute but one of degree", as rationalism believes "the key parts of language are innate – hardwired in the brain at birth as part of the human genetic inheritance" whereas empiricism believes in an innate capacity to develop language through generalizations such that "a baby's brain begins with general operations for association, pattern recognition, and generalization, and that these can be applied to a rich sensory input available to the child to learn the detailed structure of natural language".

This philosophical debate remains an open question; however, its practical utility is valuable as it theoretically underpins various areas of computer science.

More recently, statistical approaches have advanced the state-of-the-art in various NLP tasks. This progress can be attributed to, among other factors: (1) advances in computational capacity; (2) recent deep neural network models capable of retaining significantly more information than previously proposed neural models; and (3) the development of more efficient techniques for handling the vast amount of information available on the Web.

The rest of this text is organized as follows: Sections 2 and 3 give some background on foundations of Neural Networks, Deep Neural Networks, Transformers, as well as the history and development of the so-called Neural Language Models. Sections 4 and 5 review related work of two case studies that involve neural networks and transformers: Topic Modeling and Semantic Annotations of virtual patients. Section 6 briefly discusses more recent work developed in research in Language Models. This is followed by a section that critically analyzes language models addressing some interdisciplinary aspects, finishing with concluding remarks.

2.2 Neural Networks

Several recent advances in the field of Natural Language Processing (NLP) are attributed to the mellowing of Deep Neural Network models, which are more sophisticated types of Artificial Neural Networks. This section describes some relevant issues in Neural Network architectures, followed in the subsequent section by Language Models in the context of such networks

2.2.1 Artificial Neural Networks

An artificial neural network — a computational abstraction inspired by the biological nervous system — is an interconnected network of artificial neurons organized in layers [81]. Typically, neural networks perform Supervised Learning, where the network receives successive sets of pre-labeled training samples and must infer the corresponding output for each input sample. For example, a neural network can be trained to recognize cancerous tumors in computed tomography images based on labeled images previously presented to the model. After this training phase, the neural network is capable of making inferences about new images that were not observed by the model during its network training.

Example - Sentiment Analysis through Neural Networks Sentiment Analysis through neural networks is an NLP task whose objective is to classify sentences based on their sentiment polarity $C = Positive, Negative, Neutral$. In the training phase, iteratively the network is fed by pairs of $\{sentence, label\}$ contained in the training set. In each training step, let s be the sentence to be classified, y the corresponding label, and h the output of the classification, representing the class inferred for s by the algorithm. The sentence s is represented by a feature vector $x = (x_1, x_2, ..., x_n)$. Let $\theta = (\theta_1, \theta_2, ..., \theta_m)$ be a vector of parameters (weights) of each neuron. Classification works as follows. Vector x is propagated through the network's layers, adjusting the parameters θ of each neuron based on their contribution to constructing the output y. Figure 2.1 illustrates a neural network classifying the sentence "I liked this movie". The neural network produces an output vector $o = (o_1, o_2, o_3)$ containing the algorithm's hypotheses regarding the probabilities of the sentence belonging to each of the possible classes in C, where the highest one is chosen as the algorithm's hypothesis $h = positive$.

Each neuron in the network has a transfer function (or activation function) $f(\Sigma)$ that operates on the weight parameter θ_i of the neuron and the input feature x_i, as illustrated in Figure 2.2. Several transfer functions can be employed, including the sigmoid function:

$$f(\Sigma) = \frac{1}{1 + e^{-\sum_{i=1}^{n} \theta_i x_i}} \tag{2.1}$$

During the training phase, each sentence is fed and processed in a training step. A training step of a neural network involves two main mechanisms [81]:

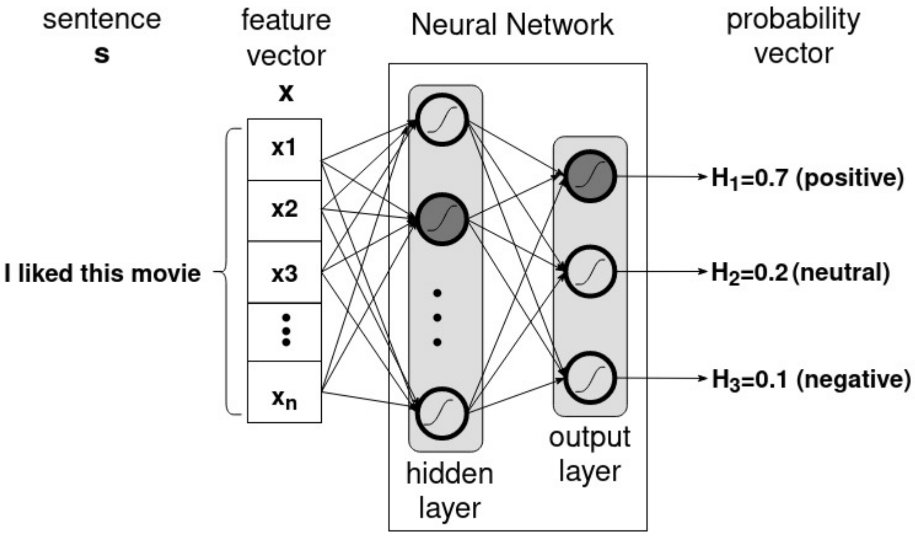


Figure 2.1: Neural network performing sentiment analysis. Adapted from [81].

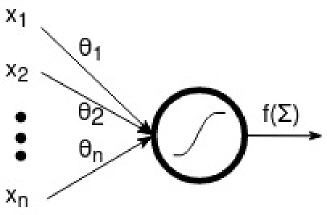


Figure 2.2: The functioning of the artificial neural unit.

- Forward Propagation: Propagates the sample sentence s through the neural network until it reaches the final layer. The final layer produces the hypothesis h containing the probability of s being classified as *positive*, *neutral* or *negative*. The difference between the hypothesis h inferred by the algorithm and the true class y annotated in the training set indicates the contribution (or responsibility) of each parameter θ_i to the error measured when classifying s.

- Backpropagation: Adjusts each parameter θ_i based on its contribution to the error calculated between the hypothesis h and the true class y annotated in the training set. The larger the contribution of the neuron, the greater the adjustment in its parameters should be. The magnitude of the adjustment in the parameters θ_i can be controlled by the hyperparameter[1] η, which typically has a value close to 0.1.

The algorithm completes one training epoch when it processes all the samples in the training set [81]. The number of epochs is also a hyperparameter. At the end of the training, the network parameters have been calibrated to solve the task for which it was trained.

An architecture with at least one hidden layer of neurons (as depicted in Figure 2.1) is

[1] Hyperparameters are variables that control the overall behavior of the network. Do not confuse with the term “parameter” that refers to the weights assigned to each neuron in the network.

known as a Multi-Layer Perceptron [81]. Other models implement different architectures, transfer functions, propagation mechanisms, etc. Deep neural networks models serve more robust architectures, including Recurrent Neural Networks, Convolutional Neural Networks, and the more recently introduced Transformers.

Recurrent Neural Networks

A recurrent neural network is suitable for solving problems with a sequential aspect [123], as observed in various NLP problems (e.g., sentiment analysis, Named Entity Recognition (NER), etc.) Recurrent neural networks leverage the inherent sequential aspect in textual constructs. Broadly speaking, the sequential aspect implies that each term w_i in a given sentence s depends on the preceding term w_{i-1}. For example, for a neural network handling the sentence "She is excellent at her role as a", the probability of the next word being "doctor" is higher than being "and" [143], as illustrated by Equation 2.2.

$$P(\texttt{doctor}|\texttt{She is excellent at her role as a}) \gg P(\texttt{and}|\texttt{She is excellent at her role as a}) \tag{2.2}$$

The Long-Short Term Memory (LSTM) model [65] is a sophisticated type of recurrent neural network that achieves considerable success in addressing NLP problems due to its mechanism for deciding which information to retain and which to discard [123]. Thus, the LSTM is capable of capturing contextual information from terms that are distant from the term being currently processed. However, the LSTM experiences performance degradation in scenarios involving very long sentences. This issue is known as Long-Term Dependencies, which arises when a word depends on words that are far apart in the sentence.

The LSTM has unidirectional contextual memory, restricting its usage to acquiring information solely from preceding terms (in the case of the left-to-right version of LSTM) or solely from subsequent terms (in the right-to-left LSTM) [46]. Additionally, as a specific type of recurrent neural network, the LSTM faces several challenges during the training phase, such as gradient explosion and gradient vanishing [112]. Furthermore, the sequential nature of the LSTM precludes the parallelization of the training process [147].

Despite these limitations, LSTM is often successfully employed in the Sequence Translation (or sequence-to-sequence) task, which aims to transform a given sequence s of arbitrary length into a corresponding sequence t of pre-defined fixed size n [136]. This fixed-size representation can be used to perform other NLP tasks (e.g., Language Modeling, Machine Translation, Speech Recognition, Question-Answering) through Transfer Learning techniques.

A significant advancement in this sequence-to-sequence task (and consequently in the NLP research field) was accomplished by the Transformer, a novel deep neural network model that overcomes the issue of long-term dependencies and provides a means to obtain bidirectional contextual memory. The next section focuses on the Transformer.

2.2.2 Transformers

The Transformer model, introduced in the paper "Attention is All You Need" [147], addresses the sequence-to-sequence problem more efficiently than the LSTM. The Transformer is a deep neural network model that captures the **context** through the Attention Mechanism. With this mechanism, the Transformer has advanced the state-of-the-art in NLP tasks that can be modeled as an instance of the sequence-to-sequence problem.

The Attention Mechanism The Transformer model overcomes the issue of long-term dependencies through the attention mechanism, which can capture global dependencies in the input sentence regardless of the distance between words [147]. Thus, this model can encode the context of the input sentence into the vector representation produced as output. The attention mechanism provides the Transformer with the ability to focus on words that are relevant to achieving the goal of the task being performed. There are various attention mechanisms [89, 58]. The Transformer specifically employs Self-Attention (or Intra-Attention), which involves a weighted sum of vectors resulting from successive linear transformations of the matrices Q, K, and V (query, key, and value, respectively) [80].

Matrices K and V comprehend the Neural Memory [60, 134][2] in which each row corresponds to a word (although, in other settings, it could be a character, a sentence, etc) within the training set, and the columns hold the features in form of distribution of weights (i.e., probabilities that a feature is relevant) [58] learned by the network. This distribution accounts for the context learned during the training phase. According to Geva et al., [60], "each key vector k_i captures a particular pattern (or set of patterns) in the input sequence, and that its corresponding value vector v_i represents the distribution of terms that follows said pattern".

The core of an attention mechanism is the computation over the matrices K, Q and V to infer an appropriate context representation. In each training step, the current word being processed (the query q) is matched against the K and V matrices to search for a key-value pair corresponding to the given query q. In this training process, the matrices undergo the series of operations illustrated in Figure 2.3, which is a visual representation of the Attention Function (Equation 2.3) to map a query and a set of key-value pairs into an output [147]. Here, we use the notation c to refer to this output vector.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \tag{2.3}$$

The attention mechanism operates over the K, Q, and V matrices in order to highlight the prominent patterns observed in the training dataset. The mechanism is about matching the context of a query q (the context of the given sentence) against the most similar context accumulated on the neural memory of $\{K, V\}$ whereupon the network gets information to accomplish the demanded task (e.g., the translation task in the original Transformer model [147]). As a byproduct of translation, the attention mechanism

[2] in fact, initially Geva et al., [60] showed the Feed-Forward sublayer can be seen as a neural memory, and [134] claimed a Feed-Forward sublayer can be seen as an attention layer, therefore we extend the concept of neural memory to refer also to the attention layer which holds the matrices K and V

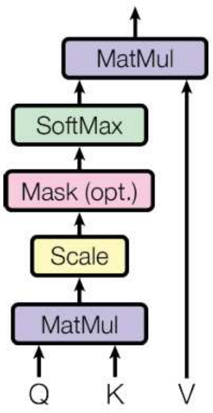


Figure 2.3: Operations performed by the layers of Self-Attention on the matrices Q, K, and V. Source: [147]

generates word alignments [89].

As an example, Figure 2.4 shows an alignment matrix derived from translating an input sentence (in the column) from English to German (in the rows). This matrix helps to visualize how the attention gets the correspondences between all words in a sentence: it highlights the "attention" each word in the target sentence pays to the words in the source sentence.

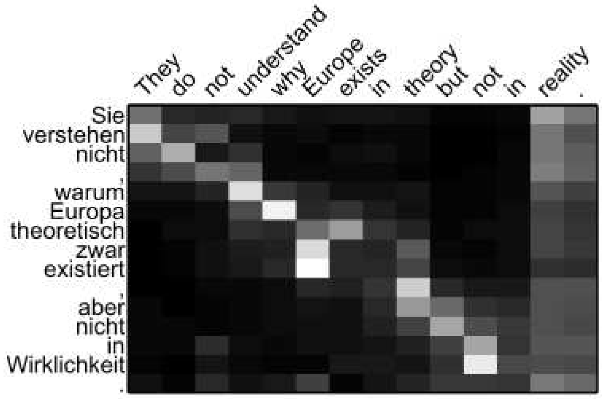


Figure 2.4: Word alignments derived from a translation task from English to German. Source: [89]

The original Transformer architecture (Figure 2.5) consists of two stacks of N layers of Encoders and Decoders. Subsequent researches proposed different architectures [144], for example encoder-only such as BERT [46], and decoder-only such as GPT [120].

Encoders An Encoder correlates each word $p_i \in s$ with all other words in the sentence s. Its result is a fixed-size vector representation c that encapsulates information regarding the sentence context.

The input sequence flows through the N layers of encoders. Each encoder receives Q, K, and V from the previous one. Within each encoder, there is a Multi-Head Attention unit (implementing Equation 2.3) and a Feed-Forward Network. The representation c

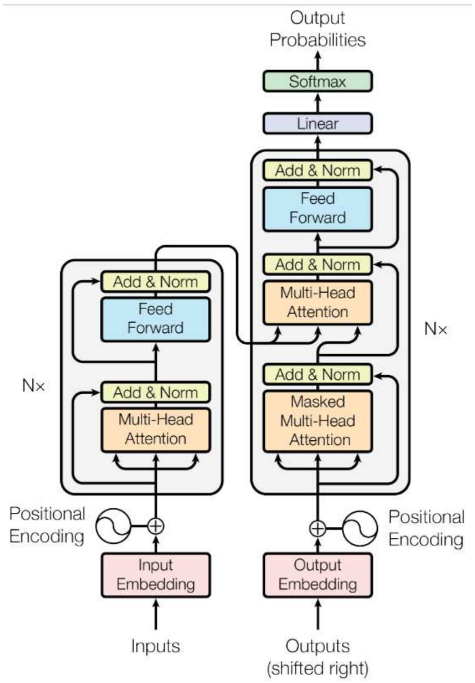


Figure 2.5: Transformer architecture. Source: [147]

generated by the last encoder is then sent to the decoder.

Decoders A Decoder uses the vector c received from the Encoder layers to generate the translated sequence t. The layers of the decoder are configured similarly to the encoder concerning the pair of multi-head attention and feed-forward layers. In addition to these, the decoder has a third layer called Masked Multi-Head Attention, responsible for ensuring that the decoder obtains information only from the preceding terms in the sequence t to preserve the auto-regressive property [147].

Multi-Head Attention In the so-called Multi-Head Attention, the Transformer projects the vectors Q, K, and V into multiple multi-heads with different learned linear projections. The resulting vectors are concatenated and projected again, resulting in the final vector representation, as depicted in Figure 2.6. Thus, the Transformer can pay attention to different representation subspaces at different positions [147], in addition to leveraging parallel computation.

Transfer Learning Through Transfer Learning techniques, a Transformer can be effectively applied to a range of NLP tasks, despite its initial design for sequence-to-sequence tasks. This versatility stems from the Transformer's capability to encode statistical patterns common to various NLP tasks into the matrices K, Q and V .

Fine-Tuning – a form of transfer learning – involves initial training on a source task τ_i followed by an adjustment of the learned parameters from τ_i to be applicable to solving a target task $\tau_j \neq \tau_i$ [67]. Through fine-tuning, the parameters of the pre-trained model are easily adaptable to other NLP tasks. There are several approaches that implement fine-tuning, such as Google BERT [46] and OpenAI GPT [120].

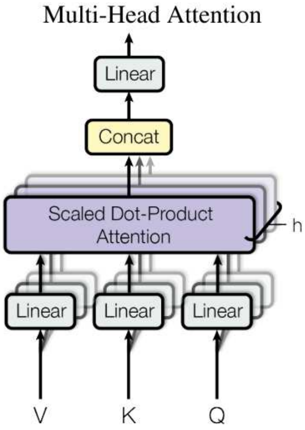


Figure 2.6: Schema of Multi-Head Attention. Source: [147]

The reuse of knowledge acquired in performing a generic task was first studied in computer vision research and, with the advent of the Transformer, has been extensively investigated in NLP [67]. One advancement in NLP was the realization that matrices K and V learned by the Transformer serve as a universal representation for textual constructs, capturing prominent syntactic and semantic aspects in the given textual construction. These aspects reveal idiosyncrasies embedded in the model during its training.

The neural memory of K and V embeds a Neural Language Model, which comprehends a new generation of Word Representations [146] (also known as Word Vectors, or Word Embeddings) for use in transfer learning across various work tasks. To our knowledge, the Transformer is currently the most robust and suitable model for pre-training these Language Models, which will be further explored in the next section.

2.3 Neural Language Models

This section discusses Language Models in the context of Neural Networks.

Given that each NLP task exhibits a distinct set of relevant features influencing its behavior and outcomes [81], it is crucial to carefully select an appropriate set of features to represent sentences given as input to the algorithm that has been selected. This stage of feature definition, known as Feature Engineering, is one of the major challenges in developing machine learning algorithms [81] and must be carefully conducted, as it has a significant impact on the algorithm's results. These features are typically stored in Feature Vectors.

In some cases, features are curated by experts, while in other scenarios features are collected through an automated process. Additionally, the criteria for feature selection are task-specific. For instance, in the case of Named Entity Recognition (NER), the features indicating whether a term should be considered a named entity encompass: term with the first letter capitalized, term preceded by a definite article, grammatical classes (e.g., noun, proper noun), prefixes and suffixes (e.g., diseases commonly ending with "it"), foreign words, term position within the sentence, term frequency in the training corpus, presence of other entities in the sentence, etc. [105].

Feature engineering often deals with the presence of redundant features, and also with features that are important in a given context but may not be as relevant in other contexts, among other obstacles. As a result, feature engineering is typically a costly and time-consuming process.

There is a recent trend on using Word Representations to alleviate the burden of effort spent in the feature engineering phase, as they provide efficient text representations that enhance the performance of classification algorithms applied to NLP tasks. Indeed, this is a rapidly growing research area that has been experiencing intense development due to recent advances in Deep Learning methods and the increased capacity for parallel processing.

2.3.1 Word Vector Representation

Also informally known as Word Embedding, word vector representation is, in summary, a numeric vector used to represent a unit of text (which can be a word, a document, a paragraph, etc.) given as input to NLP algorithms. This computational representation of textual data serves as an alternative to hand-designed feature vectors generated during the initial feature engineering phase in classical NLP pipelines.

One of the pioneer models of word representation is the One-Hot Vector, which represents each word as a vector of size $|W|$ (i.e., one vector dimension for each word in the vocabulary W) containing values of 1 in the dimension corresponding to the given word, while all other dimensions receive the value 0. Table 2.1 provides an example of a one-hot vector encoding for four words from a hypothetical vocabulary of size $|V| = 6$.

Table 2.1: One-hot representation.

	heart	drug	disease	therapy	kidney	chest
heart	1	0	0	0	0	0
drug	0	1	0	0	0	0
...			...			
disease	0	0	1	0	0	0
therapy	0	0	0	1	0	0

Unfortunately, the one-hot representation is inefficient due to its high dimensionality (one dimension for each word in the vocabulary) and its nature as a sparse matrix model, as each vector is composed of 0 values in most of its dimensions [146]. Distributional Representations are alternatives to mitigate these drawbacks.

Distributional Representations are generally based on matrices whose values are relative to the distribution of words in a specific context of that word. A context can be the entire document, a section of a document, other words nearby or around the word, among others. Typically, the context is defined in terms of window size and direction [146]. For instance, a context could be characterized by a window of the last 3 terms preceding each word, or a window of terms both to the left and right of each word.

In addition to the context, it is necessary to define the metrics to be used. One of the most common metrics is the co-occurrence of a pair of words, which can be recorded in a co-occurrence matrix of dimensions $|W| * |W|$. Another possibility is to use the widely adopted Term Frequency – Inverse Document Frequency (TF-IDF), which measures how discriminative a word is in a given collection of documents, i.e., how frequent a word is in a given document while being rare in other documents in the collection.

These distributional representations are vectors that, based on the distributional hypothesis [93], contain contextual information generated through word counting, so that words occurring in the same contexts tend to have similar meanings [13]. Such representations are suitable for tasks like classification and text retrieval; however, the challenge remains open to configure them appropriately for use in sequence labeling tasks (such as NER) [146].

An alternative approach involves generating representations through unsupervised training. These representations, commonly known as Distributed Representations, are typically induced using neural language models through training on a Language Modeling task. Unlike distributional representations that count the frequency of words in a given context, distributed representations are small – i.e., usually with a size between 50 and 1000 dimensions – and dense [146] – i.e., most dimensions contain values other than 0.

2.3.2 The Language Modeling Task

Language Modeling is an NLP task whose goal is to estimate the probability distribution of words in a given sentence. Traditionally, this estimation is achieved by predicting the next word based on the preceding words in the sentence [38] using the chain rule of probability [17]:

$$P(w_1, w_2, ..., w_{t-1}, w_t) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\dots P(w_t|w_1, w_2, \dots, w_{t-1}) \tag{2.4}$$

This Equation 2.4 predicts the probability that a word w will be used at position t following the previous words $w_1, w_2, ..., w_{t-1}$ in a given sentence. It is reasonable to expect that the term "beach" has a higher likelihood than the word "jail" of being the next word used in the sentence "I like to be in this…". Stated differently, a language model assesses the probability of a specified sentence existing within the modeled language [69].

Neural language modeling produces a set of weights (i.e., parameters) that are incrementally adjusted to minimize the loss during network training. The adjusted weights are used to induce word embeddings, whose similarity to other embeddings in the vocabulary indicates that these words occur in similar contexts in the given training set [93].

Neural language models implement a form of Unsupervised Learning and therefore do not require pre-labeled input data. The scarcity of manually annotated resources is a problem for Supervised Learning-based approaches, as is the case in many NLP algorithms, since generating pre-labeled data can be a costly and time-consuming task [120].

By using unsupervised learning, language models leverage the vast amount of unlabeled text available on the web, which is an immeasurable source of linguistic knowledge to be embedded in such models. Unsupervised training allows the model to automatically learn the latent features associated with syntactic and semantic properties.

There are many approaches to training language models, such as [18, 38, 98, 115, 23]. In a pioneering work, Bengio et al. [18] proposed a distributed representation model that effectively overcame the problem known as the Curse of Dimensionality. This problem was a barrier for training language models using neural networks; in earlier models, each word in the vocabulary was treated as a random variable which resulted in a network with a massive number of parameters, making training computationally infeasible due to the high computational cost involved.

Collobert and Weston [38] introduced a convolutional neural network to jointly train various NLP tasks through semi-supervised learning. The approach combines unsupervised learning, specifically language modeling, with supervised learning of other tasks in the pipeline, such as NER, part-of-speech tagging, etc. Therefore, this work demonstrated how to utilize embeddings learned in an unsupervised manner in the training of supervised tasks, rather than manually designed features.

