UCLouvain
Faculté de philosophie, arts et lettres – FIAL

Faculty of Philosophy, Arts and Letters

A Study of Sentence-Level Simplification of French Medical Texts

Author: **Antonella FADDA**
Supervisor: **Thomas FRANÇOIS**
Cosupervisor: **Rémi CARDON**

Academic year 2024–2025
Master [120] in Linguistics, Major in Natural Language Processing

Université catholique de Louvain

Faculty of Philosophy, Arts and Letters

A Study of Sentence-Level Simplification of French Medical Texts

Author: Antonella FADDA

Supervisor: Thomas FRANÇOIS

Cosupervisor: Rémi CARDON

Master [120] in Linguistics, Major in Natural Language Processing

Academic year 2024–2025

Declaration of Honour / Use of AI Tools

Dissertation
Faculty of Philosophy, Arts and Letters, UCLouvain

I hereby declare that this is an original and personal work, that all referenced sources have been fully cited, regardless of their origin, and that the use of artificial intelligence tools complies with the general instructions published by UCLouvain and available at the following address:

https://cdn.uclouvain.be/groups/cms-editors-p1/portail/reglement/Consignes-ChatGPT-version-etudiante.pdf?itok=SebXXm87 (French version)

I am aware that failure to cite a source, to cite it clearly and fully, to announce in a specific section at the beginning of the dissertation the use I have made of artificial intelligence tools, or to use them in a manner contrary to the guidelines set forth in the institutional note, constitutes serious irregularities in accordance with the *Academic Regulations and Procedures* (Chapter 4, section 7, articles 107 to 114) within the University. In particular, I am aware of the risks of academic and disciplinary sanctions.

In light of the above, I declare on my honour that I have not committed plagiarism or any other form of fraud and that I have not used artificial intelligence tools beyond what is permitted by UCLouvain. In the course of preparing this dissertation, I made use of artificial intelligence tools, specifically ChatGPT by OpenAI (August 13, 2025 version) for linguistic assistance as English is not my native language.

Antonella FADDA

August 13, 2025

Acknowledgments

I would first like to thank Thomas François, my thesis supervisor, for his availability and his professionalism.

I would also like to express my sincere gratitude to Rémi Cardon, my co-supervisor, who followed me step by step, offering valuable advice and moral support during the most challenging moments. His ability to encourage me helped me to stay motivated throughout this journey.

I would also like to thank the CENTAL for giving me the opportunity to complete my internship with them. It was a rich experience that allowed me to learn and grow professionally.

To my family, and especially my parents, I owe enormous gratitude for always believing in me and for their unconditional support.

Finally, thank you to my friends for always being there, and to Martina and Dan, who stood by me and supported me during the revision of my work.

To all the students who have felt illegitimate,
who have questioned whether they belong,
or whether they are enough.

This work is for you.

Table des matières

Chapitre 1

Introduction

The process of comprehending a text is often underestimated and taken for granted, even though it involves multiple skills and is an indispensable part of our daily lives. The complexity of a text can make a text incomprehensible to a significant portion of the population. Indeed, "16,7% of a population, on average, cannot understand text that goes beyond a basic vocabulary" (Štajner et al., 2022, p. 2). This leads a substantial part of the population to face challenges in "mak[ing] informed decisions regarding critical processes such as healthcare choices, legal representation, education, or democratic rights" (Štajner et al., 2022, p. 2). As a result, initiatives with this objective continue to emerge, with the aim of developing resources that simplify overly complex language. Initiatives like Simple English Wikipedia [1], FALC [2] and the *Handbook of easy languages in Europe*(Lindholm and Vanhatalo, 2021) illustrate how this can make a difference. The issue has been debated for decades and, despite the challenges of automatic implementation, it still engages a portion of the scientific community.

Text Simplification (TS) refers to the process of modifying a text in order to

1. `https://simple.wikipedia.org/wiki/Main_Page`
2. https://www.falc.be/

make it simpler for the reader and thus easier to understand. It is a field relevant for a multitude of social groups ; among others : second-language learners, children, and people with aphasia or dyslexia. The majority of TS methodologies requires extensive corpora ; however, one of the main problems of TS is the lack of data. This limitation is exacerbated for languages other than English and even more so for domain-specific contexts. Generally speaking, a corpus for text simplification is a parallel corpus composed of pairs of sentences, one being the simplification of the other. Creating such resources is not a trivial task, and it is in this context that the field of Semantic Textual Similarity (STS) provides tools and methodologies.

The objective of this Master's Thesis is to explore how a hybrid methodology could improve the research on similar sentences in order to create a parallel corpus in the medical field.

Only one french corpus that contains medical data exists : The CLEAR corpus(Grabar and Cardon, 2018). This work addresses this gap by developing and evaluating a hybrid methodology for French STS that combines sentence embeddings from large French language models (CamemBERT, Sentence-CamemBERT-Large) with sets of classic similarity measures, in the aim of exploring the extension of this resource with state-of-the-art methodologies.

This thesis is structured into four main parts. In the Background section (2), we first explore the main approaches and techniques of Text Simplification, followed by those of Semantic Textual Similarity. This section is important for understanding the domain in which the Methodology (presented in Section 3) is situated. The results obtained (Section 4) and their discussion (Section 5) are presented next, and finally, the thesis concludes with a summary of the findings.

Chapitre 2

Background

2.1 Text Simplification

Text Simplification (TS) is the task of modifying the content and structure of a text in order to make it easier to read and understand, while retaining its main idea and approximating its original meaning (Alva-Manchego et al., 2020). When simplifying, taking into consideration the target audience is a crucial initial step in this process. Indeed, the focus of the simplification can vary : Aphasic people, for example, encounter problems when reading. This can be due to lexical or syntactic nature, as they could have problems recalling terms they do not often use or particular structures can be difficult to understand (Carroll et al., 1998). The medical domain provides another example. The simplification of medical texts primarily targets an audience less familiar with intricate medical vocabulary. Other groups that are often considered for simplification encompass children, second-language learners, or people with dyslexia (North et al., 2023a).

It is important to note that simplification can occur as a preliminary step to further text manipulation. In fact, text simplification can be employed to facilitate summarization or information extraction(Chandrasekar et al., 1996) (Saggion and

Hirst, 2017).

Understanding what makes a text complex is not always simple. The concept of simplicity is intuitive, but difficult to define (Shardlow, 2014). To explain this phenomenon, let us consider the length of a sentence. A longer sentence is generally more difficult to understand. Indeed, Simple English Wikipedia follows the rule of using simpler language and shorter sentences. This is not always true, especially if a paraphrase or a definition is used to replace complex vocabulary. This concept is explained in the following example :

1. *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.*
2. *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou à ralentir le transit intestinal.*

The word *péristaltisme* has been replaced by *transit intestinal* and *inhibant* has been replaced by *destinés à bloquer ou à ralentir* ; creating a longer but simpler sentence. The concept of simplicity is therefore always bound up with that of complexity, since in textual simplification the goal is to obtain a simpler text than the original one.

Moreover, Shardlow (2014) in *Survey of Automated Text Simplification*, explains that the concepts of readability and comprehensibility are pivotal, closely related concepts. The author states that a text that is easier to read generally becomes more understandable as readers find it more manageable to navigate intricate concepts. Similarly, an easily comprehensible text motivates readers to persist even through challenging readability issues.

Thus, understanding what makes a text complex is not always clear, as multiple factors could intervene at the same time. The complexity of texts hinges significantly

on both the readers and the intended simplification purpose. Defining the target audience is a crucial initial step in this process. For instance, the simplification of medical texts primarily targets an audience less familiar with intricate medical vocabulary. The reasons are varied, and, as previously noted, it is essential to consider the specific group for whom the text is being simplified.

It is worth noting that we can simplify text either manually or through automated methods. Manual methods, however, are expensive and not cost-effective, as new sources continue to be published and their simplification would need to keep pace (Saggion and Hirst, 2017). The next section will focus on the approaches to Automatic Textual Simplification (Section 2.1.1), starting from Lexical Simplification and Syntactic simplification, and finally End-to-End Approaches (2.1.2)

2.1.1 Approaches to Simplification

The reasons behind the efforts to simplify, as well as the intended recipients of these simplifications, play a crucial role in determining the choice of methodology or combination of methodologies.

Depending on the aspect we want to simplify, we will use lexical simplification or syntactic simplification. In this section we will discuss each of these in detail.

Lexical Simplification

Lexical simplification focuses on difficult words. These words are replaced using alternatives that are considered easier to understand (Saggion and Hirst, 2017). Lexical simplification unfolds in a series of steps : it begins with the identification of complex words, followed by the generation of substitutes for each intricate term. The process then involves word sense disambiguation, which considers the context

in which the word is used, narrowing down the appropriate words to be considered. Finally, synonym ranking is conducted, arranging them from the simplest to the most intricate (Shardlow, 2014).

The following is an example of lexical simplification from *Automatic Text Simplification* (Saggion and Hirst, 2017) :

1. John composed these verses in 1995.
2. John wrote the poem in 1995.

So, the first action is to identify *composed* and *verses* as complex words, generate candidates and finally choose the best alternative, in this case *wrote* and *poem*.

Each of the stages in the lexical simplification pipeline constitutes a separate area, which has its own limitations and challenges. Three classic approaches have long been used to identify which words are complex :

1. simplify everything approach
2. threshold-based approach
3. lexicon-based approach

The first approach aims to simplify all words and thus not identify which words are complex and which are not. Its main limitation is clear : all words are simplified, even those that are already simple by nature. These are replaced with other words that are simple but less precise than the original ones, thus changing the meaning of the sentence, often producing meaningless or ungrammatical sentences (North et al., 2023b). The second approach is based on choosing a specific measure that is used to determine the threshold that dictates the complexity of a word. Word length is often used as a measure of complexity. However, this approach is ineffective because it fails to recognize complex words that are short in length. Finally, using an approach based on a list of complex words is advantageous when limiting oneself to a narrow domain or particular audience. An example of this is FACILITA (Watanabe

et al., 2010), a resource in Brazilian Portuguese specifically tailored for children with low literacy levels in Brazil that distinguishes complex words. The same resource, however, when used for other categories (second-language learners or people who have reading disabilities...), is less helpful (North et al., 2023b).

The models that work most effectively today for Lexical complexity prediction are transformer-based models. These, indeed, have "several advantages [...], namely their self-attention mechanism and their ability to more effectively capture long-term dependencies" (North et al., 2023b).

To accomplish the second action, WordNet is used as a resource to locate a list of synonyms. Let's take another example from Saggion and Hirst (2017) :

1. "My automobile is broken"

The synonyms are identified : *car*, *auto*, *machine*, and *motorcar*. They are then ranked using the *Kučera-Francis frequency count* (Saggion and Hirst, 2017) that gives *car* as the best replacement. Following this approach, Shardlow (2015) pointed out possible problems that might be encountered in the pipeline.

The disambiguation step is crucial in order to ensure that the original sense of the sentence has been preserved (Štajner et al., 2022).

An alternative avenue within lexical simplification lies in the approach of paraphrasing. As defined by Shardlow (2015), a paraphrase signifies "a pair of phrases with the same meaning, where each phrase could have one or more words." Paraphrases can be acquired through manual identification or automatically extracted from a comparable corpus. Complex terms and their corresponding paraphrases find their place in a specialized dictionary.

Syntactic simplification

The objective of Syntactic Simplification is to try and identify the syntactic structures that can impair comprehension and readability, in order to modify them to create a simpler to understand text (Saggion and Hirst, 2017). Possible transformations include converting passive sentences into active ones or simplifying subordinate clauses. Siddharthan, Advaith (2014) (p. 264) citing O'Brien (2003) adds other rules, for example "specify rules for pre and post-modifier usage, rule out ellipsis, insist on the use of article or demonstrative, restrict size of noun cluster and rule out specific prepositions, such as "of ", specify location of prepositions to reduce ambiguity, avoid use of present participle, rule out passive voice, insist on indicative mood, restrict apposition, rule out certain forms of conjunction, specify use of punctuation." The following is an example of syntactic simplification from *Automatic Text Simplification* (Saggion and Hirst, 2017) :

1. Hu Jintao, who is the current Paramount Leader of the People's Republic of China, was visiting Spain.
2. Hu Jintao was visiting Spain. Hu Jintao is the current Paramount Leader of the People's Republic of China.

In the first sentence, the relative clause "who is the current Paramount Leader of the People's Republic of China" is embedded in the main clause. This structure is more complex because it mixes the main action ("visiting Spain") with the information on Hu Jintao. The simplification separates the information into two sentences, making it easier to understand.

To complete this kind of simplification, there is a need for an experienced linguist who can extract from the texts the patterns that need to be simplified. A first approach is rule-based, which in fact relies on the knowledge of the linguist who develops rules to apply to the texts for simplification. A second approach would be

to extract these patterns through corpora (Saggion and Hirst, 2017).