Mikolov et al., [98] introduced Word2Vec, a language model with a simple and efficient neural network that has few hidden layers precisely to minimize the computational complexity caused by the non-linear hidden layer of deep neural network models. Word2Vec is provided in two similar versions: Continuous Bag-of-Words (CBOW), which predicts a target word based on the context words (4 words to the left and 4 words to the right) without considering the order of these words; and Continuous Skip-Gram, which predicts a target word based on another word within a specified range. Word2Vec captures different kinds of word similarity, going beyond basic syntax regularities. By employing simple algebraic operations on the word vectors, it is possible to observe that there is a similarity between the words "big" and "bigger" in the same way as between the words "small" and "smaller". Notably, an intriguing outcome arises from the operation $vector(King) - vector(Man) + vector(Woman) \approx vector(Queen)$.

Pennington et al., [115] introduced the GloVe model (Global Vectors for Word Representation), which generates global vectors in the sense that they contain statistical information regarding the entire training corpus. It employs a hybrid approach that combines co-occurrence matrices with distributed representations.

Bojanowski et al., [23] described FastText as an extension of the Continuous Skip-Gram model that incorporates the morphological structure of words. It represents each word based on its internal sub-terms (e.g., the vector for the word "where" is generated by summing the vectors of the n-grams <wh, whe, her, ere, re>). Thus, FastText addresses the Out-of-Vocabulary (OoV) problem by enabling the representation of words not present in the training set.

The models presented up to this point are considered static embeddings [54, 74, 138], since they assign a unique representation to each word in the vocabulary, thereby limiting their ability to handle polysemy (when a word has different meanings depending on the context in which it appears).

2.3.3 Context-aware Models

Recent work has considered the so-called context-sensitive word representations [54]. These contextualized language models have dynamic representation spaces, so that a specific term can have different representations depending on the specifics of the text in which it is found.

For instance, the Embeddings from Language Models (ELMo) model [116] derives context from the internal states of a bidirectional LSTM network that traverses both the right and left contexts of the current term. ELMo concatenates the internal states of the two layers to produce context-sensitive embeddings from both directions.

The Generative Pre-trained Transformer (GPT) [120] applies the Transformer architecture to a Language Modeling task to pre-train universal embeddings adaptable to various NLP tasks—such as Natural Language Inference, Question-Answering, Semantic Similarity, and Text Classification—through Fine-Tuning. Both ELMo and GPT use unidirectional language models to learn to represent context. The bidirectional context created by ELMo is a concatenation of two unidirectional contexts learned by different networks.

The BERT (Bidirectional Encoder Representations from Transformers) model [46] employs the Transformer architecture to train on the Masked Language Modeling task – a variation of the traditional Language Modeling seen in Equation 2.4. In this task, the model receives a training sentence with one of its terms hidden by a mask (e.g., "I like this [X] and ventilated room") and it must uncover the term hidden behind the mask. By accomplishing the task objective of language modeling, the network generates a Bidirectional Context that incorporates information statistically relating each term within the sentence to all neighboring terms present in the sentence. The language modeling captures various facets and features — e.g., long-term dependencies, hierarchical relationships, sentiments — that are relevant to task completion [67]. The awareness of the bidirectional context is a key aspect that enabled BERT to achieve the state of the art in 11 NLP tasks [46].

2.3.4 Neural Models for Sentences

The arrival of word representations has inspired other approaches to generate vector representations for larger text segments, such as phrases, sentences, paragraphs, and even entire documents.

Mikolov et al., [99] proposed a method for encoding idiomatic expressions, i.e., terms or phrases that have a meaning derived from the composition of their components, which is different from the meanings of the individual terms. For example, the expression "Boston Globe" represents the name of a newspaper, and its meaning is distinct from the simple combination of the individual terms "Boston" and "Globe". Additionally, the article describes some interesting properties of the Skip-Gram model, such as the Additive Property, which yields semantically coherent results. For example, in the vector space produced by Skip-Gram, the result of `vector(Russia) + vector(river)` is close to `vector(Volga River)`, while `vector(Germany) + vector(capital)` is close to `vector(Berlin)`. Such observations suggest that it is possible to obtain a non-obvious understanding of language by using Vector Arithmetic over word vectors [99].

Through a generalization of Skip-Gram, the SkipThought model [79] encodes a sentence by predicting the surrounding sentences. SkipThought implements an encoder-decoder model: the encoder maps words to a sentence vector, which is then utilized by the decoder to predict the surrounding sentences.

Inspired by CBOW [98], the Sent2Vec model [109] generates Sentence Vectors by averaging the vectors of the constituent n-grams in the input sentence. In fact, sentence vectors generated by averaging the vectors of all words in the sentence are quite robust models [76].

The InferSent model [39] employs a BiLSTM siamese network with a final layer of max-pooling. InterSent works as follows: the model is trained in a supervised fashion using the Stanford Natural Language Inference (SNLI) dataset [26], surpassing the results of unsupervised methods such as Skip-Thought. The SNLI dataset comprises 570,000 sentence pairs annotated with labels `contradiction`, `entailment`, and `neutral`. InferSent results suggest that Natural Language Inference (NLI) is a highly suitable task for sentence embeddings training.

The Sentence-BERT (SBERT) model [121] uses siamese and triplet networks (i.e., different networks with tied weights) to generate sentence embeddings. The training step of Sentence-BERT takes as input a pair of sentences and a similarity value between them. Initially, Sentence-BERT applies a pooling operation on the BERT embeddings to obtain a fixed-size representation (usually 768) for each sentence, as shown in Figure 2.7.

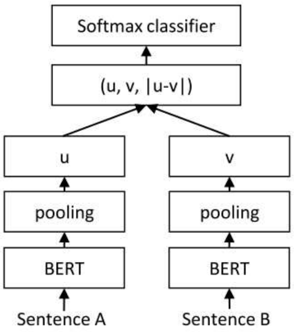


Figure 2.7: The architecture of the siamese network of SBERT. Source: [121].

2.4 Neural Networks and Topic Modeling

Topic Modeling [22, 62, 102, 68, 63] is an unsupervised Natural Language Processing challenge whose problem is discovering topics that represent an overview of the textual collection under analysis. A topic model explicitly represents the latent semantic structure [7] – or gist [62] – of a textual collection.

The Topics Model concept refers to a discrete probability distribution describing the connections between words, topics, and documents [27]. Topics are word combinations that demonstrate idiosyncrasies in the linguistic distribution of the corpus under analysis [25]. Topic models are explicit representations [22] that probabilistically associate documents with topics and topics with words.

Many methodologies address the Topic Modeling problem, such as Latent Semantic Analysis (LSA) based on linear algebra and its probabilistic version pLSA. Such methods apply dimensionality reduction to the documents represented in a Bag-of-Words format. Bag-of-words representations are adequate since, by hypothesis, word order is not a determining factor in these methods [22].

The Latent Dirichlet Allocation (LDA) [22] is a probabilistic model for discrete data collections, such as textual data. The LDA uses two complementary distributions: a topic over words distribution that describes the relationship between topics and words; and a second distribution that allocates topics to documents [27]. For LDA, a document is a random mixture of latent topics, which in turn are probability distributions over vocabulary words. The LDA is a dimensionality reduction technique [22] besides is one of the most robust and efficient methods for topic modeling.

The inferential machinery of LDA is capable of solving problems modeled in a multilevel structure, for example, Collaborative Filtering, in which the dataset comprises a collection of users, which in turn has a list of preferred objects. In this case, users and objects are analogous to documents and words in the document, respectively [22]. Therefore, LDA applies to problems beyond the textual domain.

Recent works [7, 47, 142, 114, 157, 68, 8, 82] investigate Neural Topic Modeling (NTM), a current research trend that combines topic modeling with Neural Language Models. Indeed, this is a rapidly growing area of research. [142, 157] discourse on the similarity between the topics produced via NTM and other traditional topic models, such as LDA. NTM-based works claim to produce more interpretable topics than prior methods, yielding improvements in the state-of-the-art concerning topic coherence measure [68].

Top2vec [7] infers topic vectors by applying vector algebra over the neural vectors of words and documents embedded in the same vector space. Each topic corresponds to a centroid of a cluster of documents and takes the closer word vectors as its most representative words. The approach infers the optimal number of topics through the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm.

The TopicBERT [8] topic model recognizes topics combining Transformers, community detection techniques in graphs, and named entity recognition.

[63] presented the BERTopic model based on clusters of documents grouping vectors inferred through the Sentence-BERT framework [121]. Subsequently, the approach assigns the clusters to a c-TF-IDF matrix that indicates the most representative words of each topic. c-TF-IDF is a variation of the classic TF-IDF algorithm, taking clusters as the unit of analysis instead of documents.

Although many research efforts use hierarchical clustering algorithms – such as HDBSCAN – their focus is not on the hierarchy, as they apply dimensionality reduction that interferes with hierarchical analysis. We provide an analysis of the inferred hierarchy of topics and its relationship with the arrival of Language Models based on the Transformer architecture and its attention mechanism.

Related works have studied hierarchical structures of topics [78, 71, 156]. In particular, [78] showed important aspects of the hierarchy of topics and proposed the recursive Chinese restaurant process (rCRP) method to generate hierarchical topic structures with unbounded depth and width. Their study analyzed metrics to characterize Hierarchical Structures to organize topics, such as Hierarchical Affinity and Topic Specialization.

2.5 Neural Networks to Produce Semantically Annotated Virtual Patients

This section explores the use of Neural Networks to help clinical studies. More specifically, it highlights the use of automatic semantic annotations to enhance knowledge about Virtual Patients.

2.5.1 Semantic Annotation of Virtual Patients

In this section we give a background about Semantic Web and Named Entity Recognition (NER) approaches to perform semantic annotation on free-texts.

These topics are the foundations of our method to Semantic Annotation described on Chapter 3 of this thesis. Our paper “Harena Semantics: a framework to support Semantic Annotation in Citizen Science systems” [110] published at the 15^{th} International Conference on Health Informatics 2022 (HealthInf) is based on that chapter. The paper is intended to enable easy semantic annotation over natural language sequences contained within Virtual Patients data.

Semantic Annotations

Semantic annotation [73] is the process to connect textual sentences to the resources of the Semantic Web, towards linking the natural language texts to networks of ontologies. The Semantic Web is a network designed to be manipulated both by humans and computer agents. As envisioned by Tim Berners-Lee et al., [19], it is an extension of the current Word Wide Web (not a separate one), which structures the information in a format that better enables computers to process their content in a meaningful way.

As far as we know, there is not a standardized, established definition of the approaches used for semantic annotation. Some works define them as Concept Normalization [97, 50], Entity Linking [14], Entity Typing [36], and so on.

Usually, the semantic annotation process involves some kind of Named Entity Recognition procedure. Named Entity Recognition (NER) is a Natural Language Processing (NLP) task to identify and classify entity types, such as People, Organization and Location. In the biomedical domain, research works focus on Gene, Protein, Disease, Chemical, Anatomy, etc.

There are many approaches to implement NER tasks. Recent works using statistical approaches have leveraged the NER state-of-the-art by using Deep Neural Networks to learn Language Models [38, 46]. These Language Models encode syntactic and semantic information in Word Vectors in such a way that those vectors with similar meanings have similar representations. One can build NER methods by feeding the word vectors as feature vectors to a downstream algorithm that decides if it should or not tag a given term as a named entity.

There are works [6, 84, 4] specializing BERT vectors to the biomedical and clinical domains. [84] introduce BioBERT, a BERT-based model specialized in biomedical language. BioBERT is pre-trained on large-scale biomedical corpora composed of PubMed abstracts and PMC full-text articles. BioBERT outperformed the state-of-the-art models in a bunch of experiments over NER tasks. BioBERT trained several models to recognize different named entities (e.g., one model to handle diseases, another one to deal with proteins, and so on). We based our NER implementation on the BioBERT model, extending it to recognize multiple entities (anatomy, chemicals, and disease) in a single model.

Other initiatives also address NER by extending language models, refining word vectors, or fine-tuning the model to specialized datasets [14, 91, 97]. BERN [77] is a neural biomedical multi-type NER tool based on the BioBERT model. The model uses a separate NER model for each entity type, and then combines the results using decision rules to handle overlapping entities (i.e., when the models tag a term with different tags, for instance, two models tagging the term "androgen" as gene and chemical). Differently, our NER model does not need decision rules, since the language model itself contains the statistical context to decide which entity is the most likely to occur. It is due to our model is trained on a dataset containing all the entities of interest, avoiding training one model to each entity. The model is based only on a Statistical approach (Empiricist) since does not rely on pre-defined rules.

Virtual Patients authoring

In the context of Medical Education, a clinical case comprises a medical narrative of situations occurring in real clinical environments. Lecturers use clinical cases as pedagogical resources to teach clinical practices to medical students. According to Šuster [135] "a case report is a detailed description of a clinical case that focuses on rare diseases, unusual presentation of common conditions and novel treatment methods".

There is a wide spectrum of strategies to simulate patients for students' training [40, 57, 29, 137, 34]. The adopted strategy depends on the available resources, the goal expected from the training, the level of structure in the data and the desired expressiveness of the clinical narrative of simulation.

Virtual Patients (VP) [40] are designed to present scenarios and narratives of a Clinical Case, guided by computers. They represent the Clinical Case in a graph of states affording structured guidance.

The adoption of Virtual Patients enables interesting research. For example, Hege et al. present a tool to foster the acquisition of clinical reasoning skills through Virtual Patients and Concept Maps [64].

The OpenLabyrinth [29] is a system for authoring virtual patients. The OLabX project [43] – an extension of OpenLabyrinth – uses mEducator schema to discover, retrieve, share, and reuse medical educational resources.

Our approach to Virtual Patients authoring differs from related work, as it departs from a markdown-derived language, apt for human writing, reading, and annotation, combining it with automatically generated superimposed annotations produced via the semantic annotation process. In the paper [110] we analyzed semantic annotation within the Harena system specifically designed to manage a collection of Virtual Patients.

2.6 More recent NLP research

We have witnessed a heated research focus around the release of ChatGPT. The ChatGPT system uses the models of the GPT family – GPT-3 [120], GPT-4 [3], etc – to perform text generation in a dialog style [88]. GPT is a decoder-only language model fine-tuned through Reinforcement Learning from Human Feedback [144].

Recent research focuses on increasing the network size by developing Large Language Models containing billions of parameters [144]. GPT-3 contains 175 billions of parameters, Llama 65 B [145], Chinchilla 70B, PaLM 540B, BLOOM 176B [126]. Studies [66, 145] have found not necessarily the larger model results to best performance at inference time, but there is a trade-off between the model and dataset sizes such that the best model would be a smaller model trained longer (i.e., on more samples).

A recent trend is to incorporate language models into larger systems. For example, DALL-E[3] [20] is a system to generate images from text prompts. Sora[4] is a system capable of generating high-fidelity videos from input text based on diffusion models [113]. Gemini [139] is a multimodal model trained on different modalities of data such as image, audio, and video. However, some systems and applications lack academic references to describe the techniques employed and details of integrating the theoretical models in the systems. Usually, the implementation details are referred to in web pages. Many of the recent research on language models are described in technical reports uploaded to repositories which do not account for peer-review processes. It hampers the scrutiny of the real advancements in this research area and the establishment of a reliable ground of scientific validation.

There are works focusing on demonstrating the linguistic capabilities of language models. The results of the study by Tenney et al. [141], suggest that the initial layers of BERT networks concentrate on basic syntactic information, while the higher layers focus on high-level semantic information. Ettinger [55] applies tests based on psycholinguistic studies to assess the language models' ability to capture linguistic features. The results suggest that probability distributions are sensitive to linguistic distinctions, such as semantic roles, pragmatic reasoning, common sense, etc. These aspects would be evidence of idiosyncrasies embedded in the model during its training.

Diverse surveys aim to review the methods and techniques employed on the last released large language models. Other surveys [154, 90, 33] address the evaluation of language models. The work of [90] categorizes a bunch of methods for the evaluation of language models in terms of faithful explainability.

There are works analyzing the language capabilities of language models and comparing

[3] https://openai.com/dall-e-3

[4] https://openai.com/sora#research

their procedures with the functioning of the human brain. For example, Sejnowski [129] hypothesized that the intelligence of language models is a mirror that reflects the intelligence of the person using such a system. The paper of [49] claims that, although GPT models lack mechanisms of consciousness from a cognitive science perspective, they have already passed the Turing test and therefore can successfully imitate human language capabilities.

However, there is neither consensus nor definition about which type of analysis should be employed to evaluate the language models. This state of affairs is perhaps not surprising, since neural networks are examples of Complex Systems [128] and therefore, are essentially holistic and interdisciplinary. Thus, neural networks for language models would be machines "as complex as the systems they model and therefore they will be equally difficult to analyse" [128]. This situation resembles the difficulty of validating models based on a relativist, holistic philosophy of science [12]. By such approach the "The criterion of practical use has taken the place of formal rigor [...] validation becomes a semiformal, conversational process" than "a matter of formal accuracy". Therefore, the emergence of neural language models demand the research and development of new validation methods to assess their capacities, considering their holistic and interdisciplinary nature.

2.7 Critical Analysis

Up to this point, we concentrated on the Transformer and its capability to handle global context in input sequences. This advantage enables the Transformer to successfully train Language Models, leveraging huge amounts of data. However, there are open issues: (i) the difficulty to interpret [87] the inner workings of neural networks, and consequently, of transformers; (ii) the distributed nature [128] of Language Models hinders control of the patterns represented (for example, societal biases [130]).

The Attention Mechanism and the Transformer model have elevated the state-of-the-art in various NLP tasks that have long been challenges in this research area. Such improvements suggest that a new level of language understanding can be achieved by using attention-based models to capture the patterns that structure textual sentences. The invention of attention-based language models can be stated as a revolution in the NLP research field, symbolizing a significant technological leap forward in this field. However, there are different perspectives regarding the actual advancements achieved in a scientific context beyond computer science.

In a position paper, Bender and Koller [16] argue that language models are not, a priori, capable of understanding the **real meaning** of processed texts, as they are trained only on textual forms (i.e., the linguistic signal). This is based on the definition of meaning as the relationship between a linguistic form and an intention of communication. This would imply that there is a portion of meaning attributed to extra-textual information not present in the training set. The authors draw attention to the imprudent use of certain terms (such as understanding, comprehension, etc.) as academic terminology when reporting research results in the field.

On the other hand, Sahlgren and Carlsson [124] argue that if meaning produces effects on form, then a language model should at least be able to observe and learn these effects.

This debate addresses issues that have historically been studied in various research areas. Therefore, it is necessary to analyze the results obtained by this so-called NLP revolution with caution, as it raises expectations and interests from different actors in society—companies, states, political groups, and even the public at large.

For example, AlphaFold [5] is a Google project aimed at predicting the 3D structure of proteins – an essential challenge within Biology [30] – using deep neural networks. These proteins could be applied in projects for new therapies for infectious diseases, less allergenic foods, and also for potential malicious applications—such as the development of toxic proteins as biological weapons [149]. There are also examples of work in Law [32, 53] or in Geosciences and Petroleum Engineering [94]. The work by McGuffie and Newhouse [96], warns about potential language models trained to generate content based on radical ideologies (white supremacy, anti-Semitism, etc.) with the aim of disseminating extremist thoughts.

Given that neural language models are trained to recognize prominent patterns in the training set, it is expected that they capture — and consequently reproduce — racist, classist, sexist, misogynistic, homophobic, xenophobic biases, and other patterns historically perpetuated in society. For example, the results from Silva et al. [130] indicate that Transformers exhibit a statistically significant tendency to infer female and Afro-American subjects in contexts of emotive words, thus highlighting an embedded racial bias in these opaque-box models.

The existence of bias in neural models raises concerns, in particular in sensitive situations such as the development of methods for automatic student evaluation [95], the classification of patients with Opioid Use Disorder using longitudinal health data [56], or the exploration of the connection between cannabis use and depression disorder through Twitter post content [155], and so on. Research applying language models to study such complex and interdisciplinary topics should take into account the knowledge and perspectives of other fields to avoid oversimplifications and the establishment of spurious correlations.

In this sense, it is crucial to address ethical issues in Artificial Intelligence. This includes implementing best practices for developing open-source Neural Language Models. This is essential to ensure individual freedom in an era where we store a multitude of personal information on the Internet. It also helps prevent the complexity [9] of this new technology from being used to mask biases and interests. Open-source code allows, to some extent, auditability of inferred classifications and coded rules, enabling the verification of results.

The introduction of machine learning algorithms everywhere, and NLP in particular, can impact the way we work, relate, learn, and develop. Therefore, there is a need for education at all levels aimed at teaching people how to use, understand, develop, and consume these tools in a healthy manner. Considering the breadth of the impacts that neural language models can have on human life, a pedagogical project is needed to guide towards a sustainable and ethical use of neural language models that also serve to address real and widely discussed societal problems, rather than solely serving the economic and market interests of the few who hold and dominate this technology.

2.8 Conclusion

The research efforts over the past decades on neural networks have led to the establishment of the Transformer model. A key aspect of the Transformer is its awareness of the global context within the training collection by “paying attention” to all the terms surrounding the current term being processed (not only to the n-grams to left or right as in the previously proposed approaches).

In this report, we reviewed the Transformer as a suitable approach to train Language Models that efficiently compress the global context of text collections by encoding the statistical patterns prominent in the language. Diverse long-standing NLP problems were suddenly solved by approaches applying transformers-based language models. This demonstrates and corroborates the impressive performance of Neural Language Models. Here we reviewed two of these tasks: Named Entity Recognition and Topic Modeling.

The arrival of neural language models shed many paths of discovery and improvements in processes, studies, and scientific discovery. Despite their undeniable success, the complexity and innovation of such technological tools raises concerns about: (i) evaluation of such models; and (ii) their ethical applicability – as voted by the November 2021 Unesco Assembly [10]

Chapter 3

Harena Semantics: a framework to support Semantic Annotation in Citizen Science systems

We propose a new approach to support human agents to annotate semantic concepts in free-text sentences in the biomedical domain. Using our markdown-derived language called Versum, authors can easily annotate relevant terms while producing content for Citizen Science systems. Besides, an embedded Automatic Annotation Mechanism suggests semantic concepts for the author. It implements a Named Entity Recognition task using a hybrid approach: (1) a Transformer-based Deep Neural Network and (2) an Ontology-based method. We conducted a case study running over content produced in the Harena e-learning system, which intends to teach Clinical Reasoning to students using Clinical Cases. Results of this pilot evaluation suggest the potential of Harena Semantics to engage volunteers in the production of semantic, agent-centered resources on crowdsourcing systems.

3.1 Introduction

In clinical learning environments, professors use **Clinical Cases** as pedagogical resources to teach students to solve problems and, consequently, to develop their clinical reasoning capacities. Usually, Clinical Cases have fictional narratives inspired by real situations interconnected in a network of unanticipated events commonly occurring in a clinical environment. This complex information comprises a valuable health knowledge source.

Despite the potential that Clinical Cases have to become an unprecedented Knowledge Base, there are open challenges concerning: (1) how to handle and process free-text information contained in the case narrative; and (2) how to integrate and interrelate complex information fragmented across a plethora of documents on the Web.

Envisaging these challenges, we propose Harena Semantics to construct and curate Clinical Cases delivering two main contributions:

- **Versum**: a markdown-based script language that enables authors to annotate semantic concepts inside natural language texts. Via Versum, the semantic annotation may be done manually by a human agent (e.g., professors, learners), automatically by some computer-assisted method, or by a mixture of these methods.

- An Annotation Mechanism which automatically recognizes relevant clinical concepts within a given sentence following a hybrid approach composed of two independent algorithms: (1) a **Transformer-based** Named Entity Recognition (NER) task implemented as a Deep Neural Network [147] and (2) an **Ontology-based** NER to link terms from free-text sentences to ontology-related concepts, formally defined as knowledge graph, which comprises a network of interconnected semantic resources.