2.1.2 End-to-End Text Simplification

An end-to-end simplification system is a methodology that directly transforms complex sentences or texts into simple ones in a single step. In fact, this type of methodology does not divide the process into separate lexical or syntactic simplification steps. These types of systems are the most widely used today (Kew et al., 2023)(Alva-Manchego et al., 2020) and use end-to-end neural networks to directly obtain simplified versions of the input text (Garbacea et al., 2021). These approaches are used in contrast to the lexical simplification approaches proposed by Shardlow (2014), for example.

The rise of these models is certainly due to the expansion of transformer-based architecture (Vaswani, 2017), but unfortunately, building this type of model remains a challenge. In this regard, (Martin et al., 2022) proposed a Multilingual Unsupervised Sentence Simplification system to solve the biggest problem in this area : the lack of parallel corpora for simplification. The methodology used consists of extracting sequences from CCNet (Wenzek et al., 2019) and tokenising them. In order to create the paraphrase corpus using LASER (Artetxe and Schwenk, 2019), n-dimensional embeddings are computed, which are then indexed to find the nearest neighbor through *faiss*. The results are then filtered. The corpus thus created is used to fine-tune mBART (Lewis et al., 2019), the multilingual version of BART (Liu et al., 2020), a pretrained sequence- to-sequence model. MUSS uses ACCESS to incorporate control tokens. In English and French, it achieves state-of-the-art results.

In 2023 Kew et al. (2023) published BLESS a *Benchmarking Large Language Models on Sentence Simplification*. The framework examines a wide range of models and evaluates them on commonly used simplification datasets(Section 2.1.3) such as ASSET (Alva-Manchego et al., 2020), MED-EASI (Basu et al., 2023) and Newsela (Xu

et al., 2015). Overall, the study analyses 44 LLMs with different sizes, architectures and training objectives. The authors apply in-context learning (ICL) to evaluate the models in a few-shot context, experimenting with three different types of prompts. As a baseline, they also include the MUSS model (Martin et al., 2022).

For the evaluation, simplicity is measured using SARI (Xu et al., 2016), meaning preservation with BERTScore (Zhang et al., 2020), and readability with the FKGL index (Kincaid et al., 1975). In addition, a qualitative analysis is conducted on 300 outputs produced by the models to verify whether each sentence is a valid example of simplification and, if so, to identify the type of simplification performed. In general, the results show a fairly stable balance between simplicity and preservation of meaning. Operations such as lexical simplification, paraphrasing and deletion of parts of text are performed more often, while structural interventions, such as splitting sentences or changing the order of elements, are less frequent. The analysis shows that all LLMs perform significantly better on general ASSET and Newsela texts. In contrast, MED-EASI has the lowest values for meaning preservation and the least variety of transformations.

Simplification of Medical Texts

Medical literature is characterized by an abundance of specific language that can be challenging for those outside the profession to understand.

The concept of health literacy assumes a prominent role, defined as "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" (Davis and Wolf, 2004, p. 595).

The repercussions of low medical literacy among the public are significant and cannot be underestimated. Indeed, there are studies dating back as far as two decades

such as (Davis and Wolf, 2004, 596) that link low medical literacy "to problems with using preventive services, delayed diagnoses, understanding one's medical condition, adherence to medical instructions, and self-management skills."

However, they are not the only researchers in this area. In fact, studies continue to confirm these conclusions. For example, Berkman et al. (2011, p. 103) arrived at similar results, highlighting that "low health literacy is also associated with differential use of certain health care services, including increased hospitalizations and emergency care and decreased mammography screening and influenza immunizations". Another alarming finding is that this phenomenon is inherent in all segments of the population, cutting across them. This underscores the urgent need for everyone to have access to understandable sources in order to improve the situation, which unfortunately does not appear to be progressing (Davis and Wolf, 2004)(Berkman et al., 2011).

An example of lexical simplification applied to medical literature can be found in Figure 2.1, and it clearly shows how difficult such operation can be, since many types of changes or simplifications could be applied.

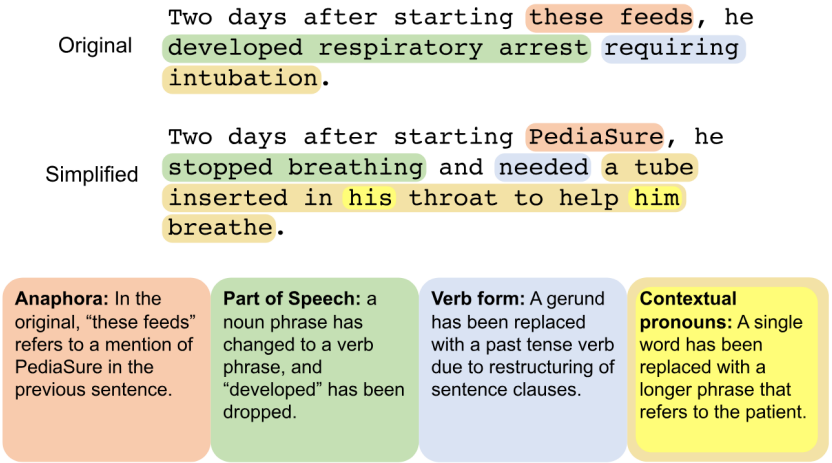

FIGURE 2.1 – Example of biomedical Text Simplification (Ondov et al., 2022)

2.1.3 Datasets and Need for Resources

The lack of resources for text simplification is particularly accentuated outside of English. Innovative approaches to text simplification often rely on language-specific resources, which unfortunately are particularly scarce. This thesis deliberately shifts the focus to the French language, a conscious choice based on the recognition that this language lacks these crucial resources, diminishing the application of advanced methods.

In almost all publications of Text Simplification the lack of data-sets and corpora is underlined. As seen in Section 2.1.2, new methodologies have arisen to solve this problem. Ondov et al. (2022) explains not only how data-driven and neural methods require parallel corpora, but also that this problem is intensified for specialized domains.

Corpus for Textual Simplification

Definition (Gena, 2010, p. 12) defines a Corpus as "a principled collection of authentic texts stored electronically that can be used to discover information about language that may not have been noticed through intuition alone." She also defines the existence of eight types of corpora :

— generalized,
— specialized,
— learner,
— pedagogic,
— historical,
— parallel,
— comparable,
— monitor.

We distinguish corpora by the way they were created : automatically, semi-automatically or manually ; and their composition : comparable or parallel.

Comparable Corpora. A comparable corpus consists of texts in two interrelated languages that discuss the same topics and themes, but which are not direct translations of each other. In the context of textual simplification, comparable corpora are not in two different languages , but in the same language ; however, they have two different lexicons because one uses a more complex language than the second. For example, using Wikipedia and Simple Wikipedia already mentioned above, a comparable corpus can be formed.

Parallel corpora. The concept of a parallel corpus, conversely, revolves around texts or sentences that exist as translations of one another—or, in the context of simplification, where one text represents the simplified version of the other. As aptly noted, "In order to build such parallel corpora, the typical approach is to start from comparable corpora and extract sentence pairs that share the same meaning". (Cardon and Grabar, 2020b, p. 44). The utility of parallel corpora lies in their inherent simplicity, facilitating the seamless realization of content alignment. However, the intricacies manifest during their creation, where the extraction of parallel sentences introduces challenges and is not immune to errors. This aspect is better explained in Section 2.2.

2.1.4 Evaluation

As mentioned earlier, the simplification of a text is intrinsically linked to the domain and audience for which the text is being simplified. This not only implies the need for specific resources, but also an evaluation that takes into consideration the characteristics listed above.

Grabar and Saggion (2022) propose two types of evaluation : human and automatic.

As for human evaluation this takes into consideration three criteria : semantics (to assess whether the meaning is unchanged), grammaticality (to assess whether the sentence is grammatically correct and understandable) and simplicity (to assess whether the simplified text is simpler than the original one) Grabar and Saggion (2022). The main problem posed by human ratings is that the ratings are subjective and the guidelines that are offered to annotators are vague, creating reproducibility problems Grabar and Saggion (2022).

Measures such as precision and accuracy, derived from comparison of the reference text and simplification, were used to assess simplification. A higher score indicates better quality in simplification.

Other measures come from the area of textual similarity or from the area of machine translation. In the former case they are highlighted by Grabar and Saggion (2022)

1. EditNTS
2. SeqLabel

In the second case we have :

1. BLEU *(bilingual evaluation understudy)*
2. TERp *(Translation Edit Rate plus)*
3. OOV *(out of vocabulary)*

2.2 Semantic Textual Similarity (STS)

We use Chandrasekaran and Mago (2021)'s definition of semantic textual similarity (STS) : "Semantic Textual Similarity (STS) is defined as the measure of semantic equivalence between two blocks of text. Semantic similarity methods usually give a ranking or percentage of similarity between texts, rather than a binary decision as similar or not similar." Determining semantic similarity between two texts has long been based on the number of words in common using methodologies such as Bag of Words (BoW) or Term Frequency - Inverse Document Frequency (TF-IDF). As explained by Chandrasekaran and Mago in their survey on semantic textual similarity, these approaches do not take into account two key aspects. In fact, words can have a different meaning and the same concept could also be explained by using very different words (Chandrasekaran and Mago, 2021). The examples used in the aforementioned survey explain this issue well :

— John and David studied Maths and Science.

— John studied Maths and David studied Science.

In this example, even though the words are exactly the same, the meaning is different ; but if we consider the next example, we realise that different words can convey the same meaning.

— Mary is allergic to dairy products.

— Mary is lactose intolerant.

The following example (already seen in the TS section), from the CLEAR corpus, will further explain the last point.

— Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.

— Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou à ralentir le transit intestinal.

Identifying the two sentences as similar can be difficult. For humans it can be

relatively simple but for a model it is not straightforward task, as the sentences use a very different vocabulary to describe the purpose of the medication and also the state of health. The verb voice and the tense are also different. The passive voice is used in the first sentence, whereas in the second sentence the voice is active, with the use of the imperative tense. It is worth noting that the second sentence is in the negative form.

Finally, the first sentence is phrased impersonally, while the second one addresses the reader directly.

All the issues listed above highlight the various challenges encountered when trying to identify the similarity between two sentences. To try and solve them from 2012 to 2017, the Shared Task Semantic Textual Similarity (STS) was held annually as part of the SemEval [1] workshop series, providing a benchmark for evaluating STS systems. In the 2017 edition, participants worked on pairs of sentences in Arabic, English and Spanish, both within the same language and across languages. For each pair, systems had to assign a continuous similarity from 0 (completely unrelated) to 5 (complete semantic equivalence). Results were then compared with human annotations, and performance was measured by the Pearson correlation between the two sets of scores.

Many methods have been developed over the years to solve these problems ; we will discuss the main ones in this section. We begin with the knowledge-based approaches (Section 2.2.1), followed by the corpus-based approaches (Section 2.2.2) and the Deep Neural Networks (Section 2.2.3) approaches. Finally, those called hybrids(Section 2.2.4), that combine both knowledge-based and corpus-based methods, will be explained.

1. https://ixa2.si.ehu.eus/stswiki/

2.2.1 Knowledge-based methods

The approaches by which semantic similarity is measured have varied greatly over time. The first approach we will consider is the *knowledge-based* one, that "[...] calculate[s] semantic similarity between two terms based on the information derived from one or more underlying knowledge sources like ontologies/lexical databases, thesauri, dictionaries, etc" (Chandrasekaran and Mago, 2021). The interest in ontologies and similar sources stems from the fact that these offer a structured representation of information "in the form of conceptualizations interconnected by means of semantic pointers" (Sánchez et al., 2012).

Before exploring the methodology, we will describe which lexical databases are most commonly used.

WordNet (Miller, 1995b) is one of the most widely used resources for knowledge-based methodologies. "[It] contains more than 118,000 different word forms and more than 90,000 different word senses, or more than 166,000 (f,s) pairs" (Miller, 1995a). WordNet can be seen as a taxonomy of concepts in which nodes denote a set of words that share a common meaning (synonyms) and edges denote hierarchical relationships between synonyms (Zhu and Iglesias, 2017).

BabelNet (Navigli and Ponzetto, 2012b) is described as an *encyclopedic dictionary*; this is done by merging WordNet and Wikipedia[2] (Navigli and Ponzetto, 2012a). BabelNet encodes knowledge as a labeled directed graph composed of a set of nodes and edges connecting pairs of concepts. It is important to note that each node contains a set of lexicalised concepts in different languages from WordNet and Wikipedia as mentioned above (Navigli and Ponzetto, 2012a).