We conducted a case study of our framework running over the Harena[1] system [45], an e-learning environment, based on cases resolution, which is used as a supporting pedagogical tool in Emergency Medicine courses. Harena represents Clinical Cases in a Virtual Patient format [40].

The Harena environment comprises a Knowledge Base of clinical cases, besides two complementary Web-based modes: (1) Interface Author, to enable one to create and curate artificial Clinical Cases inspired by real-life clinical encounters; and (2) Interface Player, to execute the simulations of the Clinical Cases narratives.

Through our approach, the Harena system adheres to the Semantic Web, enabling authors to produce semantic annotations over free-text content of their Clinical Cases. The authors of Clinical Cases (i.e., the Harena system users) are also **Citizen Scientists** who embed medical knowledge in Clinical Case narratives.

Citizen Science projects promote the collection and analysis of scientific data by members of the general public and professional scientists. By adhering to the Semantic Web, Citizen Science systems enable content producers to add a higher level of abstraction to the crowdsourced information. Harena Semantics enables volunteer users to produce agent-centered resources (e.g., Clinical Cases and beyond) and therefore cooperate on the gradual building of an interconnected network of Knowledge Bases. Agent-centered resource engagement is one of the three types of engagement in citizen science projects [72] which have the potential to increase engagement in the early stages of training in a volunteer learning scenario. Preliminary results of this study reinforce the claim about the need for mechanisms to engage users in the production of agent-centered resources.

The remaining of this paper is organized as follows: Section 3.2 gives some background foundations and related work; Section 3.3 describes the Harena Semantics framework and some results of the NER task evaluation; Section 3.4 presents a case study of our approach running over the Harena system; Section 3.5 presents our concluding remarks.

3.2 Foundations and Related Work

The scope of this review section is twofold. First, we give a background about Clinical Cases, Virtual Patients, and Semantic Web applied to the Clinical Reasoning research field. Lastly, we briefly present some concepts of Natural Language Processing (NLP), Named Entity Recognition (NER), and the just arrived Word Embeddings.

[1] `https://jacinto.harena.org/`

3.2.1 Virtual Patient

In the health context, there is a wide spectrum of strategies to simulate patients for students' training [40]. The adopted strategy depends on the available resources, the goal expected from the training, the level of structure in the data and the desired expressiveness of the clinical narrative of simulation.

Virtual Patients (VP) are designed to present scenarios and narratives of a Clinical Case, guided by computers. They represent the Clinical Case in a graph of states affording structured guidance [40].

By taking advantage of Semantic Web abstraction, OLabX (extended OpenLabyrinth) uses mEducator schema to discover, retrieve, share, and reuse medical educational resources [43]. Hege et al. present a tool to foster the acquisition of clinical reasoning skills through Virtual Patients and Concept Maps [64].

Our approach differs from related work, as it departs from a markdown-derived language, apt for human writing, reading, and annotation, combining it with automatically generated superimposed annotations.

3.2.2 Named Entity Recognition

Named Entity Recognition (NER) is a Natural Language Processing (NLP) task to identify and classify entity types, such as `People`, `Organization` and `Location`, . In the biomedical domain, research works focus on `Gene`, `Protein`, `Disease`, `Chemical`, `Anatomy`, etc.

There are many approaches to implement NER tasks. Recent works using statistical approaches have leveraged the NER state-of-the-art by using Deep Neural Networks to learn Word Embeddings [38, 46]. These neural language models encode syntactic and semantic information in vectors known as embeds in such a way that those embeds with similar meanings have similar representations. They feed the algorithm that decides if it should tag a term within the given sentence as a named entity.

Google released BERT (Bidirectional Encoder Representations from Transformers) [46], an implementation of the Transformer architecture introduced on the paper "Attention is all you need" [147]. BERT is pre-trained on the Masked Language Model (MLM), an unsupervised NLP task whose objective is to predict the hidden word in a given input sentence. MLM is an expensive task since it requires millions of sample sentences. The pre-training phase produces the Word Embeddings as a by-product of the task objective. One can easily reuse BERT embeddings by Fine-Tuning (a Transfer Learning technique) them in a downstream task (e.g., NER, Question-Answering, Natural Language Inference). Generally, Fine-Tuning is an inexpensive method since it requires a small labeled dataset.

BERT produces context-aware Word Embeddings through the Attention Mechanism, which detects the most representative parts in the whole sentence. The Attention Mechanism is a procedure to capture the sentence context based on the statistical relationship between the current word and every other word in the input sentence, providing a bidirectional context – i.e., interpreting the sentence considering the previous context (left-right direction) and posterior context (right-left direction). This bidirectional context leveraged the state-of-the-art of NER methods since the previous works were capable of dealing with the context in just one direction [46]. The NLP community sees the rising of transformer-based Word Embeddings as a revolution in this research field.

There are works [6, 84, 4] specializing BERT-embeddings to the Biomedical and clinical domains. [84] introduce BioBERT, a BERT-based model specialized in biomedical language. BioBERT is pre-trained on large-scale biomedical corpora composed of PubMed abstracts and PMC full-text articles. BioBERT outperformed the state-of-the-art models in a bunch of experiments over NER tasks. BioBERT trained several models to recognize different named entities (e.g., one model to handle diseases, another one to deal with proteins, and so on). We based our NER implementation on the BioBERT model, extending it to recognize multiple entities (anatomy, chemicals, and disease) in a single model.

Recent works attempted to join Word Embeddings and Semantic Web ideas. We intend to contribute to this endeavor by investigating how to superimpose NER annotations produced by these distinct, although related, research areas. According to our literature search, there is not a standardized, established definition of ontology-based NER methods. Some works define them as Concept Normalization [97, 50], Entity Linking [14], Entity Typing [36], and so on.

[77] present BERN, a neural biomedical multi-type NER tool based on the BioBERT model. BERN is equipped with probability-based decision rules to treat overlapping entities (polysemy – for instance, one can tag androgen as gene or chemical) and synonyms (i.e., terms described by multiple names). BERN normalizes the recognized entities assigning an ID (linking to controlled vocabularies) to each recognized entity.

Other initiatives also address NER extending BERT, refining the embeddings, or fine-tunning with specialized datasets [14, 91, 97].

3.3 Harena Semantics

This section presents Harena Semantics, a framework consisting of two complementary components: (1) a markdown-based language called Versum to enable Citizen Scientists to create and curate Clinical Cases adherents to Semantic Web; and (2) a hybrid approach to perform a Named Entity Recognition (NER) task to annotate Clinical Cases with semantic concept labels.

3.3.1 Versum

Versum enables one to add semantic structure into the free-text content of Clinical Cases, aiming to allow easy integration of semantic annotations into clinical narratives. By making explicit the semantic of Clinical Cases, Versum creates pedagogical resources adherent to the Semantic Web while providing a step forward to a machine-interpretable representation of the natural human language.

In a previous research paper, Menezes et al. developed the first version of Versum following the Narrative Design approach, which provides elements to enable scenario building and flow control of narratives.

In this paper, we release the Annotation Mechanism as an improvement feature of Versum. The process of semantic annotation using the Versum syntax is straightforward through a predefined set of reserved markups to add high-level structured information to the Clinical Case.

Using Versum markups, one can annotate a text fragment as a semantic concept enclosing it between the curly braces `{ }` followed by the concept label between parenthesis, e.g., `{heart attack}(disease)`. Moreover, one can link a free-text entity to Knowledge Bases – e.g., ontologies, controlled vocabularies, taxonomies, thesaurus – through the `namespaces` markup.

Figure 3.1 shows three stages of a clinical case being annotated using the Versum.

1. A Clinical Case narrative in free-text format. This Clinical Case was authored in the Harena system.

2. The Semantic Clinical Case produced from the original free-text Clinical Case. By annotating with Versum, the narrative becomes more structured and semantically enriched. These annotations could be manually made by a human agent or by an automatic process (like the Automatic Annotation Mechanism depicted in the next Section 3.3.2).

3. A visual representation of the Semantic Clinical Case highlighting the semantic concepts annotated.

3.3.2 Automatic Annotation Mechanism

Harena Semantics provides a mechanism to automatically annotate the concepts within the Clinical Case through a hybrid approach:

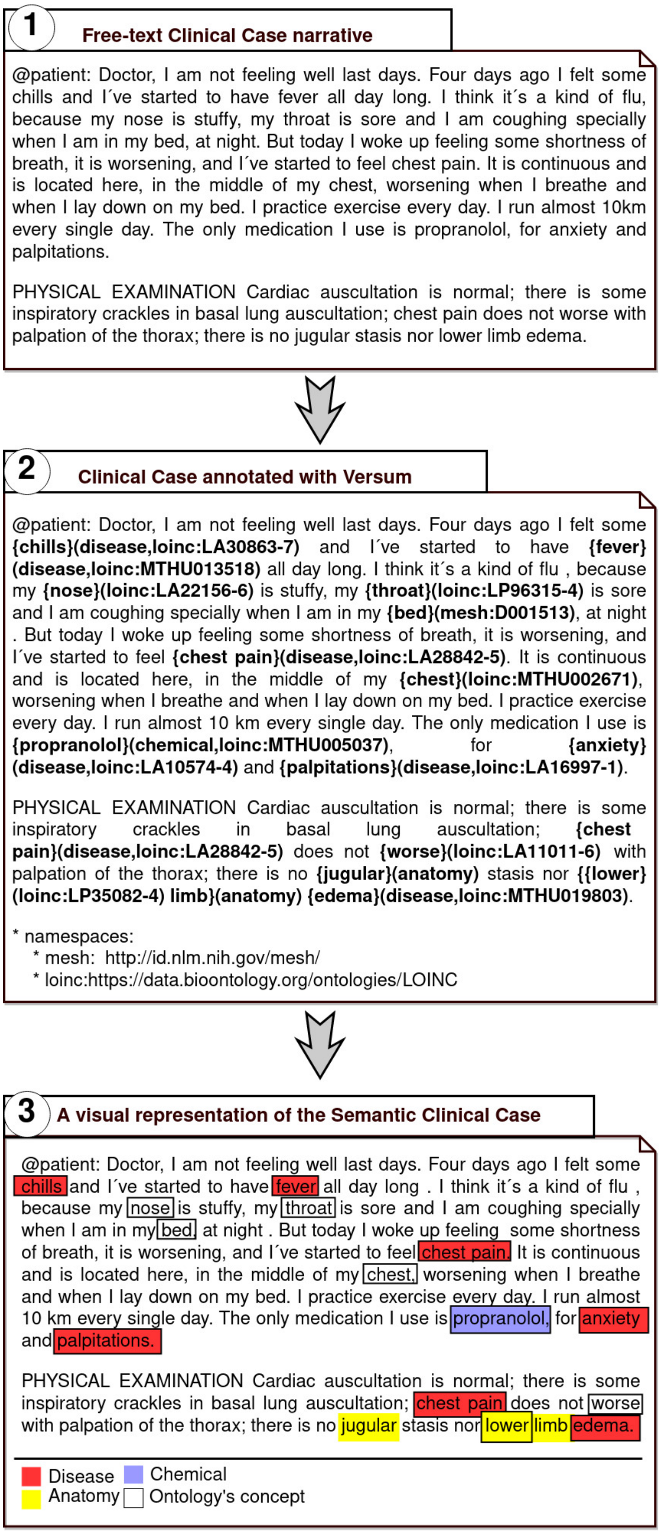
1. Transformer-based Named Entity Recognition (NER) task to assign labels to clinical terms within a given sentence. This method is based on the Transformer architecture, a Deep Neural Network capable of capturing linguistic features based on statistical inferences.

2. Ontology-based Named Entity Recognition task to link from free-text terms to concepts formally defined on biomedical ontologies.

As output, the Automatic Annotation mechanism produces the class labels to terms of the sentence given as input, as depicted by Figure 3.2. The labels may be (1) ontology concepts, (2) named entities, or (3) a combination of them.

Transformer-based Named Entity Recognition

We developed a Transformer-based NER that attends to the contextual information – i.e., syntactic and semantic aspects related to the context of a sentence – to decide if it should label or not a given term as a named entity.



@patient: Doctor, I am not feeling well last days. Four days ago I felt some chills and I´ve started to have fever all day long. I think it´s a kind of flu, because my nose is stuffy, my throat is sore and I am coughing specially when I am in my bed, at night. But today I woke up feeling some shortness of breath, it is worsening, and I´ve started to feel chest pain. It is continuous and is located here, in the middle of my chest, worsening when I breathe and when I lay down on my bed. I practice exercise every day. I run almost 10km every single day. The only medication I use is propranolol, for anxiety and palpitations.

PHYSICAL EXAMINATION Cardiac auscultation is normal; there is some inspiratory crackles in basal lung auscultation; chest pain does not worse with palpation of the thorax; there is no jugular stasis nor lower limb edema.

@patient: Doctor, I am not feeling well last days. Four days ago I felt some **{chills}(disease,loinc:LA30863-7)** and I´ve started to have **{fever}(disease,loinc:MTHU013518)** all day long. I think it´s a kind of flu , because my **{nose}(loinc:LA22156-6)** is stuffy, my **{throat}(loinc:LP96315-4)** is sore and I am coughing specially when I am in my **{bed}(mesh:D001513)**, at night . But today I woke up feeling some shortness of breath, it is worsening, and I´ve started to feel **{chest pain}(disease,loinc:LA28842-5)**. It is continuous and is located here, in the middle of my **{chest}(loinc:MTHU002671)**, worsening when I breathe and when I lay down on my bed. I practice exercise every day. I run almost 10 km every single day. The only medication I use is **{propranolol}(chemical,loinc:MTHU005037)**, for **{anxiety}(disease,loinc:LA10574-4)** and **{palpitations}(disease,loinc:LA16997-1)**.

PHYSICAL EXAMINATION Cardiac auscultation is normal; there is some inspiratory crackles in basal lung auscultation; **{chest pain}(disease,loinc:LA28842-5)** does not **{worse}(loinc:LA11011-6)** with palpation of the thorax; there is no **{jugular}(anatomy)** stasis nor **{{lower}(loinc:LP35082-4) limb}(anatomy) {edema}(disease,loinc:MTHU019803)**.

* namespaces:
  * mesh: http://id.nlm.nih.gov/mesh/
  * loinc:https://data.bioontology.org/ontologies/LOINC

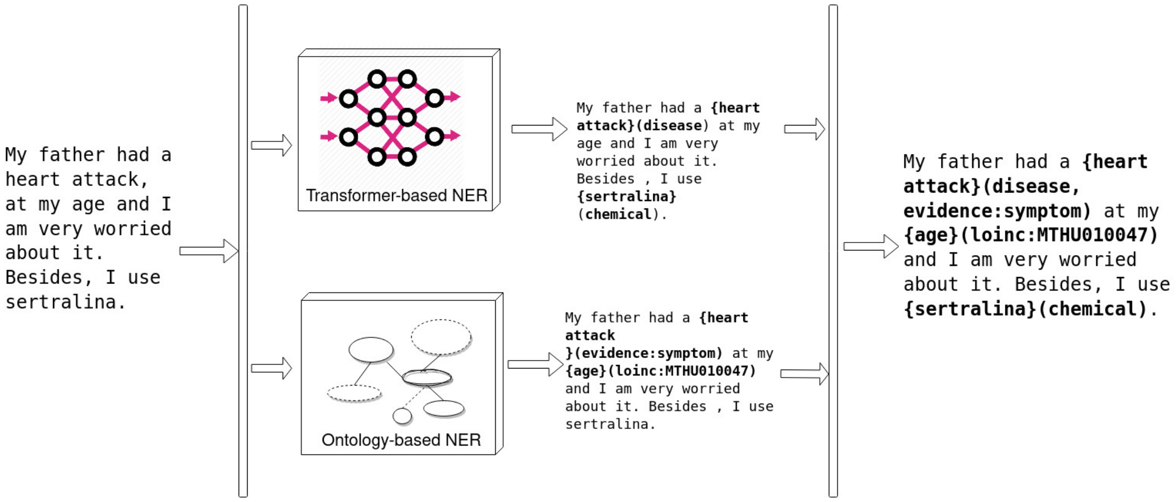
@patient: Doctor, I am not feeling well last days. Four days ago I felt some chills and I´ve started to have fever all day long . I think it´s a kind of flu , because my nose is stuffy, my throat is sore and I am coughing specially when I am in my bed at night . But today I woke up feeling some shortness of breath, it is worsening, and I´ve started to feel chest pain. It is continuous and is located here, in the middle of my chest, worsening when I breathe and when I lay down on my bed. I practice exercise every day. I run almost 10 km every single day. The only medication I use is propranolol, for anxiety and palpitations.

PHYSICAL EXAMINATION Cardiac auscultation is normal; there is some inspiratory crackles in basal lung auscultation; chest pain does not worse with palpation of the thorax; there is no jugular stasis nor lower limb edema.

🟥 Disease 🟦 Chemical
🟨 Anatomy □ Ontology's concept

Figure 3.1: (1) A Clinical Case narrative in free-text format; (2) The Semantic Clinical Case produced from 1; (3) A visual representation of the Semantic Clinical Case.

This method classifies the sentence terms according to the clinical domain-specific labels `Anatomy`, `Chemical`, and `Disease`. In order to train the model, we generated a small labeled corpus called ACD (**A**natomy, **C**hemical, **D**isease) from the concatenation of the two pre-existing labeled data sets: BC5CDR [86] and AnatEM [118]. The Disease class



My father had a heart attack, at my age and I am very worried about it. Besides, I use sertralina.

My father had a **{heart attack}(disease)** at my age and I am very worried about it. Besides , I use **{sertralina}** (**chemical**).

My father had a **{heart attack }(evidence:symptom)** at my **{age}(loinc:MTHU010047)** and I am very worried about it. Besides , I use sertralina.

My father had a **{heart attack}(disease, evidence:symptom)** at my **{age}(loinc:MTHU010047)** and I am very worried about it. Besides, I use **{sertralina}(chemical)**.

Figure 3.2: The hybrid approach to NER.

in BC5CDR also includes disease mentions [52], which comprises Signs and Symptoms. According to [42], BC5CDR and AnatEM do not exhibit a significant overlap between the training sentences of one dataset and the test sentences of the other one (it would expose the training algorithm to sentences of the validation set), which indicates the feasibility of concatenating them. Table 3.1 provides some metrics about the ACD corpus.

Table 3.1: ACD corpus statistics.

Corpus	# Sentences		Entities	# Annotations	
	Training	Test		Training	Test
ACD	21,223	10,867	Anatomy	9,085	4,616
			Chemical	10,550	5,378
			Disease	8,428	4,424

We reused the BioBERT-embeddings which are pre-trained on an unsupervised Masked language Modeling task specialized on biomedical domain (see Section 3.2.2 for more details). Then, we trained our supervised NER algorithm by adjusting the BioBERT-embeddings through a Transfer Learning technique called Fine-tuning. Therefore, this NER algorithm is a semi-supervised method once it reuses the embeddings produced by an unsupervised source task (i.e., Masked Language Modeling) in a supervised target task (i.e., NER).

This proposed Deep Neural Network – called Envoy – is a stack of 12 BioBERT layers plus an extra Fine-Tuning layer liable for specializing the network (by adjusting the parameters) to recognize the label of each term inside the given input sentence.

The Fine-Tuning process involves adding an extra linear layer to the top of the pre-trained neural model and adjusting its weights for each sentence on the training dataset. Each neuron on the first layer processes a token of the input given sentence and forwards it to the next neuron layer. This process is repeated for each sample sentence on the training data set.

We released our NER model (specialized to recognize anatomy, chemicals, and diseases) at: `https://huggingface.co/fagner/envoy`. The open-source code to extend the BioBERT language model by fine-tuning it in a multi-class NER model is available at our fork of BioBERT: `https://github.com/faguim/biobert-pytorch`.

A REST-based implementation of the Harena Semantics framework can be deployed as a container: `https://github.com/datasci4health-incubator/harena-semantics`.

Model Evaluation To validate our approach, we conducted an intrinsic evaluation [148] measuring the performance of the model at performing the NER task objective. The experimental setup is as follows:

- Pre-trained Model: We experimented with both `biobert_base_cased`[2] (containing 768 hidden states, 12 neuron layers and totaling 100 million parameters) and `biobert_large_cased`[3] (24 layers, 1024 hidden states).
- Learning rate $5e^{-6}$ using AdamW optimizer, chosen by considering a threshold between performance and stability, since high learning rate increases performance while incurs instability on the training [35].
- Batch size: 32 sentences/batch (*i.e.,* 4 sentences $*$ 128 tokens = 512 tokens/batch).

Among the two versions adopted in the evaluation, the `envoy_large` version accomplished 85,8% of success on the f1 score, while `envoy_base` 85,5% .

Although `envoy_large` present better results (lower error rate and higher f1-score), we released `biobert_base` as the official version of Harena Semantics due to its smaller model size (`biobert_base` is 432 MB, while `biobert_large` is 1,5 GB) which facilitates the deployment of Harena Semantics in personal pcs. Another relevant feature of smaller models is their stability on the training process [103].

Ontology-based Named Entity Recognition

This rule-based NER method uses ontologies as source information to label the sentence terms with the concept labels formally defined on biomedical ontologies. The algorithm looks for matches (exacts or partials) between the free-text sentence terms and ontology concepts. It provides two modes to match against ontologies:

- External ontologies: This mode uses the biontology annotator to match terms against ontologies stored on the open repository Bioportal [108].
- Local ontologies: This mode uses a RDF database to store RDF triples. We developed an API called OntoMatch to query the RDF triples through the RDFLib python library. Ontomatch enables matches based on a range of metrics such as Levenshtein, Jaccard, Cosine etc.

[2] `https://huggingface.co/dmis-lab/biobert-base-cased-v1.1`

[3] `https://huggingface.co/dmis-lab/biobert-large-cased-v1.1/tree/main`

3.4 Case Study on the Harena System

To get a pilot evaluation of Harena Semantics, we conducted a case study running it over the Harena system [45]. The Medicine course from University of Campinas uses Harena as a supporting pedagogical tool to situate the individuals in an e-learning environment, which simulates the context of an Emergency Care Unit [44].

This section presents a process to construct semantically rich Virtual Patients (VPs). We intend to reinforce the feasibility of a global knowledge network connecting the information scattered in different Virtual Patient systems.

This research paper intends to explore ways of increasing the underlying structure of Virtual Patients towards the glimpse of the Semantic Web. More structured information can be interpreted by machines, expanding the possibilities of application: (i) it becomes easier to find, reuse, and group cases and parts of cases – e.g., it becomes possible to query: cases in which the patient experienced shortness of breath; cases where the ECG was fundamental to diagnose a heart disease; (ii) data from cases can be used beyond the scope of training as a Citizen Science data source.

Our approach focuses on building a Semantic Virtual Patient from free-text Clinical Case narratives. The deployment of a Semantic Virtual Patient potentially facilitates intelligent searches, complex queries, and easy exchange between institutions. As detailed in previous sections, Harena Semantics identifies clinical concepts and links them to ontology concepts through the Automatic Annotation Mechanism and Versum tags. Therefore, it creates a RDF graph representing key knowledge about the virtual patient and integrating it to an interconnected network of concepts envisioned by the Semantic Web research area.

To evaluate the feasibility of creating and curating Semantic Virtual Patient using our framework, we departed from Virtual Patients manually annotated by doctors. These annotations are part of the case rationale, they relate relevant symptoms to the problem (disease) narrated on the Clinical Case.

At the authoring process, the author tags relevant symptoms and indicate whether they are directly related to the clinical case (e.g., arterial hypertension and acute onset of chest pain), or key to the diagnosis (e.g., pain radiating to neck and back), or just distractors to the learner which mislead her to a wrong direction (e.g., symmetric radial pulses is a specific sign but present in only one third of the patients). The diagnosis, which will be presented in the final of the case presentation as a feedback is also annotated.

Harena can superimpose several layers of annotation in the same text content and combine them throughout superimposed contexts.

It is possible to assign a context to any segment of text surrounding it by double curly braces `{{ }}`. Each context can receive an identifier prefixed by at sign `@`. Segments with the same identifier must refer to the same textual content, even though they can afford distinct superimposed annotations. For example, the three following contexts refer to the same text fragment through the identifier `@symp01`:

```
{{@symp01/evidence:finding_relevance
mesh:D000784
 A man, 52 years old, reports he is
 feeling {very strong chest pain}/
 evidence:corroborate_finding/.
}}

{{@symp01
 A man, 52 years old, reports he is
 {feeling}(loinc:MTHU021518) {very
 strong}(loinc:LA28441-6)
 {chest pain}(loinc:LA28842-5).
}}

{{@symp01
 A man, 52 years old , reports he is
 feeling very strong {chest}(anatomy)
 {pain}(disease)
}}
```

The first copy of the segment was annotated by physicians, as previously described. Besides the context id, it is possible to specify the target of the annotations. In this case, `evidence:finding_relevance mesh:D000784` indicates that the following annotations point to the relevance of the symptom to the Aortic Dissection disease (`mesh:D000784`). The second copy was annotated by the ontology-based annotation mechanism and the third by the Transformer-based mechanism.

These superimposed annotations are transformed into an RDF Graph [127] as shown in Figure 3.3. A uniquely identified RDF resource (node in the RDF graph) is associated with each word. When the annotation refers to a word – e.g., the chest is annotated with `Anatomy` (which in turn refers to `mesh:D000715`) – the related RDF node is connected by a `skos:related` association. SKOS - Simple Knowledge Organization System is a data model for knowledge organization [101].

When annotations refer to a sentence with more than one word – e.g., the sentence `chest pain` annotated by the concept Chest pain in the LOINC Document Ontology (loinc:LA28842-5) – a node aggregating the sentence's words is created and related to the concept. An aggregation will reuse already aggregated nodes whenever is possible, as in the case of the node that aggregates "very strong" and "chest pain" aggregated nodes.

Following this process, annotated content will converge to a semantic RDF profile of a virtual patient, as shown in Figure 3.4. The diagram shows a simplified version of the graph presenting an overview of what we call: Semantic Virtual Patient. The case in the example is related to an aortic dissection – a dangerous injury to the innermost layer of the aorta, which puts the life of the patient at risk.

In the long term, the produced semantic clinical cases could be used to grasp knowledge from the unstructured text within the clinical narrative. The narrative scripts in a machine-interpretable format enable sharing, versioning, and crowdsourcing. Such capabilities are needed for a system with clinical case data.

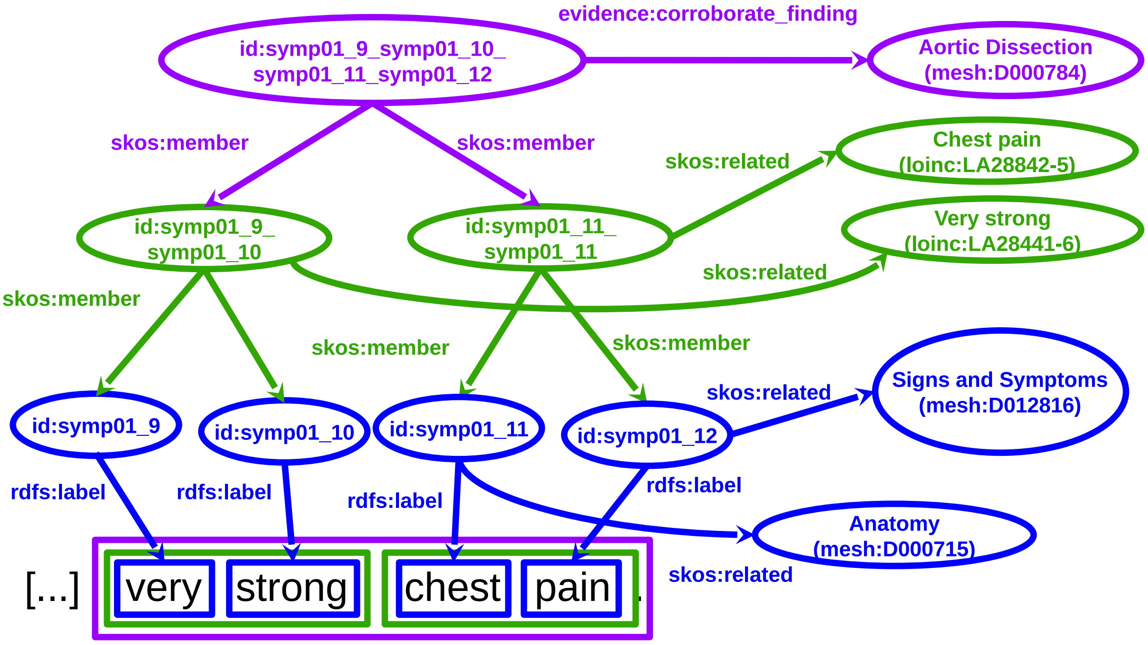
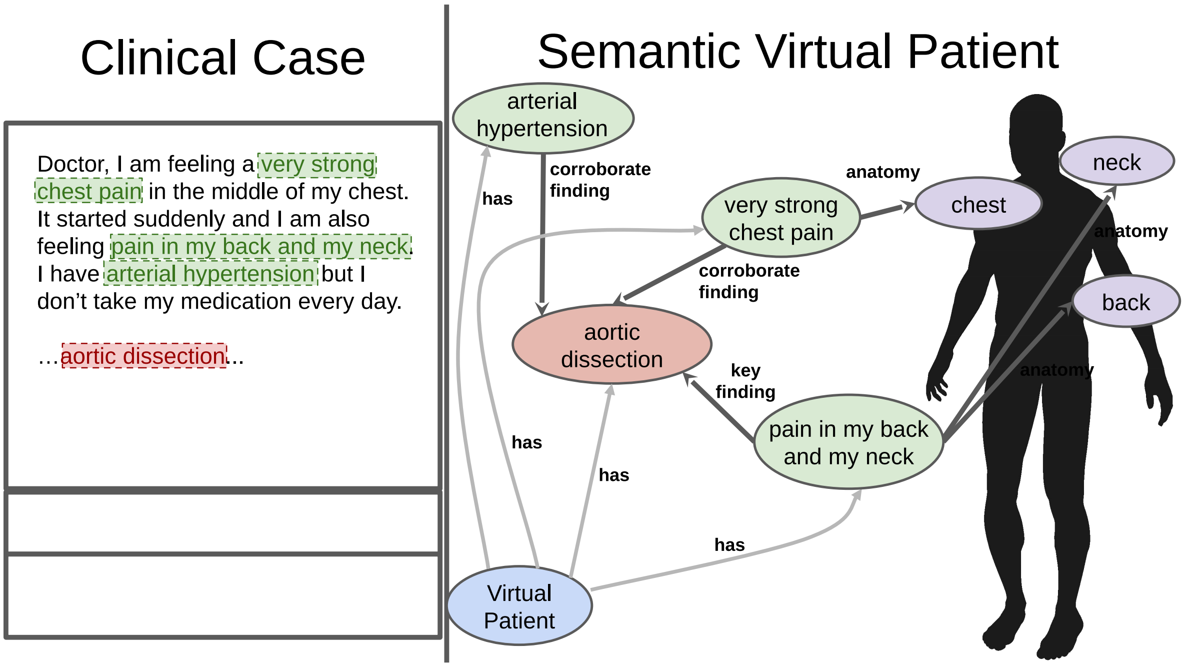


Figure 3.3: The RDF graph representing a case of Aortic Dissection, built with the support of Harena Semantics.



Doctor, I am feeling a very strong chest pain in the middle of my chest. It started suddenly and I am also feeling pain in my back and my neck. I have arterial hypertension but I don't take my medication every day.

…aortic dissection…

Figure 3.4: Key elements of a Semantic Virtual Patient RDF Graph extracted from a Clinical Case.

3.5 Conclusion

This paper presented Harena Semantics, a framework to enable Citizen Scientists to create semantic annotations directly into the text narrative of Clinical Cases. By adopting our approach, the data crowdsourced in Citizen Science systems may incorporate the information gathered in the Knowledge Network envisioned by the Semantic Web research area.

The introduced Automatic Annotation Mechanism benefits both from Rule-based (the ontology-based NER) and Statistical Learning (the Transformer-based NER) approaches. Our NER task presents results comparable to the state-of-the-art works in such research area. Our approach to superimpose annotations enables to combine human and automatic annotations to produce a knowledge network representing our Semantic Virtual Patient.

The technology stack presented in this paper could serve several purposes. As future works, we intend to implement a search engine to retrieve Clinical Cases aided by the support of semantic information. Besides, the Semantic Virtual Patient can also be used to train inference systems that automatically generate feedback to the users of Learning Environments. These educational resources must be adherent to pedagogy practices, therefore it is necessary to develop approaches to involve experts, professors, and scientists in the creation of these resources. The Harena Semantics is an initiative engaged in such effort.

Acknowledgements

This study was financed in part by CNPq grant number 428459/2018-8. We thank dr. Tiago Grangeia for the production and annotation of the Clinical Cases addressed in this work.

Chapter 4

Semantic Representations based on Neural Topic Models

A Topic Model is a sophisticated tool to extract statistics-based semantics from the inside of textual collections. However, in spite of extensive work on topic modeling, including its evaluation using Large Language Models, there is need for more research to measure, analyze and validate the topics learned. To this end, we present a Neural Topic Model called ABT whose goal is to uncover the statistical semantics latent in large corpora. ABT enables one to explore the themes covered by a given textual collection and analyze them as a hierarchical structure of topics. Our method is grounded on Neural Language Models as computational abstractions to encode – i.e., computationally represent – and to reveal semantic aspects underlying the structure of the textual data. We showcase the effectiveness of our approach by conducting a case study in the Medical domain, working with the CliCR dataset, a collection of 10.538 clinical cases, comprising textual narratives of situations that occurred in real clinical environments. The topics that resulted from our study highlighted medical concepts that are shown to be effectively used as discriminant topics within the CliCR dataset according to the language modeled and represented in the network's neurons. In this sense, a language model can reveal idiosyncrasies and patterns previously kept latent in the collection and evidenced by the neural network. We experimented with different language models to investigate the effects of using specific tasks to tune the neural network. Although each language model produces completely different topics, none of them outperforms the others regarding the Topic Coherence metric. Nevertheless, the tree structure of the topic collection evidences some interesting patterns revealed by each language model. Our study sheds light on recent findings of the related literature, which state that the topic coherence metric does not align with assessments performed by human evaluators.

4.1 Introduction

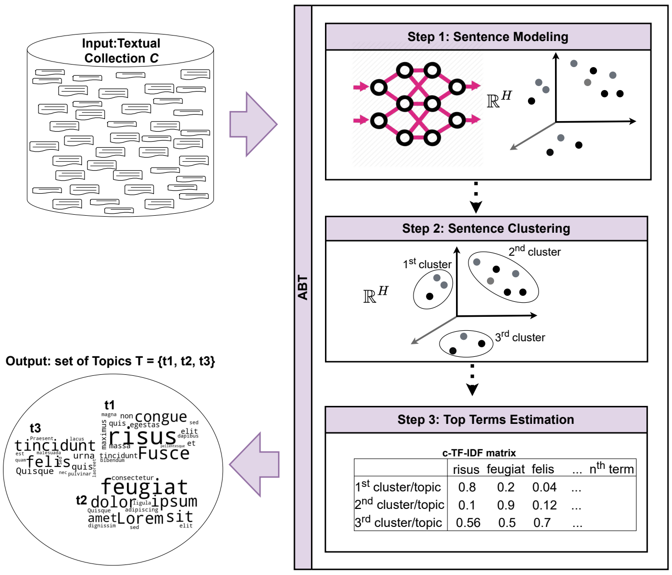
Topic Modeling [22, 62, 102, 68, 63] is a Natural Language Processing (NLP) task whose objective is to identify the main topics covered in a textual collection and encode them succinctly in a computational abstraction referred to as a Topic Model. A topic is said to be coherent if the words from such a topic are associated [28], that is, when it brings together a group of words that make sense when used in the same context, such as the Biomedical topic formed by the words: {myocardium, coronary, acute, artery, pain}.

A topic model provides an overview of a given textual collection, and is handy when the massive amount of text prevents a person from reading the entire collection. Furthermore, the resulting topics may be helpful to specific problems in the NLP research agenda. For example, they can improve diverse tasks, such as Sentiment Analysis, Machine Translation, and Domain Adaptation of Language Models [27].

Although there is extensive literature on Topic Models, there is still a lack of validation regarding the coherence, interpretability and organization of the produced topics. Therefore, in this work we analyze the hierarchical structure underlying the set of topics and the Topics Coherence measured in each level of the inferred structure.

To this effect, this article presents our Attention Based Topics (ABT) method, a Neural Topic Modeling approach [7, 63] that identifies the major themes in a given textual collection. Our method produces topics based on Neural Language Models [46, 121] to represent the textual collection through multidimensional vectors.

As depicted in Figure 4.1, our method transforms a given collection of textual sentences C into a set of topics T that encodes the gist of C, according to the following pipeline: (i) Initially, through a pre-trained language model, the sentences $s \in C$ are converted into a vector representation $\mathbb{R}^H$, which incorporates statistics-based semantics in its vector space of H dimensions; (ii) next, a Hierarchical Aggregation algorithm clusters the sentence vectors $r \in \mathbb{R}^H$ that are close together in the vector space – each cluster becomes the basis of a topic; (iii) these resulting clusters are allocated to a TFIDF matrix adapted for topics [63], which correlates them with the significance of each word within the topic regarding the vocabulary of C. As output, the method assigns to each topic $t \in T$ the set W, which contains its best-ranked words according to the cTFIDF metric.


Input:Textual Collection C
ABT
Step 1: Sentence Modeling
$\mathbb{R}^H$
Step 2: Sentence Clustering
$\mathbb{R}^H$
1st cluster
2nd cluster
3rd cluster
Step 3: Top Terms Estimation
c-TF-IDF matrix
Output: set of Topics T = {t1, t2, t3}
t1
t2
t3
risus
congue
feugiat
tincidunt
felis
Fusce
dolor
ipsum
amet
Lorem
sit

	risus	feugiat	felis	...	nth term
1st cluster/topic	0.8	0.2	0.04	...	
2nd cluster/topic	0.1	0.9	0.12	...	
3rd cluster/topic	0.56	0.5	0.7	...	

Figure 4.1: An overview of our 3-step pipeline working with a given collection of sentences. The steps operate over the input collection to produce a set of topics that summarizes it.

In this way, ABT infers latent topics from contextual information extracted from the collection, combining two representations used in the process in a complementary way: the semantic vector space $\mathbb{R}^{768}$ (related to the language model) and the matrix cTFIDF. These two representations model statistical patterns observed in the underlying structure of the textual collection. Paraphrasing the concept of the gist of a document – proposed by [62] –, we state that our approach encodes the gist of the input collection.

We conducted a case study of our approach, applied to the biomedical domain, using a collection of Clinical Cases extracted from the CliCR corpus [135]. We validated the quality of the produced topics by measuring the Topics Coherence metric [122], which assesses whether the most discriminant words in a set of topics reinforce each other.

We qualitatively evaluated the behavior of the model in terms of topic granularity. Our method produces topics organized in a Hierarchical Structure underlying the pre-trained language model. Our approach presents itself as an alternative view to the topic modeling problem, taking advantage of the context and linguistic knowledge internally incorporated into the neural network parameters to produce coherent topics with different granularity options along the path of the hierarchical structure.

The remainder of the text is organized as follows: Section 4.2 discusses background foundations and related work; Section 4.3 describes our method; Section 4.4 presents quantitative and qualitative results of a case study of our ABT approach tailored to the medical domain handling a collection of clinical cases. Finally, Section 4.5 presents our concluding remarks.

4.2 Background and Literature Review

4.2.1 Definitions and Terminology Adopted

Topic Modeling [22, 62, 102, 68, 63, 1] is an unsupervised Natural Language Processing challenge whose goal is to discover topics that represent an overview of the textual collection under analysis. A topic model explicitly represents the latent semantic structure [7] – or gist [62] – of a textual collection.

A wide range of specific-domain applications can benefit from the topics discovered using such a model, such as health, cognition studies, social studies, bioinformatics, or education [75, 82, 68, 27, 62]. For instance, [75] apply topic models to analyze the correlation between individual moral concerns and language usage. In another branch of research, [82] investigate the presence of Post Traumatic Stress Disorder in textual reports on Reddit[1] using topic models.

The concept of Topics Model refers to a discrete probability distribution describing the connections between words, topics, and documents [27]. Topics are word combinations that demonstrate idiosyncrasies in the linguistic distribution of the corpus under analysis [25]. Topic models are explicit representations that probabilistically associate documents with topics and topics with words [22].

According to [1], there are four categories of topic models: algebraic, fuzzy, probabilis-

[1] https://www.reddit.com/

tic, and neural. These different TM categories offer different advantages and are suited for particular settings. Our work can be characterized as being part of neural topic models.

Many methodologies address the Topic Modeling problem, such as Latent Semantic Analysis (LSA) based on linear algebra and its probabilistic version pLSA. Such methods apply dimensionality reduction to the documents represented in a Bag-of-Words format. Bag-of-words representations are adequate since, by hypothesis, word order is not a determining factor in these methods [22].

Latent Dirichlet Allocation (LDA) [22] is a probabilistic model for discrete data collections, such as textual data. LDA uses two complementary distributions: a topic over words distribution that describes the relationship between topics and words; and a second distribution that allocates topics to documents [27]. For LDA, a document is a random mixture of latent topics, which in turn are probability distributions over vocabulary words. LDA is a dimensionality reduction technique that is considered one of the most robust and efficient methods for topic modeling, and applies to problems beyond the textual domain [22].

4.2.2 Neural Topic Modeling

Neural Topic Modeling (NTM) is a current research trend that combines topic modeling with Neural Language Models (e.g., [7, 47, 142, 114, 157, 68, 8, 82]). Indeed, this is a rapidly growing area of research. [142, 157] discuss the similarity between the topics produced via NTM and other traditional topic models, such as LDA. NTM-based works claim to produce more interpretable topics than prior methods, yielding improvements in the state-of-the-art concerning topic coherence measure [68].

Neural Language Models

A language model is a computational abstraction that represents a unit of textual information (such as a word, a sentence, a paragraph, or a document) through a numerical vector that encodes the most relevant characteristics of this textual unit that are typically induced by a neural network [146]. This set of relevant characteristics alludes to a type of semantics grounded on the Distributional Hypothesis [93], which deems that words that occur in the same context tend to have similar meanings.

Language models (also known as feature vectors, word representations, word embeddings, vector representations, or word vectors [18, 146, 23]) are typically induced using a deep neural network [146].

Mikolov et al., [98, 99, 100] presented neural networks to train language models to represent linguistic regularities in the form of a vector space. This vector space allows one to perform vector algebra operations and, therefore, infer a kind of semantic-statistical relationship between the words based on the displacements between the vectors of the textual collection modeled in the neural network.

Another example of regularity captured through simple algebraic operations on the vectors of words is the similarity between the words big and bigger in the same way as between the words small and smaller [98]. There are indications that the linguistic regularities captured by the neural models carry apparent syntactic and semantic aspects preserved in the statistical patterns encoded in the referred semantic vector space.

To the best of our knowledge, the Transformer is currently the most robust and suitable model for pre-training Language Models, which will be further explored in the next section.

Transformers

The transformer model [147] is a deep neural network that efficiently solves long-date problems faced by the NLP research area. The Transformer architecture implements the Attention Mechanism to infer the statistical context of a sentence as a whole – i.e., a sentence's global context. This critical gap inherent to previous models has limited their contexts to a window of few terms around the focus.

The transformer produces a set of characteristic vectors, encoded in the matrix QKV [147] as a byproduct of the training activity [89]. While training, the characteristic vectors are updated with the contribution of all terms surrounding the current training term. In this way, the transformer is aware of the bi-directional context of each term at hand [123] and can guide the attention to the more discriminant terms in the entire training dataset [89] in both directions. Differently, the previously proposed approaches (e.g., LSTM) were limited to obtaining information just from one direction, left-to-right or right-to-left.

The characteristic vectors produced by the transformer can be reused as a universal representation to accomplish other tasks beyond the original task performed by the transformer, since the matrix QKV embeds syntactic and semantic aspects that are discriminant to many NLP tasks. During the training phase, the model perceives such aspects as idiosyncrasies and encodes them in the inferred vector. One can use this linguistic knowledge encapsulated in QKV to perform other tasks through Transfer Learning techniques [67].

BERT (Bidirectional Encoder Representations from Transformers)

BERT is a neural language model pre-trained on the Masked Language Modeling task through the Transformer architecture [46]. The pre-training objective is to uncover the term hidden by a MASK symbol in the sentences given as input (such as "I like this [MASK] bed"). During this pre-training phase, the network embeds the bidirectional context associated with the statistical relationship between the given masked term with all its surrounding terms within the sentence. The context emerges from patterns learned in all the sample sentences previously observed by the model by performing the language modeling during the pre-training phase. The language modeling captures a wide range of resources and facets – e.g., long-term dependencies, hierarchical relationships, sentiments, and patterns spiked at the writing – directly associated with the source task used in the pre-training [67].

BERT is a breakthrough model improving the state-of-the-art in many NLP tasks. To the best of our knowledge, BERT word vectors are one of the most powerful methods to transform textual information into an effective computational abstraction rich in context-sensitive information.

Internally, within the network parameters, BERT maintains a vector space $\mathbb{R}^H$ of multiple H dimensions. Usually, $H = 768$ depending on the selected version: BERT$_{base}$ (H=768, parameters=110M) or BERT$_{large}$ (H=1024, parameters=340M) [46].

The BERT vector space $\mathbb{R}^H$ embeds a 30000 token vocabulary. To put it less abstractly, Figure 4.2 illustrates BERT space $\mathbb{R}^{768}$ in a 3-dimensional rendering obtained by using the Embedding Projector tool[2]. This 3D picture is drawn after reducing the dimensions (through a PCA method) from $\mathbb{R}^{768}$ to $\mathbb{R}^3$, because it is impossible to draw/visualize the original 768 dimensions.