2. https://www.wikipedia.org/

Different types of methods can be identified based on the underlying principles of similarity calculation and the manner in which the ontology is utilized :

— Edge-counting methods

— Information Content-based methods

— Feature-based methods

Edge-counting methods consider the ontology as a graph and aim to find a metric that can measure the similarity between the words by calculating the distance between the edges. They are based on the basic assumption of knowledge-based approaches that if two concepts are close they are also similar (Lofi, 2015). The greater the distance between the words, the lower the similarity (Chandrasekaran and Mago, 2021). Rada et al. (1989) proposed the equation called *path* where the "the similarity is inversely proportional to the shortest path length between two terms"(Chandrasekaran and Mago, 2021) :

$$\mathrm{sim}_{\mathrm{path}}(t_1, t_2) = \frac{1}{1 + \mathrm{min\_len}(t_1, t_2)} \tag{2.1}$$

The limitations of this methodology are mainly the inability to detect "the inhomogeneous granularity of taxonomical relationships (e.g., the semantic distance from “animal” to “living thing” is significantly larger than the distance from “poodle” to “dog”)" (Lofi, 2015).

An attempt was made to solve this problem by Wu and Palmer (1994) and Li et al. (2003) :

The *wup* metric developed by Wu and Palmer (1994) counts the number of edges between each term and the common ancestor shared by both terms in the given ontology.

$$\mathrm{sim}_{wup}(t_1, t_2) = \frac{2\,\mathrm{depth}(t_{lcs})}{\mathrm{depth}(t_1) + \mathrm{depth}(t_2)} \tag{2.2}$$

Unlike methods that rely solely on the shortest path, the approach proposed by Li et al. (2003) also considers the depth of concepts, assigning a higher similarity to concept pairs located deeper in taxonomy (Zhu and Iglesias, 2017).

$$\text{sim}_{li} = e^{-\alpha \min\_\text{len}(t_1,t_2)} \cdot \frac{e^{\beta \,\text{depth}(t_{lcs})} - e^{-\beta \,\text{depth}(t_{lcs})}}{e^{\beta \,\text{depth}(t_{lcs})} + e^{-\beta \,\text{depth}(t_{lcs})}} \tag{2.3}$$

However, since these approaches failed to solve the problem, other methodologies were born. These take into consideration information content (IC) and other features.

Information Content-based methods are based on the idea that a high IC value means that the word is specific, and conversely a low value indicates a general word. Indeed, taking into consideration the context of the words will be a key element in STS. Resnik (1995) introduces the first IC-based measure (Lastra-Díaz et al., 2019) "called *res* which measures the similarity based on the idea that if two concepts share a common subsumer, then they share more information since the IC value of the [Least Common Subsumer] LCS is higher" (Chandrasekar et al., 1996).

$$sim_{res}(t_1, t_2) = IC_{t_{lcs}} \tag{2.4}$$

Since Resnik (1995)'s measure does not take into consideration the "overall information on the path linking both input concepts" (Lastra-Díaz et al., 2019), Lin et al. (1998) and Jiang and Conrath (1997) will propose other measures to overcome this problem : respectively the measures sim_{lin} and dis_{jcn}.

$$sim_{lin}(t_1, t_2) = \frac{2IC_{t_{lcs}}}{IC_{t_1} + IC_{t_2}} \tag{2.5}$$

$$dis_{jcn}(t_1, t_2) = IC_{t_1} + IC_{t_2} - 2IC_{t_{lcs}} \tag{2.6}$$

Feature-based methods "assess similarity between concepts as a function of their properties" (Sánchez et al., 2012). The basic principle for this approach is

derived from Tversky's similarity model(Tversky, 1977), which compares common and uncommon features of terms by subtracting uncommon features from common ones. The result is that common features increase similarity and non-common features decrease it (Jiang et al., 2015). The features on which this approach is based are synonyms, definitions and other semantic relations. These features are extracted from ontologies, such as WordNet in which this type of information can be found (Sánchez et al., 2012).

The main limitation of knowledge-based methods is that they are highly dependent on the resources they rely on. The lack of resources is also exacerbated for all languages other than English. For this, there is indeed Word-Net, but the same cannot be said for all other languages (Chandrasekaran and Mago, 2021).

2.2.2 Corpus-based methods

In corpus-based methodologies, the main objective is to identify the similarity between words using information extracted from large corpora (Chandrasekaran and Mago, 2021). These approaches are based on the *distributional hypothesis*, according to which "similar words appear in similar contexts"(Harris, 1954). Words, however, are not treated as simple units, but are transformed into vectors, which are then used to calculate their similarity (Gorman and Curran, 2006).

In this section, we will examine the concept of embeddings and present some examples of pre-trained embeddings models. Next, we will analyse different methodologies used in the corpus-based approach.

Embeddings. An embedding is a vector representation of a word in a continuous space. A good embedding arranges these vectors so that their geometric relationships reflect the linguistic relationships between the corresponding words (Schnabel et al., 2015). There are several approaches used to calculate embeddings, including neural

networks (Mikolov et al., 2013) or co-occurrence matrices (Pennington et al., 2014).

word2vec (Mikolov et al., 2013) is a neural network model developed and trained using the Google News Corpus. The two main architectures of Word2Vec are the continuous Bag-of-Words model, which uses the surrounding context to predict a word, and the continuous Skip-Gram model, which uses a target word to predict the context. Both architectures comprise an input layer, a hidden layer and an output layer.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) employs a multi-layer bidirectional transformer encoder, following the original design outlined by Vaswani (2017). The model was trained on the BooksCorpus, which contains 800 million words, along with text from English Wikipedia. The implementation of BERT is divided into two main phases : pre-training and fine-tuning. Pre-training is based on two unsupervised tasks. The first, the Masked Language Model (MaskedLM), involves the random masking of a percentage of tokens in the text, with the aim of predicting these missing tokens. The second task is next-sentence prediction, in which the model must determine whether a given sentence logically follows a previous sentence.

Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) has been one of the most widely used techniques for calculating semantic similarity. It uses the co-occurrence matrix technique in which rows are represented by words and columns by paragraphs, and cells represent word counts. This matrix is then subjected to a reduction in size using *Singular Value Decomposition* (SVD) which simplifies the matrix. In fact, the number of columns is reduced while the number of rows remains unchanged, thus preserving the similarity structure between words. Semantic similarity is calculated using the cosine value between the word vectors. LSA makes

it possible to assess the similarity between sentences, paragraphs and documents as it extends the approach and allows longer elements to be substituted for words in the rows (Chandrasekaran and Mago, 2021).

Word Attention Models : Building on the idea that humans naturally focus on key words when reading sentences, the incorporation of an attention mechanism(Bahdanau, 2014) has been shown to enhance the semantic representation of sentences (Wang et al., 2017a). A word attention model focuses on the most meaningful words in a sentence, by assigning different weights to words based on their contextual importance using techniques such as word alignment, word frequency or word association (Chandrasekaran and Mago, 2021).

Attention Constituency Vector Tree. Le et al. (2018) developed an Attention Constituency Vector Tree, where each node in the tree contains three core components that form a comprehensive representation of the word : a word embedding derived from a corpus, an attention score, and the word's modification relations. The similarity between two words is found using a tree kernel function :

$$\text{TreeKernel}(T_1, T_2) = \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \Delta(n_1, n_2),$$

$$\Delta(n_1, n_2) = \begin{cases} 0, & \text{if } n_1 \text{ and/or } n_2 \text{ are non-leaf nodes and } n_1 \neq n_2 \\ Aw \times \text{SIM}(vec_1, vec_2), & \text{if } n_1, n_2 \text{ are leaf nodes,} \\ \mu\left(\lambda^2 + \sum_{p=1}^{l^m} \delta_p(c_{n_1}, c_{n_2})\right), & \text{otherwise.} \end{cases}$$

Corpus-based approaches, as seen in this section, represent linguistic units as vectors, with well-known models such as LSA, Word2Vec, and BERT exemplifying this approach. Their main advantage lies in their adaptability across different languages and domains (Chandrasekaran and Mago, 2021). However, they rely heavily on

the availability of extensive, high-quality training data and require considerable computational power, which can limit their effectiveness in low-resource languages, specialized fields with scarce data, or settings with limited processing capacity.

2.2.3 Deep Neural Networks-based methods

Deep neural networks are becoming increasingly popular and now represent the state-of-the-art in textual similarity methodologies. In this section, we will examine some examples of methodologies based on this approach.

Lexical Decomposition and Composition. Wang et al. (2017b) base their work not only on the lexical similarity between two sentences, but also on their differences, which are also considered informative for assessing semantic relations. Each word is represented through embeddings, and for each word, a semantic matching vector is computed. This vector is then decomposed into two components : one similar and one dissimilar, depending on the degree of semantic alignment. As shown in Figure 2.2, the model processes the “similar” and “dissimilar” components separately and combines them to produce a global representation that allows the model to estimate sentence similarity through a two-channel CNN.

Text-to-Text Transfer Transformer. Raffel et al. (2020), with their text-to-text approach, unify all NLP tasks by formulating them as text generation problems, as shown in Figure 2.3. This significantly simplifies the overall architecture. They also introduce the C4 corpus, the "Colossal Clean Crawled Corpus", consisting of English texts collected from the web. This corpus was used to train the T5 models, "Text-to-Text Transfer Transformer". These models achieve outstanding results, as shown by Chandrasekaran and Mago (2021), and were considered state-of-the-art at the time of publication. A major recurring limitation of this and other transformer-based

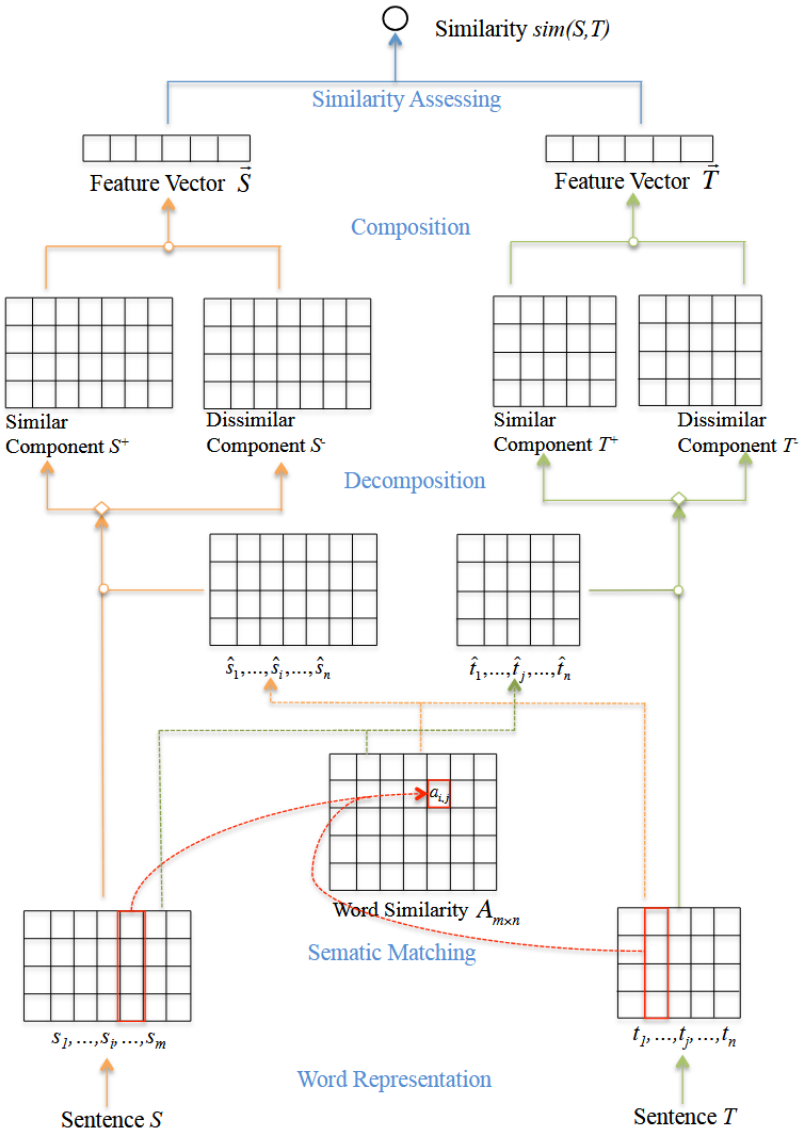

FIGURE 2.2 – Lexical Composition and Decomposition Model Overview

models is their high computational cost, which often renders them inaccessible to many researchers.