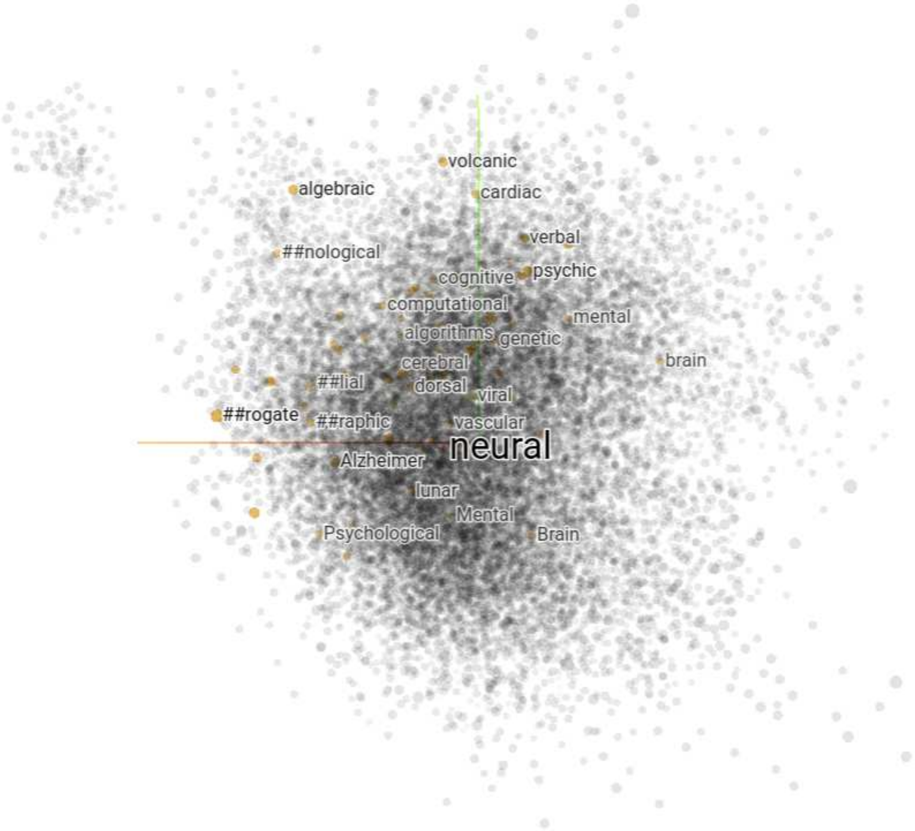


Figure 4.2: A 3-dimensional picture of the BERT vector space captured by reducing it from R^{768} dimensions into R^3. This space is internally written in the parameters of the network. In such a space, how close two terms are, more similar they are, and therefore probably occur in similar contexts.

In this vector space, the closer two terms are, the more similar they are, and therefore probably occur in similar contexts. The figure highlights the word "neural" and its more similar words. It is possible to note two distinguishable groups of words {cognitive, psychic, genetic, cerebral, Mental, ...} and {computational, algorithms} standing for two main contexts: referring to the biological Neural System and to Artificial Neural Networks from the Computer Science terminology.

It is also worth mentioning the words {##nological, ##rogate, ##raphic, ##lial} in Figure 4.2 which are used to compose rare and out-of-vocabulary words. These terms are sub-word units [153], a minimal set of terms that is capable of representing any arbitrary

[2] `https://projector.tensorflow.org/`

word. For example, the word Neurological is not in the vocabulary, it will be processed into a sequence of sub-word units: {N, ##eur, ##ological}.

The development of such a minimal set of sub-words is another important aspect of the recent NLP technological leap (besides the Attention Mechanism) and consequently the Artificial Intelligence. Since creating some kind of word inventory is a long-standing problem in NLP tasks.

Sentence-BERT

The arrival of word vector spaces (e.g., BERT) has inspired other approaches (e.g., SkipThought [79], InferSent [39], Sent2Vec [109] and Sentence-BERT [121]) to generate language models for larger text segments, such as phrases, sentences, paragraphs, and even entire documents.

Sentence-BERT (SBERT) is a sentence language model [121] which uses siamese networks (i.e., networks with tied weights) to infer sentence vectors from the BERT word vectors. The training procedure of Sentence-BERT takes a pair of sentences as input, transforms each one into a sentence vector from the averaging of its word vectors, and computes the similarity between the two sentence vectors. In the end, both sentences' vectors are tuned according to this similarity function, so that similar pairs are closer and dissimilar pairs are further apart.

4.2.3 Relating our study to other work

Our research is centered on topic modeling and on analysis of inherent latent semantics.

Top2vec [7] infers topic vectors by applying vector algebra over the neural vectors of words and documents embedded in the same vector space. Each topic corresponds to a centroid of a cluster of documents and takes the closer word vectors as its most representative words. The approach infers the optimal number of topics through the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm.

The work of [8] presented a model which recognizes topics combining Transformers, community detection techniques in graphs, and named entity recognition.

BERTopic, presented in [63], is a model based on clusters of documents grouping vectors inferred through the Sentence-BERT framework [121]. It assigns the clusters to a cTFIDF matrix that indicates the most representative words of each topic. cTFIDF is a variation of the classic TFIDF algorithm, taking clusters as the unit of analysis, instead of documents.

Top2vec and BERTopic work best when applied to low-dimensional data since both use HDBSCAN as the clustering technique. Therefore, both need to apply a dimensionality reduction on the sentence vectors before clustering them. Our work is similar to that of BERTopic in modeling the topics as clusters of sentences and using cTFIDF to estimate the words' importance in a topic/cluster. As will be seen, we adopted a Hierarchical Aggregation algorithm instead of HDBSCAN, thereby avoiding dimensionality reduction, thus working with more original information. Our approach differs as it grounds on a hierarchical aggregation algorithm. Hence, it can deal with sentence vectors in their 768 original dimensions.

Although many studies use hierarchical clustering algorithms – such as HDBSCAN – their focus is not on the hierarchy, as they apply dimensionality reduction that interferes with hierarchical analysis. We provide an analysis of the inferred hierarchy of topics and its relationship with Language Models based on the Transformer architecture and its attention mechanism.

Our analysis relies on hierarchical structures of topics, such as [78, 71, 156]. Our hierarchical analysis is based on the work of [78], who showed important aspects of the hierarchy of topics and proposed the recursive Chinese restaurant process (rCRP) method to generate hierarchical topic structures with unbounded depth and width. We, on the other hand, produced a hierarchy based on the successive steps of aggregation on the sentence collection.

4.3 The ABT Method

ABT is a Neural Topic Model proposed by us to uncover the statistical semantics latent in large corpora. To describe ABT, we formally define the following terms (extending the notation used by [22] to describe the topic modeling and by [46] to define BERT representations):

- Word w: the basic unit of a textual construction. Each w is a discrete data item from the vocabulary $V = \{w_1, w_2, ..., w_v\}$.

- Sentence s: a sequence of n words denoted by $s = (w_1, w_2, ..., w_n)$. ABT can work with words, sentences, documents and documents excerpts.

- *Corpus*: a collection of m sentences denoted by $C = \{s_1, s_2, ..., s_m\}$.

- Representation $\mathbb{R}^{768}$: a multidimensional vector space of 768 dimensions (the standard adopted by BERT-based models) representing the m sentences from *corpus* C. That is, each vector $r \in \mathbb{R}^{768}$ corresponds to a sentence $s \in C$. There are m sentence vectors of 768 dimensions.

4.3.1 Overview

The ABT pipeline consists of three successive transformation steps (as previously illustrated in Figure 4.1). It takes C as input in plain-text format and does not require any pre-processing over it.

> **Step 1:** Transform each sentence $s \in C$ into a vector representation $r \in \mathbb{R}^{768}$ using the Sentence-BERT framework [121] to model sentences. The resulting language model incorporates a semantic space of multiple dimensions – usually 768 or 1024, depending on the BERT version.

Step 2: Apply Hierarchical Clustering to $\mathbb{R}^{768}$ (i.e., the output from Step 1) to identify a set of disjoint clusters of semantically related sentences. Each cluster characterizes a latent topic t composed by a sentence set S.

Step 3: Fill the cTFIDF matrix [63] with the significance values of each term within the topic. Each row of the matrix corresponds to a topic t characterized by a list W containing the most relevant words in the global context of the collection, i.e., recurrent words in the current topic insofar as rare in the other topics.

As output, ABT produces a set of Topics T. Each topic $t \in T$ is characterized by its set of sentences S (inferred through Step 2) and its set of words W (inferred through Step 3).

The remainder of this section gives more details of each step of our method using the CliCR corpus [135] as the input collection C of a running example. The CliCR is a corpus of 12.000 case reports (from which we have used 10.538 in our running experiments) containing descriptions of situations that occur in real clinical environments. Table 4.1 shows some sentences sampled from the CliCR corpus.

Table 4.1: Textual sentences sampled from the titles extracted from the documents in the CliCR dataset titles of each document.

id	Sentence
1	Isolated cranial distortion mimicking caput succedenum from amniotic band disruption without any neurological abnormality
2	Ruptured pseudoaneurysm of the radial artery
3	Pulmonary alveolar microlithiasis
4	Successful pregnancy with autoimmune cirrhosis
...	
1000	Intraoral schwannoma – a report of two cases
...	
10538	Thrombotic thrombocytopenic purpura in a patient with HIV from Zimbabwe

4.3.2 Step 1: Sentence Modeling

Initially, the framework converts the sentences into vector representations embedding them in an H-dimensional vector space $\mathbb{R}^{768}$. It is inferred by the Sentence-BERT[3] [121] framework that allows modeling sentences – e.g., paragraphs, documents – using a model based on siamese BERT-networks. Sentence-BERT infers a neural sentence model which projects the sentences into a vector space of $H = 768$ dimensions (or 1024 depending on the BERT model version) using a base language model.

As a running example of this step, Figure 4.3 shows the output of this first step by working with the CliCR corpus using the standard BERT model (i.e, BERT-based-cases[4]. In this output vector space $\mathbb{R}^{768}$, each point (i.e,, vector) represents a clinical case from the CliCR corpus. In fact, this visualization is a 2-dimensional vector space obtained by reducing the dimensionality of $\mathbb{R}^{768}$ from its original 786 dimensions into 2 dimensions (because it is not possible to visualize a vector space of 768 dimensions). We have used

[3] `https://github.com/UKPLab/sentence-transformers`

[4] `https://huggingface.co/google-bert/bert-base-cased`

a dimensionality reduction technique called Spectral Embedding [106] implemented using the Scikit-learn[5] Python module.

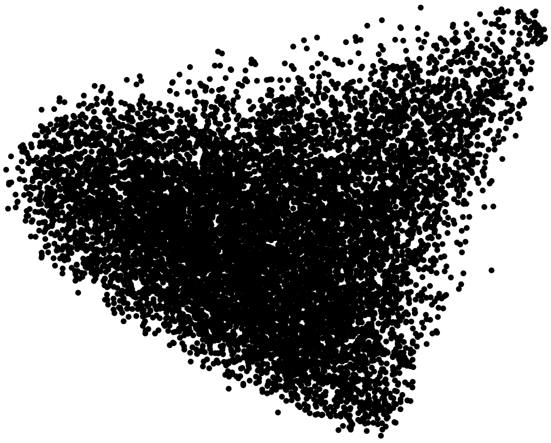


Figure 4.3: Step 1 output. It represents the CliCR corpus projected in a 2-dimensional space (obtained by reducing from the original $\mathbb{R}^{768}$ inferred by the BERT model). Each point represents a clinical case title. In such a space, the closer two points are, the more similar they are.

Sentence-BERT allows one to use any BERT-based language model. In this section we use the original BERT model as a running example to illustrate the method. Sentence-BERT allows using different language models to produce the vectors, inferring different vector spaces calibrated by specific training tasks. As described later in the Results Section 4.4, we experimented with five different BERT-based language models (BERT, BioBERT, BART, all-mpnet-base-v2 and Envoy) to investigate the effects of changing the base model used. Besides the base model, in this step one can configure other hyperparameters: the model dimensionality (768 or 1024) and the maximal input sentence size (e.g., 128, 512).

4.3.3 Step 2: Hierarchical Sentence Aggregation

This step involves applying a Hierarchical Aggregation on the set of vectors $\mathbb{R}^{768}$ to obtain clusters of similar sentences. The clustering algorithm (known as hierarchical aggregation) is an unsupervised method that identifies clusters in a bottom-up approach [81]:

1. initially, each vector $r \in \mathbb{R}^{768}$ defines its cluster;
2. iteratively, the algorithm merges pairs of clusters with the smallest distance from each other among all the cluster pairs;
3. the process continues until it reaches a certain user-defined threshold of distance d – a hyperparameter that indirectly determines the number of output clusters;

[5] `https://scikit-learn.org/stable/modules/generated/sklearn.manifold.SpectralEmbedding.html`

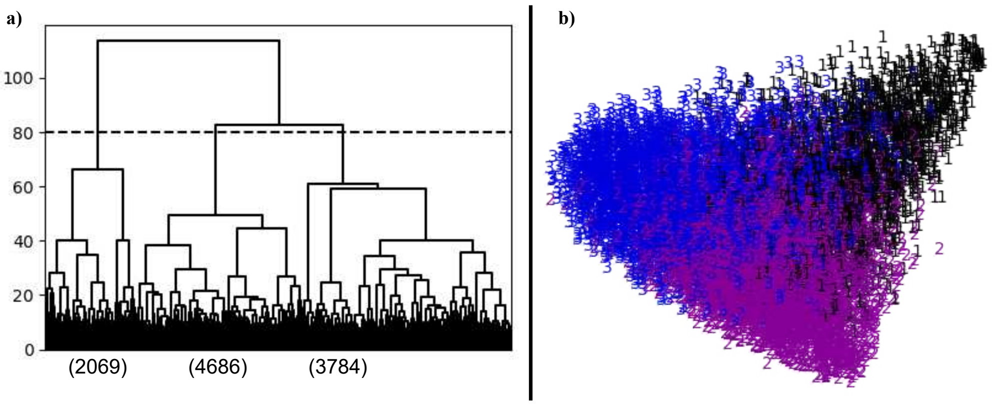


Figure 4.4: **a)** Dendrogram of the hierarchical clustering of the CliCR corpus modeled by the SentenceBERT model. The hierarchical aggregation algorithm produces 3 resulting clusters by defining the distance threshold $d = 80$. **b)** Color coded 2d visualization of the 3 clusters of clinical cases. This partitioned space (i.e., the result of step 2) makes explicit the topics encoded in the network parameters.

4. the algorithm outputs k disjoint clusters corresponding to the set of latent topics T.

In each running of the clustering algorithm, a sentence can only be associated with a single topic. A challenging aspect of topic modeling approaches is the choice of k. An analysis of the clusters hierarchy (as presented in Section 4.4.4) can help one understand the model's behavior as k increases and the sentences are partitioned into more refined clusters.

Figure 4.4 illustrates the clustering process over the space produced by the SentenceBERT model. Figure 4.4.a presents the dendrogram produced by clustering with $d = 80$ and Figure 4.4.b illustrates the $k = 3$ resulting clusters.

Table 4.2 shows selected samples of clinical case titles and their assigned cluster. The sentences shown in Table 4.2 were chosen to be representative of the output of this step and to illustrate the similarity within a cluster's sentences. The sentence from a given cluster must be more similar to other sentences within the cluster than those of other clusters. We measured the similarity using the cosine distance between the sentences within the semantic space.

Table 4.2: Samples of clinical case titles of the 3 clusters produced by the SentenceBERT model. Each cluster gathers a set of similar sentences according to such a specific language model.

Cluster	Sentence sample
1	Wind direction and mental health: a time-series analysis of weather influences in a patient with anxiety disorder. Working effectively with patients with comorbid mental illness and substance abuse: a case study using a structured motivational behavioural approach. Ineffective chronic illness behaviour in a patient with long-term non-psychotic psychiatric illness.
2	Isolated splenic tuberculosis in an immunocompetent patient. Orbital cellulitis with periorbital abscess secondary to methicillin-resistant Staphylococcus aureus (MRSA) sepsis in an immunocompetent neonate. Primary orbital chronic granulomatous reaction to deep staphylococcal infection due to trauma in immunocompetent.
3	Novel balloon application for rescue and realignment of a proximal end migrated pipeline flex embolization device into the aneurysmal sac: complication management. Subarachnoid hemorrhage then thrombosis of posterior inferior cerebellar artery dissection: is early surgical exploration warranted? Hemorrhagic collision metastasis in a cerebral arteriovenous malformation.

The inferred clusters gather sentences that are similar in certain aspects observed, captured, and featured by the language model during its pre-training phase. We notice, for instance, that cluster 3 contains sentences reporting unusual and rare clinical cases. Meanwhile, cluster 1 presents sentences associating symptoms and diagnosis.

It is worth mentioning the special scenario of $k = 1$ (i.e., a single topic to represent the entire collection). In such a case, it is not possible to assemble the cTFIDF matrix since it requires at least 2 clusters. Therefore, to create a $k = 1$ topic the ABT uses a classical TFIDF matrix, considering each input sentence $s \in C$ as a row, which implies a slightly different topic than those $k > 1$ created with the cTFIDF matrix.

The resulting clusters form the basis of the semantic representation built by ABT. We are interested in investigating whether there is a task well suited for topic modeling and apt to produce an optimal set of clusters. Therefore in the Results Section 4.4, we experiment with different threshold distances to infer a range of partitions of the input data set.

4.3.4 Step 3: Representing Topics

Lastly, ABT obtains the most descriptive terms from each cluster – through the cTFIDF measure, which indicates the relevance of each term within the cluster [63] – and assigns them to the corresponding topic $t \in T$, according to:

$$cTFIDF_{w,t} = TF_{w,t} \times \log(\frac{k}{TF_w}) \tag{4.1}$$

Where $TF_{w,t}$ models the frequency of the word w in the topic t. Here, the topic t is the concatenation of all sentences s in the cluster that corresponds to t. The Inverse Topic Frequency IDF measures how much information a word w provides to a topic t by taking the logarithm of the total number of topics k divided by TF_w the frequency of word across all topics. In other words, TF_w is the number of topics that contain w.

Then, it fills the cTFIDF matrix by allocating the cluster in the rows and the vocabulaŕy words $V = \{w_1, w_2, ..., w_v\}$ in the columns, as exemplified in Table 4.3. The cTFIDF algorithm gives information on how relevant each term is to the referred cluster, considering the common terms to the cluster insofar as uncommon in other clusters.

Table 4.3: Example of cTFIDF matrix to 3 topics inferred by using the SentenceBERT model. Each topic contains a list of top words sorted by c-TF-IDF metric. For example, the top word "mental" from Cluster 1 exhibits cTFIDF = 0.14.

Cluster	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	...	Term 7320
1	mental 0.14	guillain 0.12	your 0.12	still 0.10	illness 0.10	kikuchi 0.09	barr 0.08	this 0.08	about 0.08	forget 0.08	...	β3 0.0
2	immunocompetent 0.26	cardiomyopathy 0.19	cystic 0.15	diffuse 0.15	mutation 0.13	intravenous 0.12	month 0.11	autoimmune 0.10	chemotherapy 0.10	staphylococcus 0.10	...	laptop 0.0
3	aneurysm 0.47	dissection 0.28	arteriovenous 0.19	gastrointestinal 0.18	ruptured 0.14	occlusion 0.14	inferior 0.12	uterine 0.12	vertebral 0.12	maxillary 0.11	...	β3 0.0

The output produces a set T containing a total of k topics, where each topic $t \in T$ is characterized both by the set W (containing its most discriminant words) and by the set S (containing its clustered sentences).

Figure 4.5 shows the topics produced in the output of this step, represented in the vector space by color, and, for each, we present a word cloud with the best-ranked terms in cTFIDF.

This step allows one to filter out stop-words from the corpus by setting a frequency threshold – i.e., filter out words appearing above $x\%$ in the corpus. In this way, the

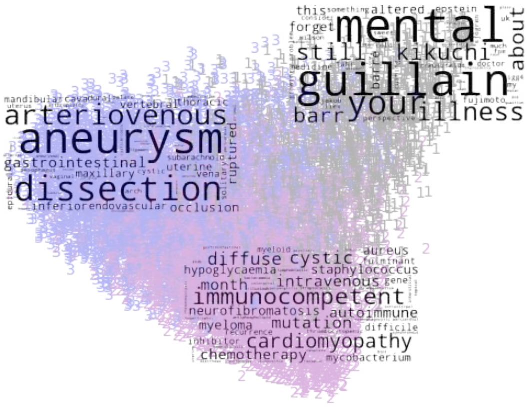


Figure 4.5: Word clouds of most relevant terms by each topic, inferred using the BERT model.

method dynamically infers a list of stop words, thus alleviating one of the bottlenecks inherent in classical approaches that use fixed and predefined stop-word lists [107].

4.4 Evaluation of Results

We applied a type of probing methodology [140] to investigate whether ABT consistently models some (if any) information associated with the topic model's phenomena. We have conducted three distinct evaluations:

1. a quantitative analysis regarding the Topics Coherence metric.
2. a qualitative analysis of the Hierarchy Structure underlying the set of discovered topics.
3. an ablation study arguing about the misalignment between the metric Topics Coherence and human evaluations. Such an issue was originally approached by [68] and reinforced by the results of our work.

4.4.1 Materials and Methods

To evaluate our approach, we conducted a case study in a collection of Clinical Cases extracted from the CliCR corpus [135]. Each clinical case from CliCR describes a medical narrative of situations occurring in real clinical environments. In medical education, professors use clinical cases as pedagogical resources to teach clinical practices to medical students. Table 4.4 illustrates some metrics of the CliCR corpus.

As proof of concept, our analysis focused on case titles instead of the entire content of a clinical case (see Table 4.4). However, ABT is flexible enough to handle larger text fragments – by improving the Sentence-BERT model to deal with more characters than the standard 128 BERT-tokens. Although such an improvement is theoretically feasible,

Table 4.4: The CliCR corpus statistics. A standard BERT-based model of 128 tokens is enough to handle sentences of the average length of case titles. In turn, the entire content of a clinical case demands a model able to deal with more BERT tokens.

Number of cases	10.538
Average cases length (characters)	1457
Average cases titles length (characters)	79,12

it would demand more computational resources (e.g., more memory and more powerful GPUs).

4.4.2 Experimental Setup

We analyzed the use of five BERT-based language models in the first step of the pipeline, run on a Linux PC with 8 GB of RAM size, Intel i5, programmed in Python.

- BERT: the standard language model [46] designed by Google. We have used the bert-base-cased version (containing 768 hidden states, 12 neuron layers, and totaling 110 million parameters).
- BioBERT: a biomedical domain-specific language model [84].
- Envoy: a fine-tuned model[6] for Named Entity Recognition (NER) task, whose objective is to identify and classify entity types. This model is tailored to recognize the following entities: Anatomy, Chemical, and Disease [110].
- BART: a model effective for summarization tasks [85].
- all-mpnet-base-v2: a model[7] produced by the sentence-BERT framework [121].

The parameter setting explored an exhaustive search of the following hyper-parameters:

- TFIDF thresholds: $\{0.1\%, 90\%, 99\%, 100\%\}$;
- Number of topics $k = \{1, 2, 3, ..., 200\}$;
- Number of top words per topic $n_w = 10, 20, 50$.

Reproducibility

Our method is an alternative software-based approach to the topic modeling problem. The ABT framework consists of software modules (i.e., Steps 1, 2 and 3) that work in tandem and can be freely parametrized and tuned.

Appendix A presents the Python code of our method. The complete code to reproduce the reported experiments is available at: `https://github.com/lealfp/labPLN/blob/master/notebooks/topic_modeling/ABT.ipynb`.

[6] `https://huggingface.co/fagner/envoy`

[7] `https://www.sbert.net/docs/pretrained_models.html`

4.4.3 Quantitative Analysis: Topics Coherence

Topics Coherence is a metric for automatically measuring the quality of a topic by measuring the interpretability of the topic model. The intuition is: a topic said to be coherent must contain words that are, in some way, associated [28]. Usually, this association is measured based on the Pointwise Mutual Information (PMI) of two randomly sampled words from the same sentence [47].

There are several metrics [1, 37] to evaluate topics such as Perplexity, Topic Diversity, Coverage, and Stability. We have chosen coherence because it is frequently used in the literature and there are implementations ready to use.

In our work, we have used the Gemsim[8] framework to measure topics' coherence [122]. We have used the c_v configuration which combines the indirect cosine measure with NPMI and Boolean Sliding Window. First, the top words W of a topic $t \in T$ are divided in the subsets W' and $W*$, according to:

$$S_{set}^{one} = \{(W', W^*) | W' = w_i; w_i \in W; W^* = W\} \tag{4.2}$$

The subset W^* is used to confirm the subset W' by comparing words to the total word set W. Next, probabilities are estimated to W' and W^* based on word co-occurrence counts using the Boolean Sliding Window approach, in which a window moves over the input Corpus C one word token per step. The j_{th} element of the context vector v_i of word w_i has NPMI:

$$v_{ij} = NPMI(w_i, w_j)^\gamma = \left(\frac{log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot (w_j)}}{-log(P(w_i, w_j) + \epsilon)} \right)^\gamma \tag{4.3}$$

The indirect cosine computes how strong the word set W^* supports W' comparing the similarity of the vectors $\vec{u}$ and $\vec{w}$ (computed with respect to the estimated probabilities), according to:

$$s_{cos}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{||\vec{u}_2|| \cdot ||\vec{w}||_2} \tag{4.4}$$

Table 4.5 shows the quality in terms of coherence of some (among fifty) topics produced in an experiment round of ABT. For instance, the best-ranked topic (row 1) exhibits coherence score $c = 0.81$ with some top words as {nose, on, ulcerated, ...}, while the worst topic (last row) achieves $c = 0.66$ with some top words as {arteritis, vasculitis, takayasu, ...}.

Table 4.6 relates the number of topics with the coherence obtained by using a selected set of language models: BERT, BioBERT, Envoy, BART and all-mpnet-base-v2. We tested various neural language models to investigate whether using them in ABT's first step affects the result, as each language model outputs a set of topics which gathers similar sentences according to the corresponding model. We are interested in observing the behavior of our method by applying language models calibrated by different training tasks: whether better-suited tasks produce more meaningful sets of topics.

[8] https://radimrehurek.com/gensim/models/coherencemodel.html

Table 4.5: Some example topics in descending sort of coherence scores. The most relevant terms describe each topic according to the cTFIDF metric.

		Top Words									
Topic	Coherence	1	2	3	4	5	6	7	8	9	10
1	0.81	nose	on	ulcerated	joints	enlarging	thumb	lip	pain	lump	nodule
2	0.79	sign	situation	ace	slips	baggage	restart	expander	unwanted	niti	aromatic
3	0.78	spotted	cat	malaria	altered	fever	status	adverse	poisoning	heart	eye
4	0.77	around	tale	just	all	not	more	what	one	double	tango
5	0.76	sarcina	fluke	celsi	lancisi	ganfort	eaten	chameleons	moth	koilonychia	fungating
	...						...				
11	0.73	being	sapho	mystery	light	eye	shining	earnest	cry	masquerade	visitor
12	0.72	tale	case	stone	rhinolith	poisoning	crying	two	dangers	horse	journey
13	0.72	genital	retinopathy	ocular	bull	mightier	jeweller	postendoscopy	teary	pence	argyrosis
14	0.72	pemphigus	gangrenosum	cutis	skin	erythema	vulgaris	vulvar	eruption	paraneoplastic	ichthyosis
15	0.71	tamponade	cardiac	hypothyroidism	sle	pancreatitis	triumph	stemi	presenting	pericardial	causes
	...						...				
36	0.68	febrile	cause	unexplained	cachexia	cyanosis	unusual	young	respiratory	neonate	premature
37	0.68	lymphohistiocytosis	haemophagocytic	langerhans	histiocytosis	cell	hlh	absent	radius	thrombocytopenia	familial
38	0.67	paediatric	population	heel	emergency	common	department	cause	thyroid	dogs	monodermal
39	0.67	malignancy	osseous	osteomalacia	hypercalcaemia	tumours	not	metaplasia	sarcoidosis	insulinoma	haematological
40	0.67	brachial	syndrome	compartment	neuropathy	calf	radiculopathy	plexus	longitudinally	claudication	quadriceps
	...						...				
46	0.66	myeloma	syphilis	presentation	atypical	paraneoplastic	multiple	syndrome	as	neurosyphili	typhoid
47	0.66	unusual	cause	foreign	mass	body	young	older	man	fossa	intraocular
48	0.66	haemangioma	capillary	sclerosing	period	haemangioblastoma	haemangioendothelioma	cavernous	multinodular	epithelioid	goitre
49	0.66	fasciitis	necrotising	thigh	limb	adenocarcinoma	patient	neutropaenia	dermatologic	traumatised	marine
50	0.66	arteritis	vasculitis	takayasu	vessel	medium	giant	retinal	as	cell	initial

Table 4.6: Topics Coherence (c) results achieved by ABT (using different neural language models) in modeling the CliCR corpus. Overall, c tends to stabilize around $k > 7$.

Number of Topics (k)	BERT	BioBERT	Envoy	BART	all-mpnet-base-v2
1	0.13	0.13	0.13	0.13	0.13
2	0.81	0.77	0.84	0.85	0.86
3	0.81	0.75	0.75	0.57	0.81
4	0.59	0.73	0.77	0.55	0.77
5	0.57	0.71	0.77	0.54	0.75
6	0.58	0.64	0.75	0.55	0.71
7	0.60	0.57	0.73	0.54	0.67
8	0.60	0.57	0.56	0.55	0.68
9	0.61	0.57	0.54	0.56	0.64
10	0.45	0.59	0.53	0.46	0.61
...			...		
50	0.60	0.60	0.61	0.61	0.62
...			...		
100	0.62	0.59	0.59	0.62	0.62
...			...		
150	0.63	0.60	0.61	0.62	0.61
...			...		
200	0.63	0.60	0.62	0.63	0.60

Each row in Table 4.6 shows the coherence achieved by the ABT when experimenting with the models standing in the columns. For instance, in the first row: all the models exhibit a similar coherence score $c = 0.13$ when producing 1 topic. By requesting $k = 50$ topics, the models show different performances, (e.g., BioBERT achieves $c = 0.60$ and Envoy $c = 0.61$). BioBERT scores $c = 0.60$ when producing $k = 200$ topics. These results are consistent with the non-monotonicity behavior [122] in growing topic sets in which the coherence of a given topic is increased by adding related sentences and decreased by adding non-related sentences.

These quantitative experiments showed that although models' performance is consistent across different language models, none completely outperforms the others regarding coherence results. However, as described in the next section, a qualitative analysis can reveal differences at the usage of different language models by inferring topics hierarchies.

4.4.4 Qualitative Analysis: Hierarchy Aggregation

The hierarchical aggregation method applied in ABT builds hierarchies (or partitions) of data that align with the statistical patterns observed by the language model used to infer the sentences representations. Such a tree-structured hierarchy directly results from the semantic vector space inferred by the language model. Therefore, the spatial arrangement of the sentences in the vector representation $\mathbb{R}^{768}$ directly determines the association between sentences and topics. In turn, the cTFIDF matrix encodes the associations between topics and terms by globally estimating the relevance of each vocabulary term within each topic.

For example, Figure 4.6 shows the hierarchical structure built by successively partitioning the sentence space inferred from the BERT model. In such a picture, we refer to each topic by its level on the hierarchy added by its index, for example $t_{3,1}$ ={doctor, fahr, my, ...} refers to the topic 1 at level 3.

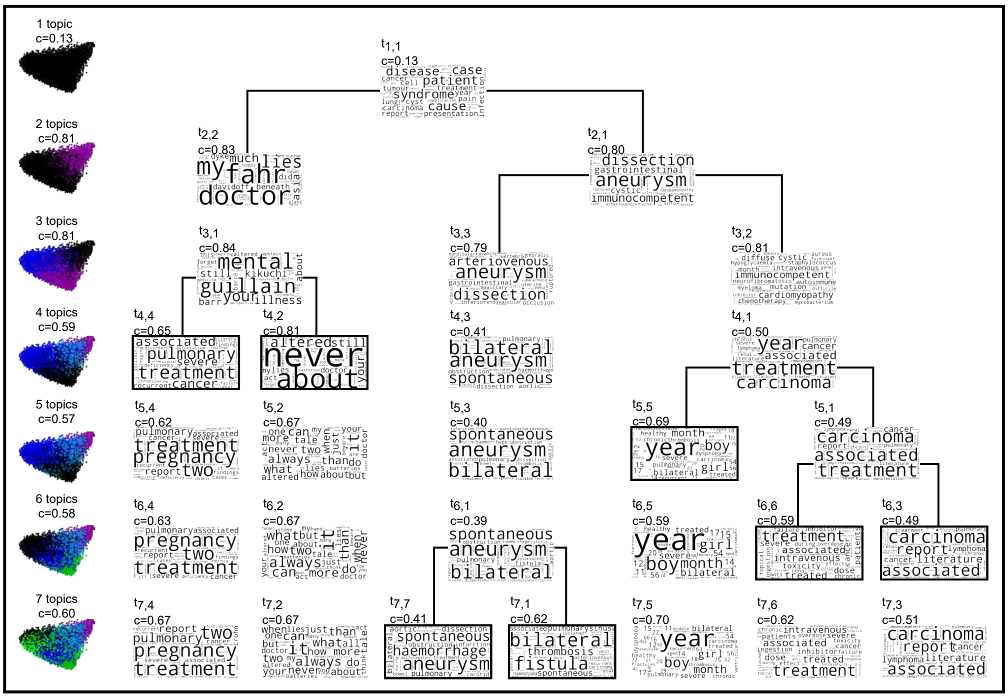

1 topic c=0.13
2 topics c=0.81
3 topics c=0.81
4 topics c=0.59
5 topics c=0.57
6 topics c=0.58
7 topics c=0.60
t1,1 c=0.13 disease case patient syndrome cause
t2,2 c=0.83 my fahr lies much doctor
t2,1 c=0.80 dissection aneurysm gastrointestinal immunocompetent
t3,1 c=0.84 mental guillain you illness barr
t3,3 c=0.79 arteriovenous aneurysm dissection
t3,2 c=0.81 diffuse cystic immunocompetent cardiomyopathy
t4,4 c=0.65 associated pulmonary treatment cancer
t4,2 c=0.81 altered still never about
t4,3 c=0.41 bilateral aneurysm spontaneous
t4,1 c=0.50 year cancer associated treatment carcinoma
t5,4 c=0.62 pulmonary treatment pregnancy two
t5,2 c=0.67 never two always
t5,3 c=0.40 spontaneous aneurysm bilateral
t5,5 c=0.69 year boy month
t5,1 c=0.49 carcinoma associated treatment
t6,4 c=0.63 pregnancy two treatment
t6,2 c=0.67 always can more
t6,1 c=0.39 spontaneous aneurysm bilateral
t6,5 c=0.59 year girl boy month
t6,6 c=0.59 treatment associated
t6,3 c=0.49 carcinoma report associated
t7,4 c=0.67 two pulmonary pregnancy treatment
t7,2 c=0.67 can it always
t7,7 c=0.41 spontaneous haemorrhage aneurysm
t7,1 c=0.62 bilateral fistula
t7,5 c=0.70 year boy month
t7,6 c=0.62 intravenous associated treatment
t7,3 c=0.51 carcinoma report literature associated

Figure 4.6: Top of the hierarchical structure produced through ABT using BERT as the language model in the first step of the pipeline. It is possible to observe some interesting patterns - such as, topics in the leaves are more specific than those near the root. Moreover, some fusion points are incurring large drops in measured coherence, which indicates essential points within the clustering process.

The hierarchy (Figure 4.6) reveals and organizes word categories particularly representative within the textual collection. Moreover, it visually evidences some latent aspects captured by such a particular model. Such patterns were hidden in the collection and discovered by the successive steps of aggregation. Each iteration (or level) of the hierarchical aggregation process discovers new latent patterns configuring a new topic.

The hierarchical structure is consistent with the property of non-monotonicity [122] in which the coherence of a given topic is increased by adding related sentences and must decrease by adding non-related sentences, as observed in the hierarchy's fusion points. For example, the fusion of topics $t_{7,7}$ and $t_{7,1}$ gives rise to topic $t_{6,1}$ having a lower value of coherence $c = 0.39$, which could indicate these sets of sentences are very unrelated. Conversely, the union of topics $t_{4,4}$ and $t_{4,2}$ increases the coherence to $c = 0.84$ and considerably changes the coherence of the remaining topics $t_{3,3}$ and $t_{3,2}$, which reflects on an overall increase of coherence from $c = 0.59$ to $c = 0.81$ by reducing from 4 to 3 topics. This indicates that in this case $k = 3$ is a well suited number of topics to the BERT vector space.

Interestingly, topic $t_{5,5}$ (and also its descendants) is formed by many numbers, which indicates that its sentences may contain numbers. Probably, at the moment to infer a representation for such sentences in Step 1, the BERT model regarded the presence of numbers as a discriminant feature in such sentences. Other models may pay attention to other features.

Departing from this supposition, next Figure 4.7 likens the hierarchies produced by other language models in order to indirectly observe possible features deemed as discriminant (i.e., latent aspects) by such models. The figure presents the hierarchical structures of the CliCR corpus according to the other 4 models: Envoy, BioBERT, BART and all-mpnet-base-v2. For simplification, the picture shows only the new topics produced in each aggregation step (i.e., one level on the hierarchy), omitting the remaining topics so as not to clutter the picture.

In these trees (Figures 4.6 and 4.7), it is possible to visualize some evident patterns, such as recurring concepts in certain tree branches – e.g., metastases in BioBERT, Envoy and all-mpnet-base-v2 models. In addition, good coherence assessments obtained in each model seem to correlate with a good number of clusters to separate the space of sentences; for example, in the BERT model, a separation of the space in 2 or 3 topics results in good coherence scores ($c = 0.81$); whereas in BART the coherence is lowered by going from 2 to 3 topics.

Moreover, topics containing numbers (e.g., $t_{5,5}, t_{6,6}$ and $t_{6,6}$) are recurring across the trees inferred by the following models, respectively: BERT, BioBERT and BART (Figures 4.6 and 4.7). This suggests these networks represent languages less domain-specific than the models (Envoy and all-mpnet-base-v2) that do not exhibit such generic topics. Besides, it can indicate outliers in the input data.

It is worth commenting on the topics produced by the Envoy model (Figure 4.7.a). For instance, almost all the topics in leaves (highlighted by borders, totaling 7 topics) seem to allude to anatomies, chemicals or diseases, except for topic $t_{4,4,}$ ={foreign, technique, approach, stent, endocarditis, hip, life, fixation, care, balloon} which stands for more generic terms than the other 6 leave topics (although also related to the biomedical context). We suppose that such topics arise as a natural reflection of the pre-training objective of the Envoy model, which consists in recognizing 4 different types of named en-

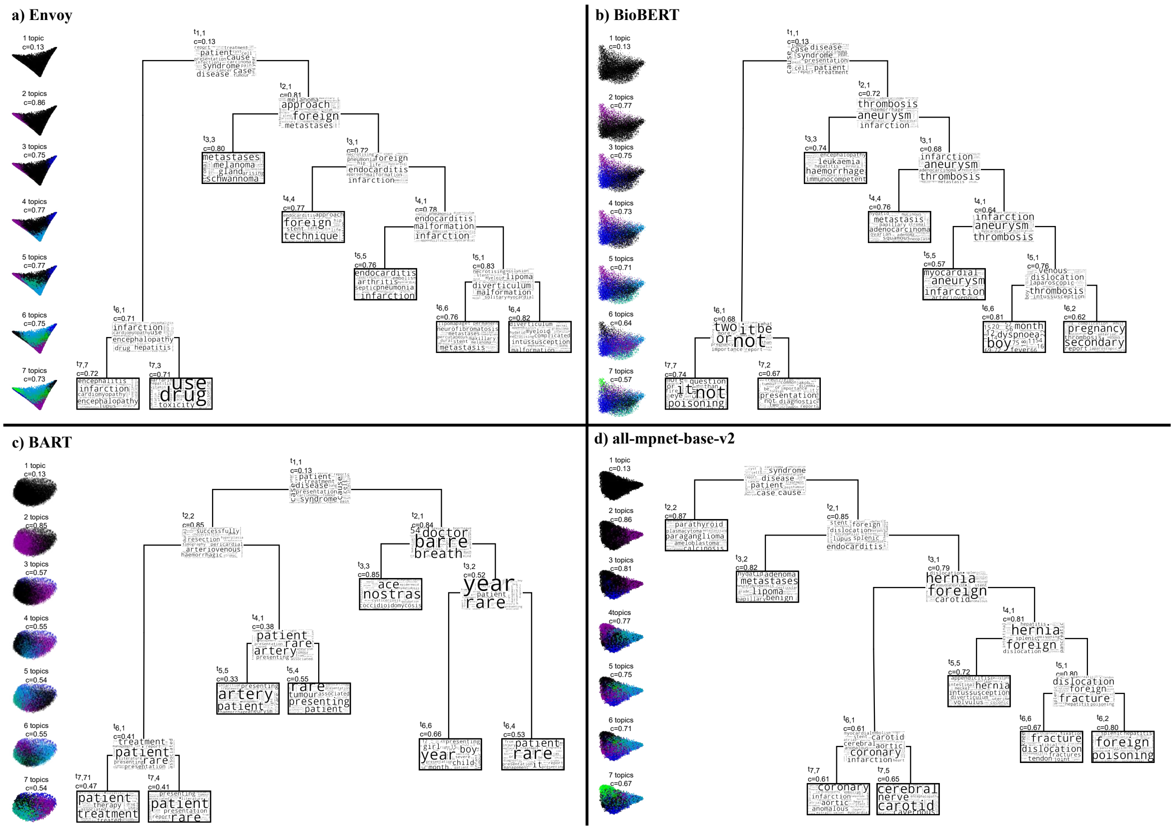

a) Envoy
1 topic c=0.13
2 topics c=0.86
3 topics c=0.75
4 topics c=0.77
5 topics c=0.77
6 topics c=0.75
7 topics c=0.73
$t_{1,1}$ c=0.13 patient cause syndrome case disease
$t_{2,1}$ c=0.81 approach foreign metastases
$t_{3,3}$ c=0.80 metastases melanoma gland schwannoma
$t_{3,1}$ c=0.72 foreign endocarditis infarction
$t_{4,4}$ c=0.77 foreign stent technique
$t_{4,1}$ c=0.78 endocarditis malformation infarction
$t_{5,5}$ c=0.76 endocarditis arthritis septic pneumonia infarction
$t_{5,1}$ c=0.83 lipoma diverticulum malformation
$t_{6,6}$ c=0.76 neurofibromatosis metastases metastasis
$t_{6,4}$ c=0.82 diverticulum myeloid complicated intussusception malformation
$t_{6,1}$ c=0.71 infarction cardiomyopathy use encephalopathy drug hepatitis
$t_{7,7}$ c=0.72 encephalitis infarction cardiomyopathy encephalopathy lupus
$t_{7,3}$ c=0.71 use drug toxicity
b) BioBERT
1 topic c=0.13
2 topics c=0.77
3 topics c=0.75
4 topics c=0.73
5 topics c=0.71
6 topics c=0.64
7 topics c=0.57
$t_{1,1}$ c=0.13 case disease syndrome presentation cell patient cause
$t_{2,1}$ c=0.72 thrombosis aneurysm infarction
$t_{3,3}$ c=0.74 leukaemia haemorrhage immunocompetent
$t_{3,1}$ c=0.68 infarction aneurysm thrombosis
$t_{4,4}$ c=0.76 metastasis adenocarcinoma
$t_{4,1}$ c=0.64 infarction aneurysm thrombosis
$t_{5,5}$ c=0.57 myocardial aneurysm infarction
$t_{5,1}$ c=0.76 venous dislocation laparoscopic thrombosis
$t_{6,6}$ c=0.81 month dyspnoea boy fever
$t_{6,2}$ c=0.62 pregnancy secondary
$t_{6,1}$ c=0.68 two it be or not
$t_{7,7}$ c=0.74 question it not poisoning
$t_{7,2}$ c=0.67 presentation diagnostic
c) BART
1 topic c=0.13
2 topics c=0.85
3 topics c=0.57
4 topics c=0.55
5 topics c=0.54
6 topics c=0.55
7 topics c=0.54
$t_{1,1}$ c=0.13 patient disease presentation syndrome case cause
$t_{2,2}$ c=0.85 successfully resection arteriovenous haemorrhagic
$t_{2,1}$ c=0.84 doctor barre breath
$t_{3,3}$ c=0.85 ace nostras coccidioidomycosis
$t_{3,2}$ c=0.52 year patient rare
$t_{4,1}$ c=0.38 patient rare artery presenting
$t_{5,5}$ c=0.33 artery patient
$t_{5,4}$ c=0.55 rare tumour presenting patient
$t_{6,6}$ c=0.66 girl boy year child month
$t_{6,4}$ c=0.53 patient rare it
$t_{6,1}$ c=0.41 treatment patient rare presentation
$t_{7,71}$ c=0.47 patient therapy treatment
$t_{7,4}$ c=0.41 patient presentation rare
d) all-mpnet-base-v2
1 topic c=0.13
2 topics c=0.86
3 topics c=0.81
4topics c=0.77
5 topics c=0.75
6 topics c=0.71
7 topics c=0.67
syndrome disease patient case cause
$t_{2,2}$ c=0.87 parathyroid plasmacytoma paraganglioma ameloblastoma calcinosis
$t_{2,1}$ c=0.85 stent foreign dislocation lupus splenic endocarditis
$t_{3,2}$ c=0.82 adenoma metastases lipoma benign
$t_{3,1}$ c=0.79 hernia foreign carotid
$t_{4,1}$ c=0.81 hernia foreign
$t_{5,5}$ c=0.72 hernia intussusception volvulus
$t_{5,1}$ c=0.80 dislocation foreign fracture
$t_{6,6}$ c=0.67 fracture dislocation tendon
$t_{6,2}$ c=0.80 foreign poisoning
$t_{6,1}$ c=0.61 carotid cerebral aortic coronary infarction
$t_{7,7}$ c=0.61 coronary infarction aortic anomalous
$t_{7,5}$ c=0.65 cerebral nerve carotid

Figure 4.7: A range of hierarchical structures drawn from the CliCR corpus encoded by four different language models. It is possible to observe that some trees exhibit a greater specialization in the language used in the training, such as Envoy and all-mpnet-base-v2.

tities: Chemical, Anatomy, Disease and Other. In this sense, the topic $t_{4,4}$ could be raised from a region in the vector space crowded with sentences containing generic terms classified as Other. Conversely, all the other topics are mixed by terms alluding to anatomies, chemicals or diseases. Therefore, the Envoy model deems the presence of terms alluding to such entities as discriminant features to infer the space vector $\mathbb{R}^{768}$ to represent the CliCR corpus.

In the Envoy model, we can observe that the root topic $t_{1,1}$ ={patient, syndrome, to, rare, case, as, unusual, acute, cause, disease} is more generic than the descent topics.

Furthermore, the ABT topic trees also exhibit the Hierarchical Affinity characteristic [78]: topics that descend from a topic t must be more similar to t than topics descendaing from other topics. For example, in all-mpnet-base-v2 (Figure 4.7.d), topic $t_{5,1}$ ={fracture, dislocation, foreign, hepatitis, poisoning, splenic, pancreatic, ...} is more similar to its children topics $t_{6,2}$ ={foreign, poisoning, hepatitis, splenic, pancreatitis, intravitreal, retinal, ...} and $t_{6,6}$ ={fracture, dislocation, tendon, fractures, knee, joint, fixation, ...} than to its non-children topic $t_{7,7}$ ={coronary, aortic, anomalous, infarction, myocardial, embolism, ...}.

Prior research has analyzed similar Hierarchical Structures [78, 71] to organize topics. Our analysis stems from [78], who proposed the Recursive Chinese Restaurant Process (rCRP), an algorithm for discovering hierarchical topic structures. Instead, our work produces a hierarchy by modeling the topic structure evidenced in the vector space $\mathbb{R}^{768}$, which is inferred by a Transformer-based neural language model. This approach enables one to design further analyses over the recognized patterns.

We argue that our approach can indicate the balance of a topic when it is split into subtopics. Let us consider two terms of similar prevalence in topic $t_{5,1}$ of Envoy: “lipoma” and “malformation”. The term “lipoma” seems to have a more balanced distribution between topics $t_{6,6}$ and $t_{6,4}$, dividing the force of its prevalence. The term “malformation” is unbalanced towards topic $t_{6,4}$ since it maintains a high prevalent in topic $t_{6,4}$ and does not appear as prevalent in topic $t_{6,6}$.