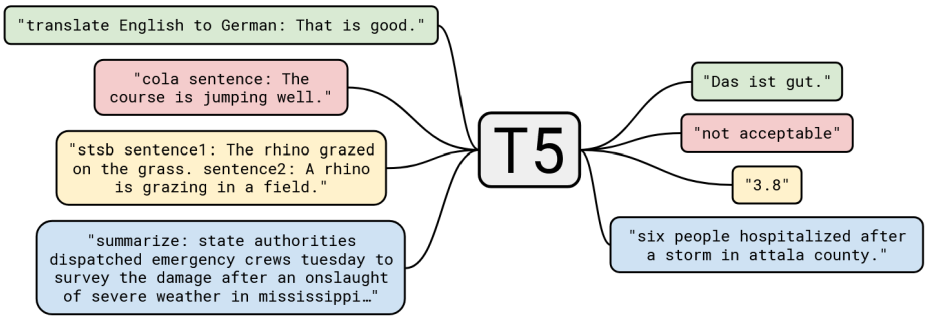

FIGURE 2.3

2.2.4 Hybrid methods

In the previous sections, we explored the main approaches to the calculation of textual semantic similarity (STS), highlighting their respective strengths and weaknesses. In recent years, however, there has been a growing trend towards the use of hybrid STS methodologies, which combine multiple approaches.

NASARI was developed by Camacho-Collados et al. (2015) and is a technique that combines lexicographic and encyclopedic resources to obtain semantic representation of concepts (Camacho-Collados et al., 2015). The resources used are WordNet and Wikipedia and to link the two, BabelNet is used. The semantic representation of the words goes through two stages : the first one consists in collecting the relevant Wikipedia pages and WordNEt information for a given concept ; the second one consists in the analysis of the information obtained and the creation of two vector representation of the concept, one word-based and one synset-based. The first vector is created using Lexixal specificity (Lafon, 1980) that is used to accurately extract the correct representative terms for a given subcorpus (Lafon, 1980), leveraging hypergeometric distribution.

$$\text{spec}(T, t, F, f) = -\log_{10} P(X \geq f) \tag{2.7}$$

As example, Camacho-Collados et al. (2015) explain that the the vector of *shore* (meaning edge of water) "has water, ocean, lake, beach and sea among its most relevant dimensions." The second vector uses as dimensions the WordNet synsets and the weights are calculated "on the basis of [...] the individual words in the group" (Camacho-Collados et al., 2015). In line with the previous example, the terms ocean, lake, and sea are categorized under their common hypernym, the synset for body of water, and are grouped within the same dimension. (Camacho-Collados et al., 2015). Using this approach, the similarity between two concepts uses the following methodology : after the extraction of the two sets of embeddings per concept, the similarity between each corresponding set of embedding is calculated. Finally, the average between the two similarity scores is computed using the Weighted Overlap :

$$WO(v_1, v_2) = \sqrt{\frac{\sum_{d \in O} \left(\text{rank}(d, \vec{v}_1) + \text{rank}(d, \vec{v}_2)\right)^{-1}}{\sum_{i=1}^{|O|} (2i)^{-1}}} \tag{2.8}$$

Most Suitable Sense Annotation (MSSA) is one of the three algorithms proposed by Ruas et al. (2019) to create word-sense embeddings. The other two are Most Suitable Sense Annotation N Refined (MSSA-NR), and Most Suitable Sense Annotation Dijkstra (MSSA-D). However, as these approaches did not provide valuable results, new methodologies were developed, integrating information content (IC) and additional features.

1. disambiguation followed by annotation ;
2. token embeddings training.

For the first one, the corpus is composed of "two Wikipedia Dumps" and each word is transformed in a synset using WordNet as lexical resource. In order to accomplish this, one of the three algorithms aforementioned can be used.
"MSSA works locally, trying to choose the best representation for a word, given its context window. MSSA-D on the other hand, [...] it considers the most similar

word-senses from the first to the last word in a document" (Ruas et al., 2019, p. 5). MSSA-NR however is an iterative model. On the first pass, synset embeddings are produced ; these are then fed back as input to the system to produce more refined synset-embeddings. For the second task, Ruas et al. (2019) use word2vec to obtain the vectors.

Unsupervised Ensemble Semantic Textual Similarity Methods (UESTS) (Hassan et al., 2019) is a methodology though which the similarity of two sentences is calculated combining 4 different measures. In the following formula :

$$\text{sim}_{\text{ESTS}}(S_1, S_2) = \alpha \cdot \text{sim}_{\text{WAL}}(S_1, S_2) + \beta \cdot \text{sim}_{\text{SC}}(S_1, S_2) + \gamma \cdot \text{sim}_{\text{embed}}(S_1, S_2) + \theta \cdot \text{sim}_{\text{ED}}(S_1, S_2) \tag{2.9}$$

we can see how the four scores WAL, SC, embed and ED are added together, and each of them is multiplied by a weight : α, β, γ, θ respectively. The WAL (Weighted Alignment-Based Similarity) approach is calculated with the following formula :

$$\text{sim}_{\text{WAL}}(S_1, S_2) = \frac{2 \cdot \sum_{(i,j) \in A_C} \min(\alpha_i, \alpha_j) \cdot \text{WordSim}(w_i, w_j)}{\sum_{i=1}^{|C_1|} \alpha_i + \sum_{j=1}^{|C_2|} \alpha_j} \tag{2.10}$$

and uses a word synset-based alligner that leverages BabelNet to identify semantically related word pairs across the two sentences (Chandrasekaran and Mago, 2021). The SC approach uses *Soft-Cardinality* to count the elements in a set, by taking their similarity into account. The idea is that highly similar items will contribute less to the total. Following Jimenez et al. (2012) the sentences are represented as sets of q-grams and the similarity is measured using the Dice coefficient (Hassan et al., 2019).

$$\text{sim}_{\text{SC}}(S_1, S_2) = \frac{|S_1 \cap S_2|' + \text{bias}}{\alpha \cdot \min(|S_1'|, |S_2'|) + (1 - \alpha) \cdot \max(|S_1'|, |S_2'|)} \tag{2.11}$$

The embed approach calculates the cosine similarity on the embeddings of the two

sentences, whereas the simED (Edit Distance) calculates the dissimilarity between two sentences (Hassan et al., 2019). To do so, the Levenshtine distance is used.

2.2.5 Conclusion

As we have seen in this section, the literature on semantic text similarity reveals strengths and weaknesses of four methodological families. Knowledge-based methods offer transparency and linguistic rigor but are limited by fixed resources and insufficient coverage, especially in less-resourced languages. Corpus-based methods capture word usage patterns statistically, and better adapt to different languages, however as knowledge-based methods they heavily rely on extensive and high-quality training data. Deep neural network–based approaches, particularly those using transformer architectures, provide state-of-the-art performance, however, these models require extensive computational resources and data, and are difficult to interpret especially in sensitive domains. Hybrid methods attempt to combine the best of both worlds, but many existing approaches are either narrowly tailored or computationally inefficient.

Chapitre 3

Methodology

3.1 Overview of Methodology

This section presents the methodology used to develop a hybrid approach for identifying the similarity between two sentences. The methodology described here was published in the article "Approaching Semantic Text Similarity with Hybrid Methods : a Case Study on French" by Fadda et al. (2024) and is divided into two separate experiments : the first is a regression task, and the second is a classification task, as shown in Figure 3.1.

The first task is based on the DEFT 2020 corpus (Cardon and Grabar, 2020a) and consists of three separate experiments. The first experiment consists in comparing the performance of three models : CamemBERT, FlauBERT and Sentence CamemBERT Large. The second one relies on similarity measures calculated both directly on the text strings and on the embeddings produced by the models mentioned above. The third experiment combines the two approaches by concatenating the embeddings and the similarity measures.

The second task is a classification task. In this case, the models previously trained on the DEFT corpus (Cardon and Grabar, 2020a) are applied to a section of 500

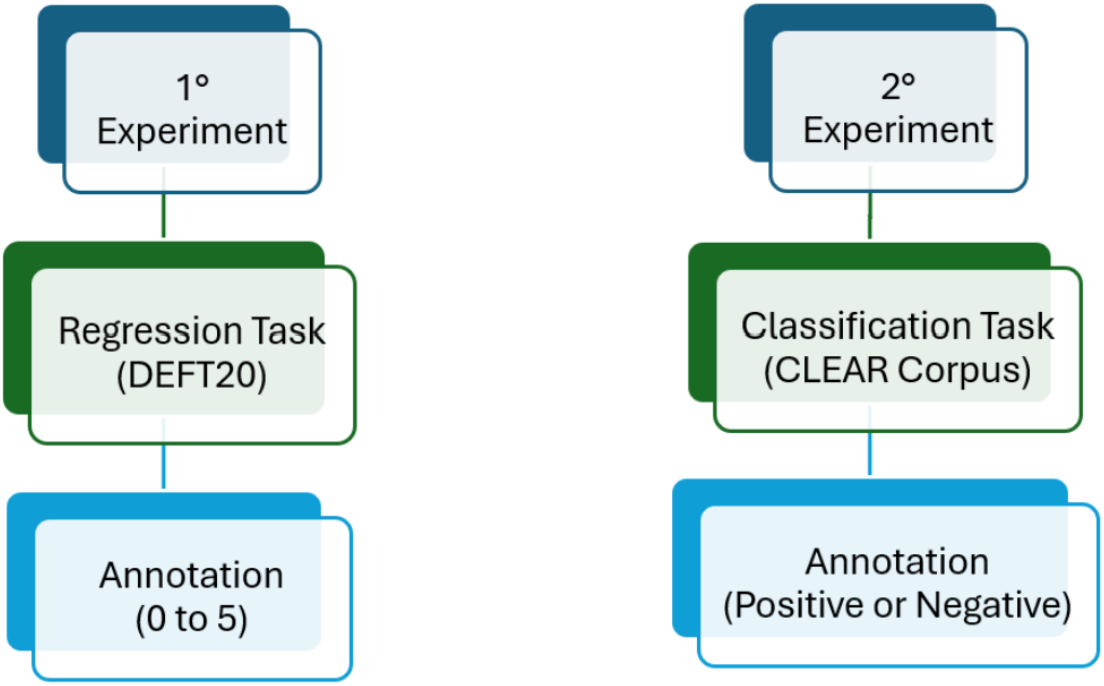

FIGURE 3.1 – Overview of the Methodology

sentences of the CLEAR corpus (Grabar and Cardon, 2018).

In this chapter, the datasets are first presented (Section 3.2). This is followed by a description of the creation of the embedding-based sentence representations (Section 3.3), the similarity measures (Section 3.4), and the hybrid model (Section 3.5). Finally, Sections 3.6 and 3.7 detail the regression task and the classification task, respectively.

3.2 Datasets

3.2.1 DEFT'20 Corpus

For our experiments we use an existing French corpus made for semantic similarity (Cardon and Grabar, 2020a). This corpus was used for an STS shared task within the DEFT 2020 challenge (Cardon et al., 2020). The corpus is composed of sentences extracted from "Wikipedia [1] and Vikidia [2] pages on various subjects [...] as well as health-related content such as drug inserts [...] and Cochrane summaries" (Cardon and Grabar, 2020a, p. 2) extracted from the CLEAR corpus. As a result, the corpus contains sentence pairs that use technical language specific to the medical domain.

It comprises a total of 1,010 sentence pairs, and is split into a training set (600 sentence pairs) and a test set (410 sentence pairs).

Each sentence pair was annotated by five annotators, who assigned them a score on a Likert scale from 0 to 5. Only four instruction points where given to the annotators : A value of 0 indicates completely dissimilar sentences, while a value of 5 represents sentences with identical meaning. For everything in between, they had to draw their own criteria and define them (Cardon and Grabar, 2020a) ; figure 3.2 summarizes them. In addition to the individual score given by each annotator, the average of

1. https://fr.wikipedia.org/
2. https://fr.vikidia.org/

	A1	A2	A3	A4	A5
0.5		A few identical segments			
1	Same topic, loose relation	One summarizes the other	Little shared information	Inference can be drawn	Almost unrelated meaning
1.5		Incomplete main information on one side and extra information missing			
2	Same topic, different information	Incomplete main information on one side	Same function, little shared information	Intermediate level	Same subject, different information
2.5		Same meaning, radically different expression			
3	Same topic, loosely shared information	Same meaning, different expression	Extra information on one side	Main concept of one sentence is missing in the other one	Extra information on one side
3.5		Same meaning, paraphrases are found			
4	Almost same content, additional information on one side	Same meaning, slight rephrasing	Same function and almost same information	Additional information on one side	One slight difference in the delivered information
4.5		Same meaning, slight syntactic difference			

FIGURE 3.2 – Annotation criteria defined by the annotators (Cardon and Grabar, 2020a).

the scores is also present. Cardon and Grabar (2020a) used the Krippendorff's α (Krippendorff, 1970) to calculate the correlation coefficient of the annotators, that produced a α score of 0.69. This score was acceptable, but it did not obtain the best results since the average of the annotators' scores obtained an α score of 0.77. Since the average of the annotators' scores was considered more reliable, we used it as target value.