Despite the satisfactory performance achieved on coherence scores, we observed that the hierarchies structures produced did not hold the generic-to-specific property (i.e., Topic specialization) elicited by [78, 71]. We suppose that to generate hierarchical structures, our method would demand a sentence model (in Step 2) that discriminates sentences according to their generality and specificity. For example, one could experience with a hypernym-hyponym detection task to better separate the sentences in the vector space according to its generality degree. Such a task could be used to infer general-to-specific topic hierarchies. However, we suspect that a perfect separation would not be possible since generality and specificity are subjective matters, which would imply it is not held by the input text collection, in other words, generality hierarchy depends on semantics outside the text collection.

Besides, the topic coherence measured did not correlate with the evaluation performed by human evaluators about the quality of the topics. This instigates us to explore the following: Whether neural topic models demand metrics distinct from those used for assessing traditional topic models. According to our investigation studies, this misalignment could be associated with an issue already observed and addressed by [68]. This will be examined next.

4.4.5 Ablation Study

We argue that the non-conformity between topics coherence metrics and human evaluation – as originally claimed by [68] and reinforced by our results – is a reflection of the impossibility for a language model to learn semantic meaning by training it just on linguistic form. Such impossibility is grounded on the concept of meaning as the relationship between a linguistic form and something external to language [16] (i.e., not present within the training dataset).

Departing from this concept of meaning assumed by Linguistics, [16] proposed the Octopus Test to illustrate the challenges of learning meaning from form alone. According to their conceptual experiment, a language model trained purely in written texts cannot pass the Octopus Test because it is not capable, a priori, of capturing all the relevant features needed to produce utterances aligned with the evaluations full of semantic meaning of human judgments. Hence, neural language models inevitably fail the Octopus Test, and consequently in the Turing Test, since the Octopus Test is a weak form of the Turing Test.

Therefore, a topic model based on neural language models does not capture fundamental features to infer semantics because it is exposed only to linguistic form inside training.

Following these premises, we assume that the coherence measures of topics may not fully correlate with the generic-to-specific properties, because generality and specificity are subjective matters which demand human assumptions constructed from our cognitive mental "database", infinitely more complex than the binary database available to the text-based coherence evaluator. The mental database of the human evaluator has access to information (such as factual knowledge, intention, memories of smell, touch, social interactions, etc) beyond that contained in the data set used to train the language model.

Scholarly work on language acquisition suggests that human children do not learn meaning from form alone. Hence we should not expect machines to do so either [16]. As a consequence, a neural topic model trained just on textual form will not learn meaningful topics perfectly aligned with human evaluations that also embed semantics.

This limit imposed on machine learning methods to produce semantic topics does not negate the utility of topic models. Indeed, topic modeling may improve many other important NLP tasks. A particular case is the use of topic modeling as a dimensionality reduction technique. ABT can be seen as a summarization framework that reduces the asymptotic complexity of a collection to k dimensions – i.e., topics – coherent with the successive partitions performed throughout the clustering process. Future analyses can investigate how much discriminatory information it loses by reducing the collection to these dimensions.

We believe that future explorations, with more grouping levels, may elicit more topics that can be intuitively associated with themes and possibly choose those levels that group topics in the desired way. In addition, the semantic representations incorporated into attention-based topics can leverage information search systems. In future work, we intend to transform the hierarchical visual rendering presented in Figures 4.6 and 4.7 into an interactive sentence exploration system based on different levels of the hierarchical grouping of topics. This will enable an information search specialist to navigate at different topic levels.

4.5 Conclusions and Future Work

The topic modeling task aims to identify predominant themes in extensive textual collections. It has the potential to improve services and systems that work with textual data. This paper presents our approach to this problem. The ABT (Attention-Based Topics) model consists of a composition of modules that can be freely parametrized and tuned to automatically extract statistical semantics in corpora. The pipeline's modules work in tandem, expanding the possibilities of the LDA's latent topic generator probabilistic process.

In our approach, a topic is a dense region of sentences coupled with semantics, characterized by its most discriminant terms, benefiting from the statistical context emerging from the structure (the vector space) inferred by the neural network. The vector space represents semantics associated with the gist [62] of the sentence collection, extracted through a statistical inference performed by a neural language model. Grounded on this concept of semantics, ABT organizes topics in a hierarchical tree, which alludes to the semantics of a massive collection of sentences; such semantics are shared by the collective consciousness of co-authors of these texts.

We report on experiments using a range of neural language models, run on the CliCR corpus. Our study shows that the proposed approach adequately fits the problem and produces results comparable to the state-of-the-art concerning topics coherence metrics. Our discussion of results analyzes claims that neural language models cannot, in principle, learn the linguistic meaning, a vital component to complete the topic modeling task successfully and also to produce generic-to-specific hierarchies. Topics coherence metrics, consequently, do not align with the topics produced and/or evaluated by humans. Additionally, we analyzed the hierarchical topic structure built as a by-product of the inferring topics process. The coherence achieved at each level of such a hierarchy structure may indicate the most appropriate levels of the hierarchy to be chosen to summarize a textual collection.

In future work, we intend to formalize ABT and analyze its complexity. Moreover, we intend to extend our tests to conduct analyses of the method's behavior when working with the full content of clinical cases and with textual collections from other domains. This, however, would require additional computing resources and adjusting Sentence-BERT to handle sequences longer than the default 512-token limit. We also plan to integrate our method into pipelines of other more complex tasks. Yet another potential extension involves filtering non-relevant terms (e.g., stopwords) by defining a list of such terms tailored to each specific domain. Furthermore, validation can use other metrics – such as Perplexity. Another extension would be to experiment with datasets containing human-annotated rationales to enhance the quality of topics.

Acknowledgements

Work partially funded by projects FAPESP 2013/08293-7, CNPq #308018/2021-4 and CNPq #428459/2018-8. We thank the *Universidade Federal do Pará* for also supporting this research.

Chapter 5

Discussion

The recent advances in Computer Science provided the means to efficiently solve some historical problems in Natural Language Processing (NLP). Such improvements suggest that a new level of language understanding can be achieved by using attention-based models to capture the patterns that structure textual sentences. The appearance of attention-based language models can be stated as a revolution in NLP research, symbolizing a significant technological leap forward in this field.

As detailed in the literature review of chapter 2, the core of the NLP revolution is the realization of a common and unified computational representation to accomplish different language tasks. Such a unified representation is obtained as a by-product of the training of the network underlying the Neural Language Model.

The other two chapters approached Neural Language Models as NLP tools to produce new information from the statistical patterns observed in an input textual source, i.e., the topics which allude to themes addressed in a massive collection of texts (chapter 4) and the named entities which refer to semantic concepts (chapter 3).

Our analyses focused on the medical domain, although they can be modified to cover other domains. The analysis results reinforce the so called NLP revolution achieving F1 (at evaluating the named entities) and coherence (at evaluating the topics) measures on par with the state-of-the-art. Nevertheless, the resulting topics are not fully aligned with human evaluations regarding semantic meaning and generic-to-specific properties. We suppose this limit is a consequence of the impossibility [16] of language models to learn the complete semantics of the processed text. It demands different methods to better evaluate the results of the methods based on language models.

In the academic literature there is neither consensus about which type of analysis should be employed to evaluate language models. In fact, there are different perspectives regarding the actual advances achieved in a scientific context beyond computer science.

In the linguistic realm, there are researches arguing that language models are not, a priori, capable of understanding the **real meaning** (in a linguistic sense) [16] of processed texts, as they are trained only on textual forms (i.e., the linguistic signal), and therefore, the current language models cannot pass in the Turing test. According to this, there is a portion of meaning attributed to extra-textual information not present in the training set.

On the other hand, Sahlgren and Carlsson [124] argue that if meaning produces effects on form, then a language model should at least be able to observe and learn these effects. The paper of [49] claims that, although GPT models lack mechanisms of consciousness from a cognitive science perspective, they have already passed the Turing test and therefore can successfully imitate human language capabilities.

This debate addresses issues that have historically been studied in various research areas. Therefore, it is need to analyze the results obtained by this so-called NLP revolution with caution, as it raises expectations and interests from different actors in society—companies, states, political groups, and even the public at large.

The difficult to analyze the language models demands more research to investigate the extension to which they can imitate the human language capabilities. The lack of consensus about evaluation of language models is perhaps not surprising, since neural networks are examples of Complex Systems [128] and therefore, are essentially holistic and interdisciplinary. Thus, neural networks for language models would be machines "as complex as the systems they model and therefore they will be equally difficult to analyse" [128]. This situation resembles the difficulty of validating models based on a relativist, holistic philosophy of science [12]. By such approach the "The criterion of practical use has taken the place of formal rigor [...] validation becomes a semiformal, conversational process" than "a matter of formal accuracy". Therefore, the emergence of neural language models demands research and development of new validation methods to assess their capacities, considering their holistic and interdisciplinary nature.

In our work, we evaluated the language models at producing sets of topics to summarize the themes covered in a text collection. We observed the language models lack important information to perfectly infer the semantics from textual collections. We assume that the topics coherence measured may not fully correlate with the human evaluations because topic quality is a subjective matter which demands human assumptions constructed from the human cognitive mental "database", infinitely more complex than the binary database available to the text-based coherence evaluator. The mental database of the human evaluator has access to information (such as factual knowledge, intention, memories of smell, touch, social interactions, etc) beyond that contained in the data set used to train and evaluate the language model.

This limit imposed on machine learning methods does not deny the utility of neural language models. The Language models open up vast possibilities to unprecedented applications. Indeed, language models may improve many important human tasks.

Accordingly it is crucial to implementing best practices for developing open-source Neural Language Models. It helps prevent the complexity [9] of this new technology from being used to mask biases and interests. For example, the financial market speculation (which has led to higher stock prices for AI technologies) which was exposed with the release of Deepseek [21], a cheaper, smaller (i.e., fewer parameters), and open-source model. The arrival of such a Chinese model with fewer parameters call into question the term used to label this research field — Large Language Models — since it was not the model size that revolutionized this branch of research, but the dense and distributed aspect of Transformer models. This is why we refer to it as Neural Language Models instead of Large Language Models, this label wrongly directs the focus to a matter of size.

Chapter 6

Conclusions and Future Perspectives

This thesis concentrated on analysis of Neural Language Models applied to problems of linguistic nature, applied to a set of case studies involving clinical data. It was organized based on a set of publications, as defined by the Institute of Computing of Universidade Estadual de Campinas. This research covers intersections among many research areas, in particular involving some of the topics covered by machine learning and generative AI - and thus computational linguistics and statistics. Moreover, given the nature of the case studies, there was the need to work with domain experts in the field of clinical data analysis.

Clinical texts contain sensitive, private data, which would imply additional pre-processing for ethical reasons. In our research, this does not apply. First, we used only fictitious cases from the Harena repository. In ABT tests, we point out that CliCR is an open anonymized data set and hence the problem does not apply either.

Our work tackled some problems related to the processing of textual big data, involving the extraction of the embodied semantics and its incorporation in machine-interpretable data formats. The thesis was concerned with two research questions:

(**RQ1**): How to encode structured semantics directly into textual constructions?

(**RQ2**): How to reveal and extract the latent semantics which is kept hidden underlying the statistical patterns occurring in textual constructions?

To deal with these research questions we leveraged Semantic Representations to provide computers with means to handle semantic meaning. Research Question 1 (RQ1) is addressed through ontologies, as it concerns so-called external semantics. Conversely, the second Research Question (RQ2) is tackled using statistical semantics to uncover the latent, or inner, semantics within texts.

We point out that the introduction of artificial intelligence everywhere can impact the way we work, relate, learn, and develop. Therefore, there is a need for education at all levels aimed at teaching people how to use, understand, develop, and consume these tools in a healthy manner. Considering the breadth of the impacts that neural language models can have on human life, a pedagogical project is needed to guide towards a sustainable and ethical use of neural language models that also serve to address real and widely discussed societal problems, rather than solely serving the economic interests of the few who hold and dominate this technology and perform financial speculation over their closed-source products.

6.1 Main contributions

Chapter 2 presented a summary of the various associated concepts needed to design and implement our algorithms and develop our research. Departing from this conceptual basis, we analyzed language models applied to two research problems: Semantic Annotations and Topic Modeling, which are related to the main contributions of this thesis.

Chapter 3 presents our first three contributions (Envoy, Harena Semantics and the Semantic Virtual Patient), being related to RQ1, and Chapter 4 concerns the fourth contribution (Attention-based Topics), being focused on the answer RQ2.

Contributions to Semantic Annotations Our first contribution is a Named Entity Recognition (NER) method – which we call *Envoy* – to recognize and classify biomedical concepts inside the input texts. In the best setting, Envoy achieved results comparable to the state-of-the-art NER models at the time (2020).

The Envoy model is invoked by the second contribution – the *Harena Semantics* – in order to annotate the clinical cases drawn from the Harena data repository, described in the Introduction to this thesis. Harena Semantics combines the biomedical entities (recognized by Envoy) with ontology concepts to produce the third contribution: the *Semantic Virtual Patient*, a semantic representation that models clinical cases as a network of concepts linked to the Semantic Web.

Results of a pilot evaluation suggest the framework's potential to engage users in the collaborative creation of semantic resources. A differential of our approach is the superimposition of human-made annotations with the machine-inferred annotations automatically suggested by the framework.

Contributions to Topic Modeling Our fourth contribution is an approach to the topic modeling problem, called ABT. It is a method that infers a set of topics from an input textual collection, according to three successive steps that perform a series of transformations over the input collection. The implementation of the ABT model consists of a composition of modules that can be freely parameterized and tuned.

In our approach, a topic is a dense region of sentences coupled with a statistical semantic, represented by its most discriminant terms, benefiting from statistical context emerging from the structure (the vector space) inferred by the language model. Grounded on this concept of semantics, ABT organizes topics in a hierarchical tree, which alludes to the semantics of a massive collection of sentences; such semantics are supposed to be shared by the collective consciousness of co-authors of the input texts.

The inferred output topics are organized in a hierarchical structure. We have evaluated our approach over the CliCR dataset containing clinical case reports using the Topics Coherence metric to measure the quality of the inferred topics. Our experiments used a range of neural language models, including our Envoy model (see above contribution) showing that the proposed approach adequately fits the problem and produces results comparable to the state-of-the-art concerning topics coherence measure. The coherence achieved at each level of such a hierarchy structure may indicate which levels of the hierarchy should be chosen to better summarize the input textual collection.

Additionally, we proposed a theoretical discussion on the issue of non-conformity between topics coherence and human evaluations. Our hypothesis is that language models cannot, in principle, learn the linguistic meaning, which is a vital component to successfully complete the topic modeling task – as discussed, among others, by Bender et al., [15]. This might be the reason why the extracted measure of topics coherence does not align with the topics produced and/or evaluated by human agents.

6.2 Extensions

Our work has both theoretical and practical extensions.

Our case study is centered on the health domain and clinical case studies. However, our methods are generic enough to deal with data from other domains; hence, our work can be extended to other datasets and branches of research such as Chemistry, Physics, Education, Social Sciences, and others. Indeed, there are a vast amount of processes that can be modeled as a language, and therefore, amenable to use within the Transformer architecture.

ABT, Harena Semantics and Envoy showcase how neural language models can be used to compound diverse complex tasks. Such methods have the potential to optimize a variety of services and systems that work with domain textual data. For instance, in a collaborative filtering problem to recommend movies, a user and the movies would be analogous to a document and the words in the document, respectively [22].

In yet another extension, the topics produced by ABT can be leveraged as features in other downstream tasks, such as Discourse Analysis, Content Recommendation, Summarization and Search Engines. ABT can thus be leveraged into pipelines of other more complex tasks.

From a theoretical point of view, ABT should be formalized and its complexity be analyzed, in particular formally comparing it to LDA. This will require defining similarities and differences in order to identify strengths and weaknesses of each method. Furthermore, there should be investigation into ways to improve ABT with topics from LDA and vice-versa – e.g., to analyze whether the LDA method can be used to fine-tune the pre-trained language models.

Another possible direction is to experiment with other tasks to finetune the sentence model. For example, one can experience with a hypernym-hyponym detection task to better discriminate sentences according to their generality and specificity. Such a task could be used to infer general-to-specific topic hierarchies, extending the experiments of this thesis.

Another possibility of improving ABT is to involve metadata such as keywords, category labels and semantic annotations. Besides, there should be experiments to conduct analyses of the method's behavior when working with larger textual fragments, such as paragraphs, books and documents in general.

While ABT was conceived to work with textual corpora, it can be adapted to work on extraction of topics from collections of other kinds of media – such as audio and video to produce more meaningful topics. This would require initial preprocessing of such media – e.g., by sound-to-text transformations.

Yet another possible direction would be to experiment with Brazilian Portuguese LLM implementations, such as Sabiá [117], Tucano [41], Seriema [131], Capivara [51]. A possible target application might be to analyze texts concerning tropical, regional diseases. Furthermore, the employment of multimodal data to align images and captions – as investigated by Santos et al., [51] – could lead to improvements in NER and topic modeling.

Our evaluation of ABT can also be extended by, for instance, adopting other evaluation metrics, such as Perplexity, Hierarchical Affinity and Topic Specialization. Also, further investigation is needed on the incoherence of coherence [68]. Another possibility is to extrinsically evaluate ABT at performing other tasks, such as Summarization, Dimensionality Reduction and Content Recommendation.

The topic modeling problem might also be analyzed as an instance of a Complex System. The related work has already approached the problem as a bipartite network of words and documents linked by the topics [59]. The ABT topic hierarchy can be transformed into a network, which can then be analyzed to see whether it is adherent to the properties of complex systems, such as Zipf's law [92].

ABT can also be extended to help the user figure out the "best" k number of topics. Hierarchical clustering produced by ABT can also be combined with ontologies, extending the topic's semantic scope.

ABT topics could also be considered as part of the Harena annotation framework. Conversely, annotations might be used to help topic generation. The annotations of Virtual Patient are stored with the document. An extension in this sense might be to create a specific store of annotations, to allow future semantic queries that would take such annotation store into account.

Harena Semantics can be embedded into a search engine to retrieve clinical case reports aided both by the semantic annotations and by topic models.

It is also possible to extend our Envoy model of Named Entity Recognition to identify more types of biomedical entities (beyond those we used, namely Anatomy, Chemical and Disease), including entities in the Brazilian Portuguese language. Besides, our approach can be further evaluated using other metrics such as Interpretability.

Results of fine-tuning a model have costs, no matter how light the model. Thus, another extension would be to check the impact of adding or decreasing layers in model execution, in particular for Envoy.

Additional future work, in particular for the clinical domain, would be to consider multimodal data sets (e.g., images).

A possible additional investigation concerns checking evidence of linguistic semantics within the Vector Space produced by language models. An evidence would be, for instance, an aspect of the space that clearly reflects the language semantic properties – such as the proximity of the sentences within the semantic space investigated in this study.

Given the interdisciplinary nature of language models, there is a need for an alignment of concepts and technical terms between the NLP research field and other fields, to leverage the real scientific potential of language models under an interdisciplinary and complex perspective of sciences. For example, understanding and comprehension of terms should be standardized according to the linguistic perspective [16] in academically-oriented publications.

Bibliography

[1] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. Topic modeling algorithms and applications: A survey. *Information Systems*, pages 1–17, 2023.

[2] Omri Abend and Ari Rappoport. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 77–89, 2017.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774v6*, pages 1–100, 2023.

[4] Liliya Akhtyamova, Paloma Martínez, Karin Verspoor, and John Cardiff. Testing contextualized word embeddings to improve ner in spanish clinical case narratives. *IEEE Access*, pages 164718–164726, 2020.

[5] Mohammed AlQuraishi. Alphafold at CASP13. *Bioinformatics*, pages 4862–4865, 2019.

[6] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323v3*, pages 1–7, 2019.

[7] Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470v1*, pages 1–25, 2020.

[8] Meysam Asgari-Chenaghlu, Mohammad-Reza Feizi-Derakhshi, Leili farzinvash, Mohammad-Ali Balafar, and Cina Motamed. Topicbert: A cognitive approach for topic detection from multimodal post stream using bert and memory–graph. *Chaos, Solitons & Fractals*, pages 1–13, 2021.

[9] Julian Assange. *Cypherpunks: liberdade e o futuro da internet*. Boitempo Editorial, 1st edition, 2015. pages 1–148.

[10] Unesco Assembly. Ethics of artificial intelligence – a unesco recommendation. *unesco*, 2021. `www.unesco.org/en/artificial-intelligence/recommendation-ethics`.

[11] Parikshit Bansal and Amit Sharma. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766v1*, pages 1–16, 2023.

[12] Yaman Barlas and Stanley Carpenter. Philosophical roots of model validation: Two paradigms. *System Dynamics Review*, pages 148–166, 1990.

[13] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 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*, pages 238–247, 2014.

[14] Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. Cometa: a corpus for medical entity linking in the social media. *arXiv preprint arXiv:2010.03295v2*, pages 3122–3137, 2020.

[15] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[16] Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198, 2020.

[17] Yoshua Bengio. Neural net language models. *Scholarpedia*, 2008. `http://www.scholarpedia.org/article/Neural_net_language_models`.

[18] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, pages 1137–1155, 2000.

[19] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, pages 34–43, 2001.

[20] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. *OpenAI*, pages 1–8, 2023. `https://cdn.openai.com/papers/dall-e-3.pdf`.

[21] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954v1*, pages 1–48, 2024.

[22] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[23] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, pages 135–146, 2017.

[24] Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports*, pages 1–16, 2025.

[25] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of German Society for Computational Linguistics*, pages 1–11, 2009.

[26] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326v1*, pages 1–11, 2015.

[27] Jordan Boyd-Graber, Yuening Hu, and David Mimno. Applications of topic models. *Foundations and Trends® in Information Retrieval*, pages 1–154, 2017.

[28] Jordan Boyd-Graber, David Mimno, and David Newman. *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*. Chapman and Hall/CRC, 1st edition, 2014. pages 225–254.

[29] Tudor Călinici and Valentin Muntean. Open labyrinth – a web application for medical education using virtual patients. *Applied Medical Informatics*, pages 15–20, 2010.

[30] Ewen Callaway. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, pages 203–205, 2020. `https://www.nature.com/articles/d41586-020-03348-4`.

[31] Leonardo Campillos-Llanos, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. Hybrid natural language processing tool for semantic annotation of medical texts in spanish. *BMC bioinformatics*, pages 1–39, 2025.

[32] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559v1*, pages 1–7, 2020.

[33] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, pages 1–45, 2024.

[34] Timothy Chaplin, Brent Thoma, Andrew Petrosoniak, Kyla Caners, Tamara McColl, Chantal Forristal, Christa Dakin, Jean-Francois Deshaies, Eliane Raymond-Dufresne, Mary Fotheringham, David Ha, Nicole Holm, James Huffman, Ann-Marie

Lonergan, George Mastoras, Michael O'Brien, Marie-Rose Paradis, Nicholas Sowers, Errol Stern, and Andrew K. Hall. Simulation-based research in emergency medicine in canada: priorities and perspectives. *Canadian Journal of Emergency Medicine*, pages 103–111, 2020.

[35] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174, 2016.

[36] Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. Ultra-fine entity typing. *arXiv preprint arXiv:1807.04905v1*, pages 1–10, 2018.

[37] Rob Churchill and Lisa Singh. The evolution of topic modeling. *ACM Computing Surveys*, pages 1–35, 2022.

[38] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

[39] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364v5*, pages 670–680, 2017.

[40] David A Cook and Marc M Triola. Virtual patients: a critical literature review and proposed next steps. *Medical Education*, pages 303–311, 2009.