3.2.2 CLEAR Corpus

The CLEAR corpus is a comparable corpus aligned at document level, it contains both the simplified and the complex versions of documents in the medical domain (Grabar and Cardon, 2018). The corpus includes :

Encyclopedia articles from French Wikipedia and French Vikidia (the Wikipedia for children aged from 8 to 13). These articles discuss various topics including politics, economics, medicine, etc. From these sources, a total of 575 topics are extracted. Their composition is shown in Table3.1

Source	**Word Occurrences**
Wikipedia	2,293,078
Vikidia	197,672

TABLE 3.1 – Word occurrences in the Encyclopedia articles.

Drug Leaflets that contain information such as composition, prescription indications, precautions etc... . Each leaflet is produced in two versions, one for the health provider, and one for the public. The second one uses a simpler jargon and a more personal style (Grabar and Cardon, 2018). Word occurrences in technical and simplified medical leaflets from 11,800 drugs are reported in Table 3.2.

Leaflet Type	Word Occurrences
Technical Leaflets	52,313,126
Simplified Leaflets	33,682,889

TABLE 3.2 – Word occurrences in medical leaflets

Cochrane summaries are summaries of reviews of different medical researches that are created in two forms, one for experts and a simplified version for non-experts. A total of 3,815 reviews contain both versions, and Table 3.3 reports the corresponding word counts.

Summary Type	Word Occurrences
Technical Summaries	2,840,003
Simplified Summaries	1,515,051

TABLE 3.3 – Word occurrences Cochrane summaries

From the comparable data of CLEAR, a subset has been extracted. It is a parallel corpus containing a total of 663 pairs of sentences that have been manually aligned (Grabar and Cardon, 2018).

3.3 Embedding-Based Sentence Representations

In this section, we will explain the Regression experiment in detail. We proceed with three sets of experiments with regression. For each, we use the train and test splits of the DEFT 2020 corpus. One set of experiments consists in comparing the performance of the three language models mentioned before : CamemBERT (Martin et al., 2020), FlauBERT (Le et al., 2020), and Sentence-CamemBERT-Large (Reimers and Gurevych, 2019). To do so, we train a Random Forest regression model on top

of the concatenation of the embeddings of each sentence for a given pair.

Camembert [3] (Martin et al., 2020) is a multilayer bidirectional Transformer model based on the RoBERTa architecture (Liu et al., 2019), which represents an evolution of the original BERT implementation. Compared to BERT (Devlin et al., 2019), RoBERTa introduces improvements such as the use of larger training batches and the removal of the Next Sentence Prediction task. The main difference between CamemBERT and RoBERTa is that the former uses SentencePiece tokenisation (Kudo and Richardson, 2018) and is trained on the French portion of the OSCAR corpus (Suárez et al., 2019). The model was evaluated on four main tasks : POS tagging, dependency parsing, named entity recognition, and natural language inference.

Flaubert [4] is also based on the BERT and RoBERTa approaches, but uses a smaller corpus. The training data comes from 24 distinct sub-corpora which, after being preprocessed and cleaned, were reduced to a total of 71 GB (Le et al., 2020), as we can see in Table 3.3.

	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$	CamemBERT	$\text{FlauBERT}_{\text{BASE}}$/$\text{FlauBERT}_{\text{LARGE}}$
Language	English	English	French	French
Training data	13 GB	160 GB	138 GB†	71 GB‡
Pre-training objectives	NSP and MLM	MLM	MLM	MLM
Total parameters	110 M	125 M	110 M	138 M/ 373 M
Tokenizer	WordPiece 30K	BPE 50K	SentencePiece 32K	BPE 50K
Masking strategy	Static + Sub-word masking	Dynamic + Sub-word masking	Dynamic + Whole-word masking	Dynamic + Sub-word masking

†, ‡: 282 GB, 270 GB before filtering/cleaning.

FIGURE 3.3 – Comparison between the models. (Le et al., 2020)

Sentence-CamemBERT-Large is a model available on Hugging Face [5] that applies the Sentence-Transformers (SBERT) (Nils Reimers, 2019) architecture to

3. https://huggingface.co/docs/transformers/model_doc/camembert
4. https://huggingface.co/docs/transformers/model_doc/flaubert
5. https ://huggingface.co/dangvantuan/sentence-camembert-large

a CamemBERT-large base, allowing it to generate fixed-size semantic embeddings in French. SBERT is a variant of BERT that uses Siamese and triplet networks to generate meaningful fixed-size semantic embeddings. To obtain these embeddings, SBERT adds a pooling operation to the outputs of BERT. SBERT training is mainly based on a combination of the SNLI (Bowman et al., 2015) and Multi-Genre NLI datasets (Williams et al., 2017).

3.3.1 Sentence Embedding Generation

As a first step for both CamemBERT and FlauBERT models we tokenize the input sentence with their respective tokenizers. Special tokens such as [CLS] (classification) and [SEP] (separator) are added. Due to architectural constraints proper to *BERT, the max length has been set to 512 tokens. Padding is applied to ensure uniformity. We generate *BERT embeddings for each tokenized sentence, and the hidden states of the last layer are extracted. We calculate the average of those hidden states for all words in the sentence, considering the attention mask associated with each sentence in order to avoid including padding tokens in the calculation.

Unlike CamemBERT and FlauBERT, Sentence-BERT directly produces a single vector for the entire sentence. The model is specifically designed to directly produce fixed-size sentence embeddings (Nils Reimers, 2019). As a result, once a sentence has passed through Sentence-BERT, the output is already a semantically meaningful vector representation of the entire sentence, eliminating the need for further averaging or padding calculations.

3.4 Similarity measures

The second set of experiments studies the various classic similarity measures. The similarity measures we choose are those that were also used in the DEFT 2020

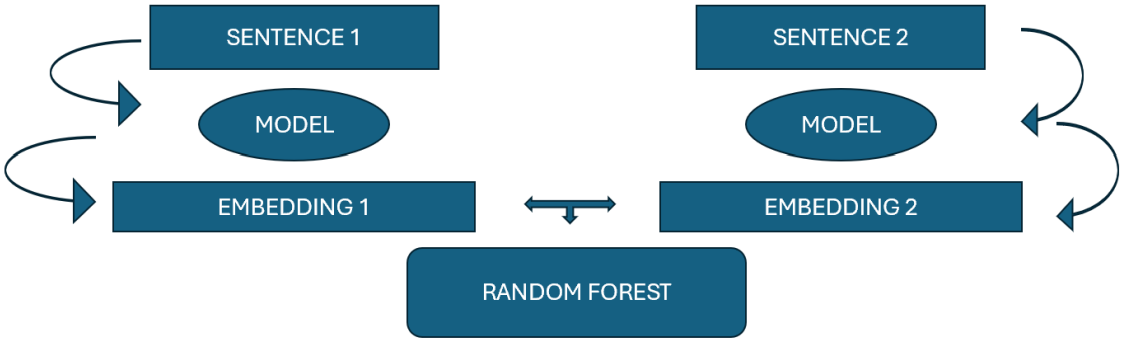

FIGURE 3.4 – First Experiment

campaign (Cardon et al., 2020) and that produced good results. Among others Buscaldi et al. (2020), Cao et al. (2020), Dramé et al. (2020) use similarity measures. On the one hand, we include similarity measures calculated on the embeddings, such as the Manhattan distance, the Jaccard distance (Jaccard, 1912), the Euclidean distance, the cosine similarity and dissimilarity, the Bray-Curtis dissimilarity (Bray and Curtis, 1957), and Ochiai coefficient (Ochiai, 1957). On the other hand, other measures were calculated comparing the two strings of characters, and they include the Sorensen Dice coefficient (Dice, 1945), the Levenshtein distance (Levenshtein, 1966), the Qgram coefficient (Ukkonen, 1992), the number of words in common, the length of each sentence and the absolute difference between the length of the two sentences.

3.4.1 Embeddings similarity measures

Listed below are the formulas used to calculate the similarity measures.

Manhattan Distance :

$$\text{Dist}_{\text{Manh}} = \sum_{i=1}^{n} |x_i - y_i|$$

Euclidean distance :

$$\text{Dist}_{\text{Eucl}} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Cosine similarity :

$$\text{Cosinus} = \frac{\mathbf{V} \cdot \mathbf{W}}{\|\mathbf{V}\| \cdot \|\mathbf{W}\|}$$

Cosine Dissimilarity :

$$\text{Cosine Dissimilarity} = 1 - \frac{\mathbf{V} \cdot \mathbf{W}}{\|\mathbf{V}\| \cdot \|\mathbf{W}\|}$$

Jaccard index (Jaccard, 1912) :

$$\text{Jaccard} = \frac{|\text{set}(\mathbf{V}) \cap \text{set}(\mathbf{W})|}{|\text{set}(\mathbf{V}) \cup \text{set}(\mathbf{W})|}$$

Bray-Curtis dissimilarity (Bray and Curtis, 1957) :

$$\text{Bray-Curtis Dissimilarity} = \frac{\sum_{i=1}^{n} |V[i] - W[i]|}{\sum_{i=1}^{n}(V[i] + W[i])}$$

Ochiai coeffiecient (Ochiai, 1957) :

$$\text{Ochiai} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

3.4.2 Classic similarity measures

For the measures calculated directly from the sentence and not from its latent representation of the embeddings, we performed the pre-processing directly on the sentences. The pre-processing involves the removal of stopwords using the NLTK

package (Bird et al., 2009) and the removal of all non-alphanumeric characters.

Sorensen Dice coefficient (Dice, 1945) :

$$\text{Dice} = \frac{2 \cdot |\text{set}(V) \cap \text{set}(W)|}{|\text{set}(V) \cup \text{set}(W)|}$$

Levenshtein distance (Levenshtein, 1966) :

The formula was retrieved from the Wikipedia page. [6]

$$\text{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if head}(a) = \text{head}(b), \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(a), b), \\ \text{lev}(a, \text{tail}(b)), \\ \text{lev}(\text{tail}(a), \text{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$

Q-gram coefficient (Ukkonen, 1992) :

$$D_q(x,y) = \sum_{v \in \Sigma^q} |G(x)[v] - G(y)[v]| \,.$$

Number of words in common :

$$N_{\text{common}} = |S_1 \cap S_2|$$

Length of each sentence measured in tokens :

$$\text{len}_1 = |S_1| \qquad \text{(Length of sentence 1)}$$

$$\text{len}_2 = |S_2| \qquad \text{(Length of sentence 2)}$$

6. `https://en.wikipedia.org/wiki/Levenshtein_distance`

Absolute difference between two sentences :

$$D_{\text{abs}} = |L(S_1) - L(S_2)|$$

In order to find the best combination of features, in this set of experiments we test all the possible subsets of similarity measures. Since there are 14 measures, a total of 16,383 Random Forest models are trained. Referring to the Spearman's correlation measure calculated for every Random Forest model, we aim to identify the best combination of measures.

3.5 Hybrid Modeling

The third set of experiments consists in using the best set of measures we obtained after the second set of experiments, combined with the top two out of three language models. Indeed, due to a limited availability of computing power, we are not able to re-train more than sixteen thousand Random Forests that include the embeddings and all possible combinations of measures. We therefore chose to limit this last step to the best combination of measures. We concatenate the embeddings with the vector of features produced by the best combination of similarity measures and train a Random Forest model with this approach.

3.6 Regression task

As explained in the Data section (3.2), the corpus used for this task is the one used in the DEFT2020 campaign (Cardon and Grabar, 2020a). In the corpus the label given to each sentence pair is the average of the scores given by the annotators, and it is represented by values ranging from 0 to 5. Since the score lies on a continuous numerical scale, we frame the problem as a regression task.

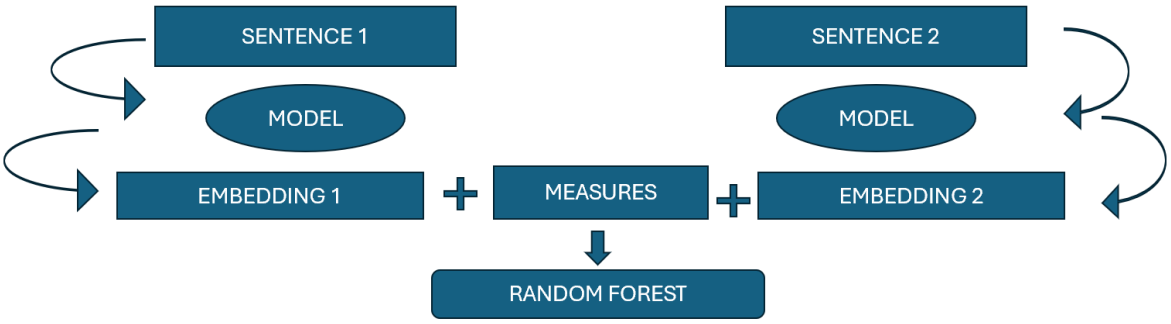

FIGURE 3.5 – Third Experiment

Given that the Random Forest was both the most frequently used and one of the top-performing approaches in the DEFT 2020 campaign, we adopt it as the regression method in the current experiment.

We use the RandomForestRegressor class from the Scikit-learn library (Pedregosa et al., 2011), version 1.3.2. As for the hyperparameters, we set the number of estimators to 500 and a random state of 0 for reproducibility, and kept the other hyper-parameters at their default settings. This decision is supported by a grid search over selected hyperparameters, which confirmed the effectiveness of the chosen configuration. Variations in the 'max depth' and 'max features' parameters had no significant impact on the results. Specifically, the 'max depth' parameters were tested with values of None (default), 50, 100 and 150, while for 'max features' the choices included 1.0 (default), sqrt and log2.

After calculating the embeddings and the similarity measures, we proceed to train the Random Forest regressor using the hyperparameters specified above.

We evaluate our experiments using Spearman's correlation coefficient. The correlation is calculated between the target value, that corresponds to the average of the annotations of the pair of sentences, and the prediction calculated by the different models. Since it is widely adopted in semantic textual similarity tasks, including the DEFT 2020 campaign, it is the most appropriate choice for ensuring direct

comparability with previously published results.

3.7 Classification Task

As a last series of experiments, we apply our best models for each set of experiments, trained on the DEFT20 corpus, to the labeled sentence pairs from the CLEAR corpus. Since our models are trained on a regression task, whereas those pairs of sentences are tailored for a binary classification task (either they are similar or they are not), we test different thresholds. If the score is higher than the threshold, the pair is marked as similar, if the score is equal to or lower than the threshold, the pair is considered as not similar. We run this set of experiments with all threshold values between 0.1 and 4.9 in steps of 0.1. We evaluate those experiments with accuracy.

Chapitre 4

Results

4.1 Regression Task

In this section, we report the results obtained with the Regression models. Our evaluation makes use of Spearman's correlation and is structured in three parts : (1) the results obtained from the Random Forest trained on embeddings only (Section 4.1.1), (2) the results obtained from the similarity measures (Section 4.1.2) and (3) finally the results obtained from the concatenation of the best combination of measures and embeddings (Section 4.1.3).

4.1.1 Random Forest on Embeddings only

Table 4.1 reports the Spearman correlations obtained by training a Random Forest model using only the embeddings produced by the aforementioned language models. Among the tested models, CamemBERT embeddings yield the highest correlation (0.76), slightly outperforming Sentence-CamemBERT-Large (0.75) and Flaubert (0.74).

Given the low scores obtained by Flaubert, we will no longer use this model.

TABLE 4.1 – Results of the RF training on embeddings only

Data	**Spearman**
CamemBERT embeddings	**0.76**
Flaubert embeddings	0.74
Sentence Camembert Large	0.75

4.1.2 Random Forest on Measures only

As far as the results of the combinations of measures are concerned, we decided to present only the combination that obtained the very best results. This choice is due to the fact that we do not observe a significant divergence between the Spearman correlations of the various combinations. However, specific measures can be identified, that demonstrate suboptimal performance relative to others. These include the length of the two sentences and the Ochiai coefficient. In particular, these metrics consistently rank in the bottom 10% of the results, each with a frequency of 0.08. They also consistently remain among the least effective metrics in the bottom 20%, where the length of the two sentences has a frequency of 0.16 and the Ochiai coefficient has a frequency of 0.17.

The frequencies of the most effective measures follow a similar pattern. Specifically, the Manhattan distance, the absolute difference between the length of the two sentences, the Levenshtein distance, and the Qgram similarity all rank in the top 10%. The Manhattan distance is the most present with a frequency of 0.83, followed by the absolute difference with 0.74, the Levenshtein distance with a frequency of 0.69 and the Qgram similarity with 0.69. Within the 20% range, the ranking does not change : the Manhattan distance is still the most frequent with 0.72 followed by the absolute difference with 0.70. The Levenshtein distance has a frequency of 0.66

and the Qgram similarity follows with 0.63.

Our experiments indicate that the best combination is composed of the following measures :

— Manhattan distance,

— Euclidean distance,

— Sorensen-Dice coefficient,

— Common words,

— length of the first sentence,

— length of the second sentence,

— absolute difference between the two sentences,

— Levenshtein distance,

— Qgram similarity.

Spearman's correlation values on the best concatenation on measures obtain two different results when the embeddings are calculated using the Camembert model or the Sentence-CamemBERT-Large model. As shown in Table 4.2, the combination of measures that uses Sentence-CamemBERT-Large slightly outperforms the one that uses CamemBERT of 0.04.

TABLE 4.2 – Spearman correlation results for the best concatenation of similarity measures.

Embedding model	**Spearman correlation**
CamemBERT	0.85
Sentence-CamemBERT-Large	**0.89**

4.1.3 Random Forest on Hybrid task

The results of concatenation of embeddings and the best combination of measures calculated using the Camembert model has a Spearman's correlation of 0.88, which is 0.11 above the best performance obtained by the DEFT 2020 participants. It is also 0.04 above the model trained only on similarity measures, and 0.12 above the model trained on CamemBERT embeddings only.
On the other hand, The model that uses the embeddings calculated with Sentence-Camembert-Large has a Spearman's correlation of 0.91, outperforming all the other results.

TABLE 4.3 – Results of the RF training on the concatenation of Embeddings and measures

Data	**Spearman**
CamemBERT embeddings + Measures	0.88
Sentence-Camembert-Large + Measures	**0.91**

4.2 Classification

We report results on six classification models in total, based on the results described in the previous section. The models are :

1. the model trained only on CamemBERT embeddings ;
2. the model trained only on the Sentence-CamemeBERT-Large ;
3. the best performing model trained on similarity measures only (for both models) ;
4. the model that is trained on the concatenation of CamemBERT embeddings and the best performing combination of similarity measures ;

5. the model that is trained on the concatenation of Sentence-CamemeBERT-Large embeddings and the best performing combination of similarity measures.

We apply those models to the 500 sentence pairs (100 positive, 400 negative) from CLEAR and report the accuracy for each threshold. We report only results for thresholds between 1.5 and 3.5 for readability fro the CamemBERT model, as we do not have a good performing threshold outside of this scope. However, for the Sentence-CamemBERT-Large we expand the range to 3.7, to include all important results. We also report the results for the minimum (0.1) and maximum (4.9) thresholds.

The results based on the CamemBERT model are displayed in Table 4.4, whereas the ones based on the Sentence-CamemBERT-Large model are shown in Table 4.5.

For the CamemBERT model, we can see that the best accuracy is obtained by the model trained on measures only, with a score of 97.2 at threshold 2.2. The CamemBERT+measures model is not far behind as it peaks at 96.0, at thresholds 2.9 and 3.0. The model including only CamemBERT embeddings performs the worst. These results suggest that it cannot be leveraged to produce a threshold that distinguishes the two classes : its highest score (accuracy of 80%) is at the maximum threshold (4.9), which means that everything is predicted as non-similar (0). As 80% of examples are negative examples, this model cannot do better than a system that would assign everything to the negative class, despite its Spearman's correlation suggesting it would achieve good results.

For the Sentence-CamemBERT-Large model, the highest accuracy is achieved by the model trained on both embeddings and similarity measures, with a score of 97.8 at threshold 2.5. The model trained on measures only performs slightly worse than the CamemBERT counterpart, reaching 96.2 at threshold 2.3. As with CamemBERT,

the model trained only on Sentence-CamemBERT-Large embeddings produces the lowest accuracy. Its best score is 83.8 at threshold 3.7, which is considerably lower than the other two variants.

TABLE 4.4 – Accuracy results for the three retained methods on the binary classification task

Threshold	**CamemBERT**	**Measures**	**CamemBERT+Measures**
0.1	20.0	27.4	20.0
1.5	20.4	87.8	53.8
1.6	21.2	90.8	58.6
1.7	21.6	94.6	62.2
1.8	23.0	96.6	68.2
1.9	24.0	96.4	70.0
2.0	25.6	97.0	73.8
2.1	26.8	97.0	78.2
2.2	29.6	**97.2**	82.4
2.3	32.2	96.8	87.0
2.4	34.8	96.6	90.6
2.5	36.4	96.4	93.6
2.6	40.2	95.8	94.8
2.7	44.4	95.6	95.4
2.8	47.8	95.2	95.8
2.9	51.6	94.4	96.0
3.0	55.0	92.4	**96.0**
3.1	62.0	91.4	95.0
3.2	68.2	90.4	93.2
3.3	73.2	89.2	91.4
3.4	75.6	88.4	89.6
3.5	78.8	87.0	88.6
4.9	**80.0**	80.0	80.0

TABLE 4.5 – Accuracy (%) for SBERT, measures_SBERT, and hybrid_SBERT across thresholds

Threshold	**SBERT (%)**	**Measures (%)**	**SBERT+Measures (%)**
0.1	20.0	22.4	20.0
1.5	20.0	96.2	91.2
1.6	20.0	96.2	94.6
1.7	20.2	96.0	96.0
1.8	21.4	96.2	97.2
1.9	22.4	96.4	97.8
2.0	23.8	96.4	98.0
2.1	26.0	96.4	98.2
2.2	27.4	96.2	98.4
2.3	29.8	**96.2**	98.6
2.4	32.6	95.8	97.8
2.5	35.4	96.0	**97.8**
2.6	37.8	94.2	97.0
2.7	41.4	92.6	97.0
2.8	44.8	89.4	96.6
2.9	46.6	86.0	96.0
3.0	50.2	83.6	95.6
3.1	55.4	82.8	94.8
3.2	61.4	82.6	93.4
3.3	67.6	80.8	92.6
3.4	76.0	80.4	91.2
3.5	81.0	80.2	90.6
3.6	83.0	80.0	90.0
3.7	**83.8**	80.0	88.4
4.9	80.0	80.0	80.0

Chapitre 5

Discussion

In this chapter, we will discuss the results obtained in the experiments presented in the methodology chapter, with particular emphasis on the analysis of errors made by the models during classification.

5.1 Regression Task

As for the first experiment, in which regression is used to predict the similarity score between pairs of sentences, the experiments revealed clear trends in model performance. In fact, the hybrid approach combining *Sentence-CamemBERT-Large* embeddings with the best-performing combination of similarity measures achieved the highest Spearman correlation (0.91), outperforming all other configurations tested. Indeed, the Spearman correlation of *Camembert* alone (0.85) and *Sentence-Camembert-Large* (0.89) lower. The combination of *CamemBERT* with the measures also produced good results (0.88). These results highlight the robustness of hybrid architectures and the complementarity between embeddings and similarity measures.

5.2 Classification Task

In the classification task, the same trends are confirmed, with the best results being obtained by hybrid models. In particular, *Hybrid_ SBERT* at threshold 2.5 achieves an accuracy of 97.8. Even models based solely on measurements achieve good results in the classification task. For example, the *Measures_ CBERT* model achieved an accuracy of 97.2% at the optimal threshold. On the other hand, models based solely on embeddings do not perform well and, although they achieved acceptable results in the regression task (as shown in section 4.1.1), their suitability for the classification task is insufficient.

5.3 Error Analysis

In this section, we examine four error batches, two from the models that rely exclusively on similarity measures, and two from the hybrid models. For clarity, each batch is assigned a specific designation. *Measures_ CBERT* refers to the similarity measures model in which the similarity scores are computed on embeddings generated by CamemBERT. *Measures_ SBERT* refers to the similarity measures model where the similarity scores are computed on embeddings generated by Sentence-CamemBERT-Large.

The hybrid models follow the same naming logic : *Hybrid_ CBERT* denotes the hybrid model that combines both similarity measures and classification using CamemBERT embeddings, while *Hybrid_ SBERT* designates the corresponding hybrid model using Sentence-CamemBERT-Large embeddings.

The error analysis is conducted using the optimal thresholds determined for each model :

— **2.2** for the *Measures_ CBERT* model

— **2.9** for the *Hybrid_ CBERT* model

— **2.3** for the *Measures_ SBERT* model

— **2.5** for the *Hybrid_ SBERT* model

These manual observations are straightforward to make due to the low number of errors.

The *Measures_ CBERT* model produced 11 false negatives (equivalent pairs that are considered unrelated), resulting in a recall of 0.89 for the positive class, and 3 false positives (nrelated pairs that are considered equivalen), leading to a precision of 0.97 for the positive class.

The *Measures_ SBERT* model produced 16 false negatives, corresponding to a recall of 0.84, and 3 false positives, with a precision of 0.97. This indicates slightly lower sensitivity compared to *Measures_ CBERT*, although the precision remains equally high.