[41] Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. Tucano: Advancing neural text generation for portuguese. *Patterns*, page 101325, 2025.

[42] Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, pages 1–14, 2017.

[43] Eleni Dafli, Panagiotis Antoniou, Lazaros Ioannidis, Nicholas Dombros, David Topps, and Panagiotis D Bamidis. Virtual patients on the semantic web: a proof-of-application study. *Journal of medical Internet research*, pages 1–18, 2015.

[44] Tiago de Araujo Guerra Grangeia, Bruno de Jorge, Daniel Franci, Thiago Martins Santos, Maria Silvia Vellutini Setubal, Marcelo Schweller, and Marco Antonio de Carvalho-Filho. Cognitive load and self-determination theories applied to e-learning: impact on students' participation and academic performance. *PloS one*, pages 1–21, 2016.

[45] Marcos Felipe de Menezes Mota, Fagner Leal Pantoja, Matheus Silva Mota, Tiago de Araujo Guerra Grangeia, Marco Antonio de Carvalho Filho, and André Santanchè. Analytical design of clinical cases for educational games. In *Joint International Conference on Entertainment Computing and Serious Games*, pages 353–365, 2019.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[47] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, pages 439–453, 2020.

[48] Carla Geovana do Nascimento Macário. *Anotação semântica de dados geoespaciais.* PhD thesis, University of Campinas, Brazil, 2009. pages 1–126.

[49] Gordana Dodig-Crnkovic. How gpt realizes Leibniz's dream and passes the turing test without being conscious. In *Computer Sciences & Mathematics Forum*, pages 1–6, 2023.

[50] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, pages 1–10, 2014.

[51] Gabriel Oliveira dos Santos, Diego A. B. Moreira, Alef Iury Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia Da Silva, Esther Colombini, Helio Pedrini, and Sandra Avila. Capivara: Cost-efficient approach for improving multilingual clip performance on low-resource languages. *arXiv preprint arXiv:2310.13683*, 2023.

[52] Rezarta Islamaj Doğan and Zhiyong Lu. An improved corpus of disease mentions in pubmed citations. In *Proceedings of the Workshop on Biomedical Natural Language Processing*, pages 91–99, 2012.

[53] Emad Elwany, Dave Moore, and Gaurav Oberoi. Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. *arXiv preprint arXiv:1911.00473v1*, pages 1–4, 2019.

[54] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 55–65, 2019.

[55] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, pages 34–48, 2020.

[56] Sajjad Fouladvand, Jeffery Talbert, Linda P Dwoskin, Heather Bush, Amy Lynn Meadows, Lars E Peterson, Ramakanth Kavuluru, and Jin Chen. Predicting opioid use disorder from longitudinal healthcare data using multi-stream transformer. *arXiv preprint arXiv:2103.08800v2*, pages 1–10, 2021.

[57] Karen M Freeman, Scott F Thompson, Eric B Allely, Annette L Sobel, Sharon A Stansfield, and William M Pugh. A virtual reality patient simulation system for teaching emergency response skills to US navy medical providers. *Prehospital and Disaster medicine*, pages 3–8, 2001.

[58] Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, pages 4291–4308, 2021.

[59] Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. A network approach to topic models. *Science Advances*, pages 1–11, 2018.

[60] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913v2*, pages 1–12, 2020.

[61] Ben Goertzel. *Chaotic logic: Language, thought, and reality from the perspective of complex systems science*, volume 9. Springer Science & Business Media, 2013. pages 1–278.

[62] Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. Topics in semantic representation. *Psychological review*, pages 211–244, 2007.

[63] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794v1*, pages 1–10, 2022.

[64] Inga Hege, Andrzej A Kononowicz, Martin Adler, et al. A clinical reasoning tool for virtual patients: design-based research study. *JMIR medical education*, pages 1–11, 2017.

[65] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, pages 37–45, 1997.

[66] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556v1*, pages 1–36, 2022.

[67] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146v5*, pages 1–12, 2018.

[68] Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. Is automated topic model evaluation broken? The incoherence of coherence. *Advances in neural information processing systems*, pages 2018–2033, 2021.

[69] Chip Huyen. Evaluation metrics for language modeling. *The Gradient*, 2019. `https://thegradient.pub/understanding-evaluation-metrics-for-language-models/`.

[70] Jeannie Y Irwin, Henk Harkema, Lee M Christensen, Titus Schleyer, Peter J Haug, and Wendy W Chapman. Methodology to develop and evaluate a semantic representation for nlp. In *AMIA Annual Symposium Proceedings*, pages 271–275, 2009.

[71] Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. Tree-structured neural topic model. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 800–806, 2020.

[72] Corey Brian Jackson, Carsten Østerlund, Kevin Crowston, Mahboobeh Harandi, and Laura Trouille. Shifting forms of engagement: Volunteer learning in online citizen science. *Proceedings of the ACM on Human-Computer Interaction*, pages 1–19, 2020.

[73] Clement Jonquet, Nigam H Shah, Cherie H Youn, Mark A Musen, Chris Callendar, and Margaret-Anne Storey. Ncbo annotator: semantic annotation of biomedical data. In *8th International Semantic Web Conference, Poster and Demo Session (ISWC)*, pages 1–3, 2009.

[74] Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics*, pages 1–21, 2020.

[75] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. Moral concerns are differentially observable in language. *Cognition*, pages 1–12, 2021.

[76] Tom Kenter, Alexey Borisov, and Maarten De Rijke. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640v1*, pages 1–11, 2016.

[77] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, pages 73729–73740, 2019.

[78] Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 783–792, 2012.

[79] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. *Advances in neural information processing systems*, pages 1–9, 2015.

[80] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102v2*, pages 1–19, 2020.

[81] Miroslav Kubat. *An introduction to machine learning*. Springer, 2nd edition, 2017. pages 1–348.

[82] Atharva Kulkarni, Amey Hengle, Pradnya Kulkarni, and Manisha Marathe. Cluster analysis of online mental health discourse using topic-infused deep contextualized representations. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 83–93, 2021.

[83] Manish Kumar, Pardeep Singh, and Poonam Kashtriya. Enhanced biomedical named entity recognition using spacy and bert models. *Procedia Computer Science*, pages 1954–1961, 2025.

[84] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, pages 1234–1240, 2020.

[85] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461v1*, pages 1–10, 2019.

[86] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, pages 1–10, 2016.

[87] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, pages 31–57, 2018.

[88] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, pages 1–14, 2023.

[89] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025v5*, pages 1–11, 2015.

[90] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, pages 657–723, 2024.

[91] Yufeng Lyu and Jiang Zhong. Dsmer: A deep semantic matching based framework for named entity recognition. In *European Conference on Information Retrieval*, pages 419–432, 2021.

[92] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing.* The MIT Press, 2nd edition, 1999. pages 1–720.

[93] James H Martin and Daniel Jurafsky. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 3rd draft edition, 2019. pages 1–613.

[94] J Massot. How named entity recognition and document comprehension unlock geosciences and engineering semantic search without big data. In *First EAGE Digitalization Conference and Exhibition*, pages 1–5, 2020.

[95] Elijah Mayfield and Alan W Black. Should you fine-tune bert for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–12, 2020.

[96] Kris McGuffie and Alex Newhouse. The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807v1*, pages 1–12, 2020.

[97] Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, and Elena Tutubalina. Drug and disease interpretation learning with biomedical entity representation transformer. *European Conference on Information Retrieval*, pages 1–15, 2021.

[98] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, pages 1–12, 2013.

[99] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pages 1–9, 2013.

[100] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[101] Alistair Miles and Sean Bechhofer. Skos simple knowledge organization system reference. *W3C*, 2009. `https://www.w3.org/TR/skos-reference/`.

[102] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272, 2011.

[103] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884v3*, pages 1–19, 2020.

[104] Tahsir Ahmed Munna, Filipe Cunha, António Leal, Ricardo Campos, and Alípio Jorge. Human experts vs. large language models: Evaluating annotation scheme and guidelines development for clinical narratives. *Proceedings of the Text2Story'25 Workshop*, pages 149–160, 2025.

[105] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, pages 1–20, 2007.

[106] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, pages 1–8, 2001.

[107] Joel Nothman, Hanmin Qin, and Roman Yurchak. Stop word lists in free open-source software packages. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 7–12, 2018.

[108] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, pages 170–173, 2009.

[109] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507v3*, pages 1–13, 2017.

[110] Fagner Leal Pantoja, Marco Antonio de Carvalho Filho, and André Santanchè. Harena semantics: A framework to support semantic annotation in citizen science systems. In *5th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF)*, pages 336–343, 2022.

[111] Fagner Leal Pantoja, André Santanchè, and Claudia Bauzer Medeiros. A bibliographic survey of neural language models with applications in topic modeling and clinical studies. *Technical Report - Instituto de Computação, Universidade Estadual de Campinas*, pages 1–26, 2024. `https://www.ic.unicamp.br/~reltech/2024/24-01.pdf`.

[112] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1310–1318, 2013.

[113] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[114] Nicole Peinelt, Dong Nguyen, and Maria Liakata. tbert: Topic models and bert joining forces for semantic similarity detection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7047–7075, 2020.

[115] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[116] ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, pages 1–15, 2018.

[117] Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240. Springer Nature Switzerland, 2023.

[118] Sampo Pyysalo and Sophia Ananiadou. Anatomical entity mention recognition at literature scale. *Bioinformatics*, pages 868–675, 2014.

[119] Mingjie Qiu, Wenzhong Yang, Fuyuan Wei, and Mingliang Chen. A topic modeling based on prompt learning. *Electronics*, pages 1–16, 2024.

[120] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, pages 1–12, 2018.

[121] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, pages 1–11, 2019.

[122] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, 2015.

[123] Sebastian Ruder. *Neural transfer learning for natural language processing.* PhD thesis, NUI Galway, 2019. pages 1–311.

[124] Magnus Sahlgren and Fredrik Carlsson. The singleton fallacy: Why current critiques of language models miss the point. *Frontiers in Artificial Intelligence*, pages 1–9, 2021.

[125] André Santanchè, Heitor Soares Mattosinho, Marcos Felipe De Menezes Mota, Fagner Leal Pantoja, Gabriel De Freitas Leite, Ana Claudia Tonelli, Fernando Salvetti Valente, Juliana De Castro Solano Martins, Sandro Queirós, Tiago De Araujo Guerra Grangeia, and Marco Antonio de Carvalho Filho. Virtual patient platform and data space for sharing, learning, discussing, and researching. In *IEEE 19th International Conference on e-Science (e-Science)*, pages 1–10, 2023.

[126] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100v4*, pages 1–73, 2023.

[127] Guus Schreiber and Yves Raimond. Rdf 1.1 primer. *W3C*, 2014. `https://www.w3.org/TR/rdf11-primer/`.

[128] Stephen K Scott. *Approaching complexity.* Faraday Discussions, 1995. pages 1–156.

[129] Terrence J Sejnowski. Large language models and the reverse turing test. *Neural computation*, pages 309–342, 2023.

[130] Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, 2021.

[131] Fillipe dos Santos Silva, Julio Cesar dos Reis, and Marcelo S Reis. Seriema: A framework to enhance clustering stability, compactness, and separation by fusing multimodal data. In *International Conference on Applications of Natural Language to Information Systems*, pages 394–408, 2024.

[132] Ziyang Song, Yuanyi Hu, Aman Verma, David L Buckeridge, and Yue Li. Automatic phenotyping by a seed-guided topic model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4713–4723, 2022.

[133] Ziyang Song, Xavier Sumba Toral, Yixin Xu, Aihua Liu, Liming Guo, Guido Powell, Aman Verma, David Buckeridge, Ariane Marelli, and Yue Li. Supervised multi-specialist topic model with applications on large-scale electronic health record data. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–26, 2021.

[134] Sainbayar Sukhbaatar, Edouard Grave, Guillaume Lample, Herve Jegou, and Armand Joulin. Augmenting self-attention with persistent memory. *arXiv preprint arXiv:1907.01470v1*, pages 1–11, 2019.

[135] Simon Šuster and Walter Daelemans. Clicr: a dataset of clinical case reports for machine reading comprehension. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1551–1563, 2018.

[136] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 1–9, 2014.

[137] Akihiro Takeuchi, Tomomi Kobayashi, Minoru Hirose, Takashi Masuda, Toshiro Sato, and Noriaki Ikeda. Arterial pulsation on a human patient simulator improved students' pulse assessment. *Journal of Biomedical Science and Engineering*, pages 285–289, 2012.

[138] Noha S Tawfik and Marco R Spruit. Evaluating sentence representations for biomedical text: Methods and experimental results. *Journal of biomedical informatics*, pages 1–10, 2020.

[139] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805v5*, pages 1–90, 2023.

[140] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *arXiv preprint arXiv:1905.05950v2*, pages 1–9, 2019.

[141] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316v1*, pages 1–17, 2019.

[142] Laure Thompson and David Mimno. Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626v1*, pages 1–12, 2020.

[143] Christos Thrampoulidis. Implicit bias of next-token prediction. *arXiv preprint arXiv:2402.18551v2*, pages 1–25, 2024.

[144] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, pages 1–13, 2024.

[145] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971v1*, pages 1–27, 2023.

[146] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394, 2010.

[147] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, pages 1–11, 2017.

[148] Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, et al. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *Journal of biomedical informatics*, pages 11–19, 2018.

[149] Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222v3*, pages 1–21, 2020.

[150] Gabriella Vigliocco and David P Vinson. Semantic representation. *The Oxford handbook of psycholinguistics*, pages 195–215, 2007.

[151] Ruohan Wang, Zilong Wang, Ziyang Song, David Buckeridge, and Yue Li. Mixehrnest: Identifying subphenotypes within electronic health records through hierarchical guided-topic modeling. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–8, 2024.

[152] Xiaobao Wu, Fengjun Pan, Thong Nguyen, Yichao Feng, Chaoqun Liu, Cong-Duy Nguyen, and Anh Tuan Luu. On the affinity, rationality, and diversity of hierarchical topic modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19261–19269, 2024.

[153] Yonghui Wu. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144v2*, pages 1–23, 2016.

[154] Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? A comprehensive evaluation from deductive, inductive and abductive views. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–37, 2023.

[155] Shweta Yadav, Usha Lokala, Raminta Daniulaityte, Krishnaprasad Thirunarayan, Francois Lamy, and Amit Sheth. "When they say weed causes depression, but it's your fav antidepressant": Knowledge-aware attention framework for relationship extraction. *PloS one*, pages 1–18, 2021.

[156] Hanqi Yan, Lin Gui, and Yulan He. Hierarchical interpretation of neural text classification. *Computational Linguistics*, pages 987–1020, 2022.

[157] He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498v1*, pages 1–8, 2021.

[158] Shuxin Zhou, Hao Liu, Pritam Sen, Yehoshua Perl, and Mahshad Dehkordi. Cfc annotator: A cluster-focused combination algorithm for annotating electronic health records by referencing interface terminology. In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF)*, pages 195–206, 2025.

Appendix A

Source Code for Chapter 4

This appendix presents the Python codes for the pipeline of ABT (A.1) and for the validation by topic coherence (A.2).

A.1 Pipeline

```
corpus = []
vocabulary = []
topics = []
topics_coherence = 0.0

class Topic:
        def __init__(self, index):
        self.index = index
        self.sentences = ""
        self.words = []
        self.coherence = 0.0

DATASET_FILE = "clicr-cases-titles.txt"
with open(DATASET_FILE) as f:
        corpus = f.readlines()


### STEP 1
from sentence_transformers import SentenceTransformer

BERT_MODEL = "bert-base-cased"
language_model = SentenceTransformer(BERT_MODEL)
vector_space = language_model.encode(corpus)


### STEP 2
DISTANCE_THRESHOLD =300
```

```
from sklearn.cluster import AgglomerativeClustering

clustering_model = AgglomerativeClustering(distance_threshold=
    ↪ DISTANCE_THRESHOLD, n_clusters=None, linkage="ward")
clustering_model = clustering_model.fit(vector_space)
k = clustering_model.n_clusters_

topics = [Topic(i) for i in range(k)]

for i in range(len(corpus)):
        inferred_cluster_index = clustering_model.labels_[i]
        topics[inferred_cluster_index].sentences += corpus[i] + "␣"


### STEP 3
MAX_DF = 0.1

import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer

 if (k > 1):
        c_MAX_DF = 0.5
        c_tfidf_model = TfidfVectorizer(max_df=c_MAX_DF)
        c_tfidf = c_tfidf_model.fit_transform([topic.sentences for
            ↪ topic in topics])
        c_tfidf_matrix = c_tfidf.toarray()
        words = c_tfidf_model.get_feature_names_out()

        for i, topic in enumerate(topics):
            sorted_term_indexes = np.argsort(-1*c_tfidf_matrix[topic.
                ↪ index])
            topic.words = [words[j] for j in sorted_term_indexes]
 else:
        c_tfidf_model = TfidfVectorizer(max_df=MAX_DF)
        c_tfidf = c_tfidf_model.fit_transform(corpus)
        c_tfidf_matrix = c_tfidf.toarray()
        words = c_tfidf_model.get_feature_names_out()

        mean_tfidf = np.array(c_tfidf_matrix.mean(axis=0)).flatten()
        sorted_term_indexes = np.argsort(-1*mean_tfidf)
        topics[0].words = [words[j] for j in sorted_term_indexes]
```

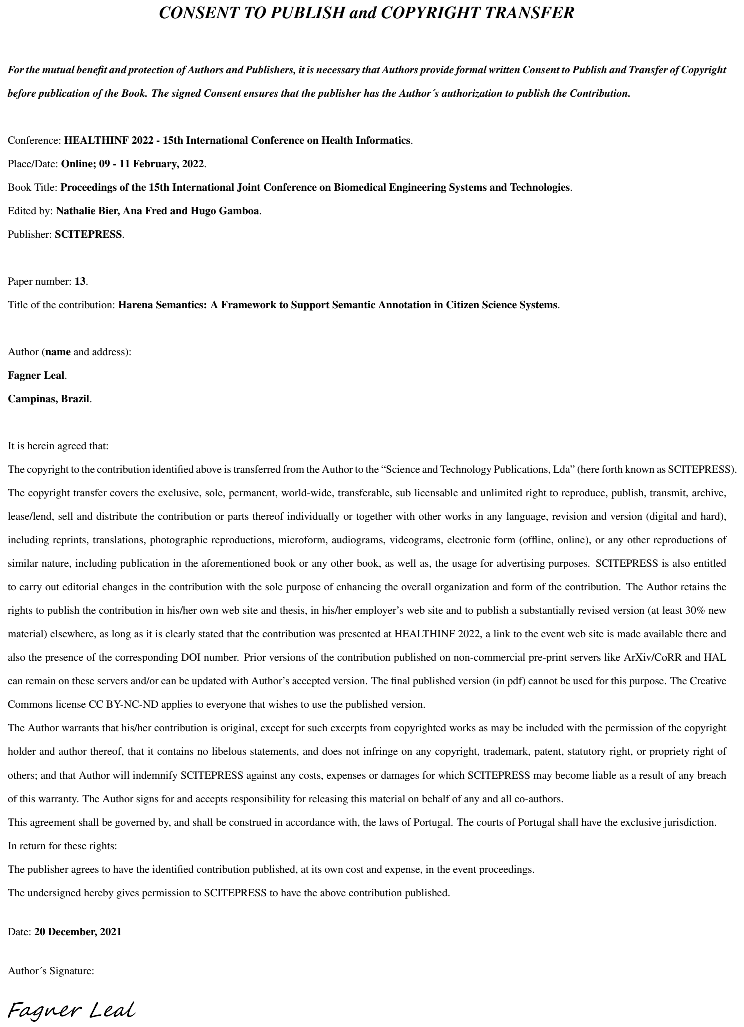
A.2 Validation

```
TOP_WORDS = 50
```

```
from gensim.models import CoherenceModel
from gensim.corpora import Dictionary

tfidf_model = TfidfVectorizer(max_df=MAX_DF)
tfidf = tfidf_model.fit_transform(corpus)
vocabulary = tfidf_model.get_feature_names_out()
terms_by_sentence = tfidf_model.inverse_transform(tfidf)

dictionary = Dictionary(terms_by_sentence)
cm = CoherenceModel(topics=[topic.words[:TOP_WORDS] for topic in
    ↪ topics], texts=terms_by_sentence, dictionary=dictionary,
    ↪ coherence='c_v', topn=TOP_WORDS)
topics_coherence = cm.get_coherence()
coherence_per_topic = cm.get_coherence_per_topic()

for i, coherence in enumerate(coherence_per_topic):
        topics[i].coherence = coherence
        print("Topic␣", str(i))
        print(topics[i].words[:TOP_WORDS])

print("Total␣coherence:␣", cm.get_coherence())
print("Coherence␣by␣topic:␣", cm.get_coherence_per_topic())
```

Appendix B

Publisher Authorization



CONSENT TO PUBLISH and COPYRIGHT TRANSFER

For the mutual benefit and protection of Authors and Publishers, it is necessary that Authors provide formal written Consent to Publish and Transfer of Copyright before publication of the Book. The signed Consent ensures that the publisher has the Author´s authorization to publish the Contribution.

Conference: **HEALTHINF 2022 - 15th International Conference on Health Informatics**.
Place/Date: **Online; 09 - 11 February, 2022**.
Book Title: **Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies**.
Edited by: **Nathalie Bier, Ana Fred and Hugo Gamboa**.
Publisher: **SCITEPRESS**.

Paper number: **13**.
Title of the contribution: **Harena Semantics: A Framework to Support Semantic Annotation in Citizen Science Systems**.

Author (**name** and address):
Fagner Leal.
Campinas, Brazil.

It is herein agreed that:

The copyright to the contribution identified above is transferred from the Author to the "Science and Technology Publications, Lda" (here forth known as SCITEPRESS). The copyright transfer covers the exclusive, sole, permanent, world-wide, transferable, sub licensable and unlimited right to reproduce, publish, transmit, archive, lease/lend, sell and distribute the contribution or parts thereof individually or together with other works in any language, revision and version (digital and hard), including reprints, translations, photographic reproductions, microform, audiograms, videograms, electronic form (offline, online), or any other reproductions of similar nature, including publication in the aforementioned book or any other book, as well as, the usage for advertising purposes. SCITEPRESS is also entitled to carry out editorial changes in the contribution with the sole purpose of enhancing the overall organization and form of the contribution. The Author retains the rights to publish the contribution in his/her own web site and thesis, in his/her employer's web site and to publish a substantially revised version (at least 30% new material) elsewhere, as long as it is clearly stated that the contribution was presented at HEALTHINF 2022, a link to the event web site is made available there and also the presence of the corresponding DOI number. Prior versions of the contribution published on non-commercial pre-print servers like ArXiv/CoRR and HAL can remain on these servers and/or can be updated with Author's accepted version. The final published version (in pdf) cannot be used for this purpose. The Creative Commons license CC BY-NC-ND applies to everyone that wishes to use the published version.

The Author warrants that his/her contribution is original, except for such excerpts from copyrighted works as may be included with the permission of the copyright holder and author thereof, that it contains no libelous statements, and does not infringe on any copyright, trademark, patent, statutory right, or propriety right of others; and that Author will indemnify SCITEPRESS against any costs, expenses or damages for which SCITEPRESS may become liable as a result of any breach of this warranty. The Author signs for and accepts responsibility for releasing this material on behalf of any and all co-authors.

This agreement shall be governed by, and shall be construed in accordance with, the laws of Portugal. The courts of Portugal shall have the exclusive jurisdiction.

In return for these rights:

The publisher agrees to have the identified contribution published, at its own cost and expense, in the event proceedings.

The undersigned hereby gives permission to SCITEPRESS to have the above contribution published.

Date: **20 December, 2021**

Author´s Signature:

Fagner Leal