The *Hybrid_ CBERT* model yielded 13 false negatives, giving a recall of 0.87 for the negative class, and 7 false positives, resulting in a precision of 0.93 for the positive class.

Finally, the *Hybrid_ SBERT* model produced a total of 11 errors, 10 of which were false negatives. This corresponds to a recall of 0.84, while the single false positive leads to a precision of 0.99.

5.3.1 False Negatives

For the false negatives, the first plot 5.1 shows that *Measures_ SBERT* accounts for the highest number of errors, followed by *Hybrid_ CBERT*, *Measures_ CBERT*, and *Hybrid_ SBERT*. The plot in figure 5.1 examines how these errors are dis-

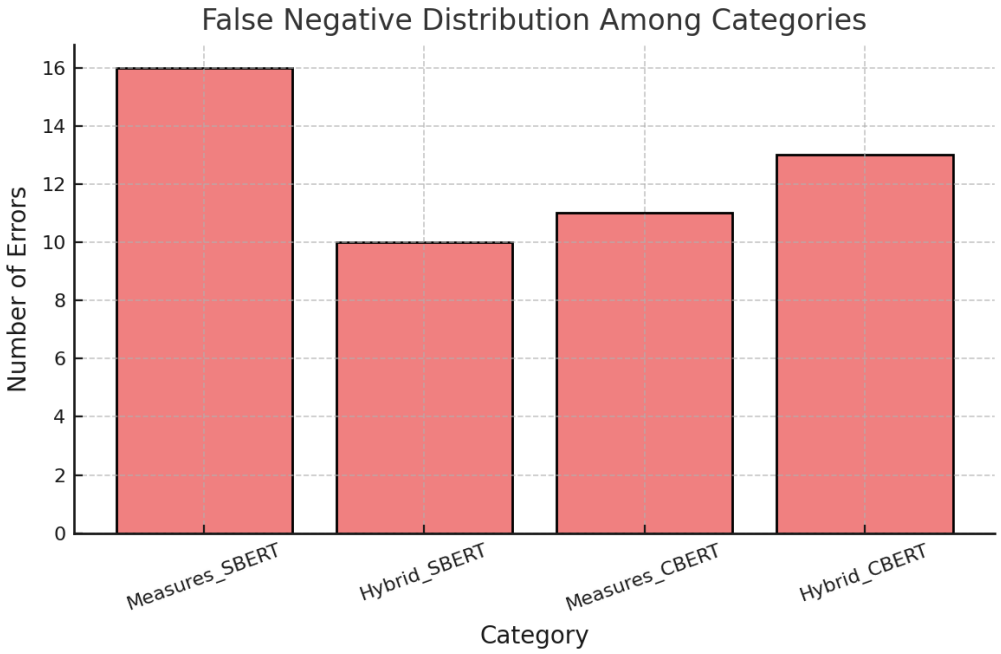

FIGURE 5.1 – Distribution of false negatives across the four evaluated categories.

TABLE 5.1 – False negatives intersections with their corresponding error counts.

Intersection	Errors
Hybrid_SBERT	6
Measures_SBERT	3
Measures_SBERT ∩ Hybrid_CBERT	2
Measures_SBERT ∩ Measures_CBERT	1
Hybrid_SBERT ∩ Hybrid_CBERT	1
Measures_CBERT ∩ Hybrid_CBERT ∩ Measures_SBERT	7
Measures_SBERT ∩ Hybrid_SBERT ∩ Measures_CBERT ∩ Hybrid_CBERT	3

tributed across intersections of the categories. Six false negatives occur only in *Hybrid_SBERT* and three only in *Measures_SBERT*. A further seven are shared between *Measures_CBERT*, *Hybrid_CBERT*, and *Measures_SBERT*, while three are common to all four categories. Smaller overlaps are also present between *Measures_SBERT* and *Hybrid_CBERT*, *Measures_SBERT* and *Measures_CBERT*, and *Hybrid_SBERT* with *Hybrid_CBERT*. These results indicate that while certain categories are prone to unique false negatives, a considerable number arise from multi-category overlaps, particularly among the measure-based and hybrid CBERT models.

There seems to be a common pattern, as the sentences in those pairs differ by the way they organize the delivery of the message, for instance, the voice of the verb (passive vs. active) or an impersonal form vs. a sentence that addresses the reader. The following error is shared between all models :

1. Sentence 1 : *les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation*

   Sentence 2 : *dans ce cas , ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal*

2. Sentence 1 : *la prudence doit être observée en cas d' atteinte du système nerveux central ou périphérique*

   Sentence 2 : *si vous avez des troubles cérébraux (par exemple un cancer qui s' est propagé au niveau du cerveau) , ou des lésions nerveuses (neuropathie)*

Interestingly, the six false negatives that are shared between all models except *Hybrid_SBERT* display the same patterns, indicating that some of those cases are recognized by the models. This tends to suggest that adding information on the syntactic organization of the constituents (possibly with their semantic roles) would be the next step to improve the classification.

5.3.2 False Positives

Regarding the false positives, they are typically long sentences that share words in common even though they do not convey the same message. The example shown below is common to all 4 models (we put in **bold** the phrases present in the two sentences that we suspect to be the origin of the error) :

4. Sentence 1 : *Tous les essais contrôlés randomisés (ECR) ou quasi-ECR de l'adjonction de* ***corticostéroïdes*** *dans le traitement des* ***nouveau-nés atteints de méningite bactérienne****.*
   Sentence 2 : *Est-ce que l'utilisation de* ***corticostéroïdes*** *adjuvants chez les* ***nouveau-nés atteints de méningite bactérienne*** *réduit le risque de décès et la possibilité d'avoir des séquelles neurodéveloppementales ?*

The next two pair of sentences were common to three models : Measures_CBERT, Measures_SBERT and Hybrid_CBERT. The first pair of sentences shows the same pattern as the example shown before, namely the two sentences had words in common but they did not have the same meaning.

4. Sentence 1 : *so14796) de protocole identique , multicentriques , randomisées et contrôlées supportent l' utilisation de capécitabine en première ligne dans le traitement du* ***cancer colorectal*** *métastatique*
   Sentence 2 : *capecitabine eg est utilisé dans le traitement du* ***cancer du côlon*** *, du cancer rectal , du cancer de l' estomac ou du cancer du sein*

4. Sentence 1 : *l' altération de la vigilance peut rendre dangereuses la conduite de véhicules et l' utilisation de machines*
   Sentence 2 : - *troubles de la coordination et de l' équilibre*

Interestingly the second pair of sentences doesn't have any words in common, but both describe adverse effects, and that could be the reason they are shown as similar.

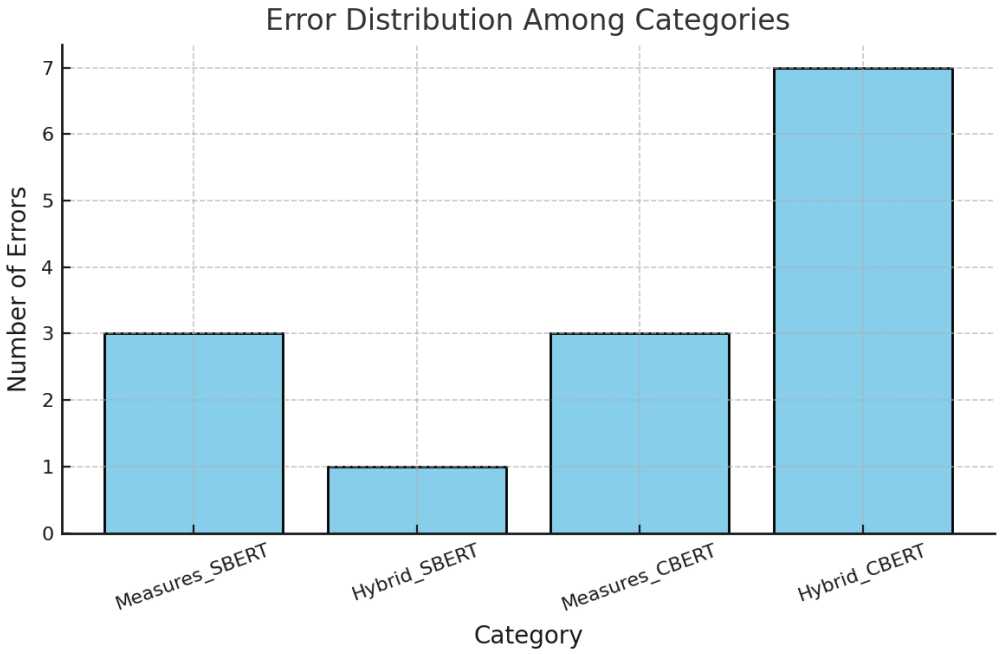

FIGURE 5.2 – False Positives error distribution among categories.

TABLE 5.2 – False positives intersections with their corresponding error counts.

Intersection	Errors
SBERT_MES ∩ SBERT_HYB ∩ CBERT_MES ∩ CBERT_HYBRID	1
SBERT_MES ∩ CBERT_MES ∩ CBERT_HYBRID	2
CBERT_HYBRID	4

Figure 5.2 shows the distribution of False Positives across the four evaluated categories, showing that *Hybrid_ CBERT* produced the largest number of errors, followed by *Measures_ SBERT* and *Measures_ CBERT*, while *Hybrid_ SBERT* contributed the least. Table 5.2 illustrates how these errors overlap between categories. One error is shared by all four categories, two errors are shared by *Measures_ SBERT*, *Measures_ CBERT*, and *Hybrid_ CBERT*, and four errors are unique to *Hybrid_ CBERT*.

5.4 Limitations and Future Work

Although the results are good, two main limitations should be highlighted. First, the corpus on which the models were trained must be taken into account. In fact, it should always be borne in mind that a larger corpus could lead to better results and expanding the training dataset to include more varied paraphrastic constructions could, indeed, improve the robustness of the model.

Secondly, computational constraints limited the exploration of model architectures, the tuning of hyperparameters, and the identification of the best combinations of measures to concatenate with the model embeddings. Although *CamemBERT* and *Sentence-CamemBERT-Large* provided solid benchmarks, it is possible that newer models could further improve the results.

Error analysis shows that false positives often result from high lexical overlap without semantic equivalence, while false negatives often involve paraphrases with significant syntactic variations. To improve these results, adding information about syntactic organization and semantic roles could help models better capture equivalence in structurally different sentences, reducing false negatives. Similarly, it would also allow for the recognition of differences in seemingly identical sentences in false positives.

Chapitre 6

Conclusion

This thesis aimed to address the scarcity of resources for textual similarity (TS) in the French language, particularly in the medical field. Based on the CLEAR corpus, this work proposed and evaluated a hybrid approach that combines sentence embedding from French large language models (CamemBERT, Sentence-CamemBERT-Large) with selected similarity measures. The methodology was tested on both the DEFT 2020 and CLEAR datasets, and the results showed that hybrid models consistently outperform approaches based solely on embeddings or similarity measures. The combination of Sentence-CamemBERT-Large embeddings with an optimized set of measures achieved the highest performance, reaching a Spearman correlation of 0.91 in the regression task, surpassing the best DEFT 2020 results. In the classification task on CLEAR, the same hybrid model achieved 97.8% accuracy by applying an optimal threshold. This model has been applied to the CLEAR (comparable) corpus, resulting in a total of 163,645 pairs of sentences. Considering that the model's accuracy is 97.8%, we can be considered satisfied with the results obtained.

Bibliographie

F. Alva-Manchego, C. Scarton, and L. Specia. Data-driven sentence simplification : Survey and benchmark. *Computational Linguistics*, 46(1) :135–187, 2020. doi : 10.1162/coli_a_00370. URL `https://aclanthology.org/2020.cl-1.4`.

M. Artetxe and H. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7 :597–610, 2019.

D. Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*, 2014.

C. Basu, R. Vasu, M. Yasunaga, and Q. Yang. Med-easi : Finely annotated dataset and models for controllable simplification of medical texts. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, 2023.

N. D. Berkman, S. L. Sheridan, K. E. Donahue, D. J. Halpern, and K. Crotty. Low health literacy and health outcomes : an updated systematic review. *Annals of internal medicine*, 155(2) :97–107, 2011.

S. Bird, E. Loper, and E. Klein. *Natural Language Processing with Python.* O'Reilly Media, Inc., 2009.

S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv :1508.05326*, 2015.

J. R. Bray and J. T. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4) :325–349, 1957.

D. Buscaldi, G. Felhi, D. Ghoul, J. Le Roux, G. Lejeune, and X. Zhang. Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? (sentence similarity : a study on similarity metrics with words and character strings). In R. Cardon, N. Grabar, C. Grouin, and T. Hamon, editors, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 14–25, Nancy, France, 6 2020. ATALA et AFCP. URL `https://aclanthology.org/2020.jeptalnrecital-deft.2/`.

J. Camacho-Collados, M. T. Pilehvar, and R. Navigli. NASARI : a novel approach to a semantically-aware representation of items. In R. Mihalcea, J. Chai, and A. Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 567–577, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi : 10.3115/v1/N15-1059. URL `https://aclanthology.org/N15-1059`.

D. Cao, A. Benamar, M. Boumghar, M. Bothua, L. Ould Ouali, and P. Suignard. Participation d'EDF R&D à DEFT 2020 (this paper describes the participation of EDF R&D at DEFT 2020 evaluation campaign). In R. Cardon, N. Grabar, C. Grouin, and T. Hamon, editors, *Actes de la 6e conférence conjointe Journées*

d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes, pages 26–35, Nancy, France, 6 2020. ATALA et AFCP. URL https://aclanthology.org/2020.jeptalnrecital-deft.3/.

R. Cardon and N. Grabar. A French corpus for semantic similarity. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6889–6894, Marseille, France, May 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.851.

R. Cardon and N. Grabar. Reducing the search space for parallel sentences in comparable corpora. In R. Rapp, P. Zweigenbaum, and S. Sharoff, editors, *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 44–48, Marseille, France, May 2020b. European Language Resources Association. ISBN 979-10-95546-42-9. URL https://aclanthology.org/2020.bucc-1.7/.

R. Cardon, N. Grabar, C. Grouin, and T. Hamon. Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases). In R. Cardon, N. Grabar, C. Grouin, and T. Hamon, editors, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 1–13,

Nancy, France, 6 2020. ATALA et AFCP. URL https://aclanthology.org/2020.jeptalnrecital-deft.1.

J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 workshop on integrating artificial intelligence and assistive technology*, pages 7–10. Madison, WI, 1998.

R. Chandrasekar, C. Doran, and S. Bangalore. Motivations and methods for text simplification. In *COLING 1996 Volume 2 : The 16th International Conference on Computational Linguistics*, 1996.

D. Chandrasekaran and V. Mago. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2) :1–37, 2021.

T. C. Davis and M. S. Wolf. Health literacy : implications for family medicine. *FAMILY MEDICINE-KANSAS CITY-*, 36(8) :595–598, 2004.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi : 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297–302, 1945.

K. Dramé, G. Sambe, I. Diop, and L. Faty. Approche supervisée de calcul de similarité sémantique entre paires de phrases (supervised approach to compute semantic

similarity between sentence pairs). In R. Cardon, N. Grabar, C. Grouin, and T. Hamon, editors, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 49–54, Nancy, France, 6 2020. ATALA et AFCP. URL https://aclanthology.org/2020.jeptalnrecital-deft.5/.

A. Fadda, R. Cardon, N. Grabar, and T. François. Approaching semantic text similarity with hybrid methods : a case study on french. In *JADT*, 2024.

C. Garbacea, M. Guo, S. Carton, and Q. Mei. Explainable prediction of text complexity : The missing preliminaries for text simplification, 2021. URL https://arxiv.org/abs/2007.15823.

B. R. Gena. Using corpora in language learning classroom : Corpus linguistics for teachers. *Michigan ELT*, 2010.

J. Gorman and J. R. Curran. Scaling distributional similarity to large corpora. In N. Calzolari, C. Cardie, and P. Isabelle, editors, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 361–368, Sydney, Australia, July 2006. Association for Computational Linguistics. doi : 10.3115/1220175.1220221. URL https://aclanthology.org/P06-1046.

N. Grabar and R. Cardon. CLEAR – simple corpus for medical French. In A. Jönsson, E. Rennes, H. Saggion, S. Stajner, and V. Yaneva, editors, *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands, Nov. 2018. Association for Computational Linguistics. doi : 10.18653/v1/W18-7002. URL https://aclanthology.org/W18-7002.

N. Grabar and H. Saggion. Evaluation of automatic text simplification : Where are we now, where should we go from here. In *Traitement Automatique des Langues Naturelles*, pages 453–463. ATALA, 2022.

Z. S. Harris. Distributional structure. *Word*, 10(2-3) :146–162, 1954.

B. Hassan, S. E. Abdelrahman, R. Bahgat, and I. Farag. Uests : An unsupervised ensemble semantic textual similarity method. *IEEE Access*, 7 :85462–85482, 2019. doi : 10.1109/ACCESS.2019.2925006.

P. Jaccard. The distribution of the flora in the alpine zone. 1. *New Phytologist*, 11 (2) :37–50, 1912.

J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

Y. Jiang, X. Zhang, Y. Tang, and R. Nie. Feature-based approaches to semantic similarity assessment of concepts using wikipedia. *Information Processing Management*, 51(3) :215–234, 2015. ISSN 0306-4573. doi : https://doi.org/10.1016/j.ipm.2015.01.001. URL `https://www.sciencedirect.com/science/article/pii/S0306457315000023`.

S. Jimenez, C. Becerra, and A. Gelbukh. Soft cardinality : A parameterized similarity function for text comparison. In E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, and D. Yuret, editors, **SEM 2012 : The First Joint Conference on Lexical and Computational Semantics – Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 449–453, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL `https://aclanthology.org/S12-1061/`.

T. Kew, A. Chi, L. Vásquez-Rodríguez, S. Agrawal, D. Aumiller, F. Alva-Manchego, and M. Shardlow. Bless : Benchmarking large language models on sentence simplification, 2023. URL https://arxiv.org/abs/2310.15773.

J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training*, pages 1–49, 1975.

K. Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1) :61–70, 1970.

T. Kudo and J. Richardson. Sentencepiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv :1808.06226*, 2018.

P. Lafon. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1) :127–165, 1980.

T. K. Landauer and S. T. Dumais. A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104 :211–240, 1997. URL https://api.semanticscholar.org/CorpusID:1144461.

J. J. Lastra-Díaz, J. Goikoetxea, M. A. Hadj Taieb, A. García-Serrano, M. Ben Aouicha, and E. Agirre. A reproducible survey on word embeddings and ontology-based methods for word similarity : Linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence*, 85 :645–665, 2019. ISSN 0952-1976. doi : https://doi.org/10.1016/j.engappai.2019.07.010. URL https://www.sciencedirect.com/science/article/pii/S0952197619301745.

H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab. FlauBERT : Unsupervised language model pre-training for French. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.302/`.

Y. Le, Z.-J. Wang, Z. Quan, J. He, and B. Yao. Acv-tree : A new method for sentence similarity modeling. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4137–4143. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi : 10.24963/ijcai.2018/575. URL `https://doi.org/10.24963/ijcai.2018/575`.

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10 :707, 1966.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv :1910.13461*, 2019.

Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. on Knowl. and Data Eng.*, 15(4) :871–882, jul 2003. ISSN 1041-4347. doi : 10.1109/TKDE.2003.1209005. URL `https://doi.org/10.1109/TKDE.2003.1209005`.

D. Lin et al. An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304, 1998.

C. Lindholm and U. Vanhatalo. *Handbook of easy languages in Europe*. Frank & Timme, 2021.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*, 2019.

Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv :2001.08210*, 2020.

C. Lofi. Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. *Information and Media Technologies*, pages 493–501, 09 2015. doi : 10.11185/imt.10.493.

L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, de la Clergerie, D. Seddah, and B. Sagot. Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. doi : 10.18653/v1/2020.acl-main.645. URL `http://dx.doi.org/10.18653/v1/2020.acl-main.645`.

L. Martin, A. Fan, É. de la Clergerie, A. Bordes, and B. Sagot. MUSS : Multilingual unsupervised sentence simplification by mining paraphrases. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France, June 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.176/`.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

G. A. Miller. Wordnet : a lexical database for english. *Commun. ACM*, 38(11) : 39–41, nov 1995a. ISSN 0001-0782. doi : 10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

G. A. Miller. Wordnet : a lexical database for english. *Communications of the ACM*, 38(11) :39–41, 1995b.

R. Navigli and S. P. Ponzetto. Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193 :217–250, 2012a. ISSN 0004-3702. doi : https://doi.org/10.1016/j.artint.2012.07.001. URL https://www.sciencedirect.com/science/article/pii/S0004370212000793.

R. Navigli and S. P. Ponzetto. Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193 :217–250, 2012b.

I. G. Nils Reimers. Sentence-bert : Sentence embeddings using siamese bert-networks. *https ://arxiv.org/abs/1908.10084*, 2019.

K. North, T. Ranasinghe, M. Shardlow, and M. Zampieri. Deep learning approaches to lexical simplification : A survey. *arXiv preprint arXiv :2305.12000*, 2023a.

K. North, M. Zampieri, and M. Shardlow. Lexical complexity prediction : An overview. *ACM Computing Surveys*, 55(9) :1–42, 2023b.

A. Ochiai. Zoogeographical studies on the soleoid fishes found in japan and its neighhouring regions-ii. *Nippon Suisan Gakkaishi*, 22(9) :526–530, 1957.

B. Ondov, K. Attal, and D. Demner-Fushman. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11) :1976–1988, 2022.

S. O'Brien. Controlling controlled english. In *EAMT Workshop : Improving MT through other language technology tools : resources and tools for building MT*, 2003.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn : Machine learning in python. *the Journal of machine Learning research*, 12 :2825–2830, 2011.

J. Pennington, R. Socher, and C. Manning. GloVe : Global vectors for word representation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi : 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19 :17 – 30, 02 1989. doi : 10.1109/21.24528.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140) :1–67, 2020.

N. Reimers and I. Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks, 2019. URL `https://arxiv.org/abs/1908.10084`.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *CoRR*, abs/cmp-lg/9511007, 1995. URL `http://arxiv.org/abs/cmp-lg/9511007`.

T. Ruas, W. Grosky, and A. Aizawa. Multi-sense embeddings through a word sense disambiguation process. *Expert Systems with Applications*, 136 :288–303, 2019.

H. Saggion and G. Hirst. *Automatic text simplification*, volume 32. Springer, 2017.

T. Schnabel, I. Labutov, D. Mimno, and T. Joachims. Evaluation methods for unsupervised word embeddings. In L. Màrquez, C. Callison-Burch, and J. Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. doi : 10.18653/v1/D15-1036. URL `https://aclanthology.org/D15-1036`.

M. Shardlow. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1) :58–70, 2014.

M. Shardlow. *Lexical simplification : optimising the pipeline*. The University of Manchester (United Kingdom), 2015.

Siddharthan, Advaith. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2) :259–298, 2014.

S. Štajner, D. Ferrés, M. Shardlow, K. North, M. Zampieri, and H. Saggion. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in Artificial Intelligence*, 5 :991242, 2022.

P. J. O. Suárez, B. Sagot, and L. Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, 2019.

D. Sánchez, M. Batet, D. Isern, and A. Valls. Ontology-based semantic similarity : A new feature-based approach. *Expert Systems with Applications*, 39

(9) :7718–7728, 2012. ISSN 0957-4174. doi : https://doi.org/10.1016/j.eswa.2012.01.082. URL https://www.sciencedirect.com/science/article/pii/S0957417412000954.

A. Tversky. Features of similarity. *Psychological review*, 84(4) :327, 1977.

E. Ukkonen. Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, 92(1) :191–211, 1992.

A. Vaswani. Attention is all you need. *arXiv preprint arXiv :1706.03762*, 2017.

S. Wang, J. Zhang, and C. Zong. Learning sentence representation with guidance of human attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4137–4143, 2017a. doi : 10.24963/ijcai.2017/578. URL https://doi.org/10.24963/ijcai.2017/578.

Z. Wang, H. Mi, and A. Ittycheriah. Sentence similarity learning by lexical decomposition and composition, 2017b. URL https://arxiv.org/abs/1602.07019.

W. M. Watanabe, A. Candido Jr, M. A. Amâncio, M. De Oliveira, T. A. Pardo, R. P. Fortes, and S. M. Aluísio. Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. In *Proceedings of the 2010 international cross disciplinary conference on web accessibility (W4A)*, pages 1–9, 2010.

G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzman, A. Joulin, and E. Grave. Ccnet : Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv :1911.00359*, 2019.

A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv :1704.05426*, 2017.

Z. Wu and M. Palmer. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA, June 1994. Association for Computational Linguistics. doi : 10.3115/981732.981751. URL `https://aclanthology.org/P94-1019`.

W. Xu, C. Callison-Burch, and C. Napoles. Problems in current text simplification research : New data can help. *Transactions of the Association for Computational Linguistics*, 3 :283–297, 2015.

W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4 :401–415, 2016.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore : Evaluating text generation with bert. *8th International Conference on Learning Representations (ICLR 2020)*, Apr. 2020.

G. Zhu and C. A. Iglesias. Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1) :72–85, 2017. doi : 10.1109/TKDE.2016.2610428.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté de philosophie, arts et lettres
Place Blaise Pascal, 1 bte L3.03.11, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/fial