A Format-sensitive BERT-based Approach to Resume Segmentation

Albeiro Espinal
IMT Atlantique
Lab-STICC,
CNRS UMR 6285
Brest, France

Yannis Haralambous
IMT Atlantique
Lab-STICC,
CNRS UMR 6285
Brest, France

Dominique Bedart
DSI Global Services
Le Plessis Robinson, France

John Puentes
IMT Atlantique
Lab-STICC,
CNRS UMR 6285
Brest, France

***Abstract*—In the early stages of a recruitment process, recruiters can spend a lot of time analyzing resumes (CVs) manually. This has led to the development of machine learning methods for the automated analysis of such documents, which currently besides text encompass rich formatting. Since rich formatting is not considered in any of the automated analysis stages and its possible impact has not been studied, this article investigates how to extract, transform, and apply grapholinguistic content. To this end, we propose a format sensitive and BERT-based framework for the essential first step in CV analysis, i.e. segmentation, relating the automatic description of graphic and textual markers, transformed in linguistic variables by means of fuzzification, to identify dependencies and semantic relationships with the recruiters' manual segmentation. Using a training dataset of 150 resumes, our approach achieved an F1-Score of 89% when segmenting 153 new samples.**

I. Introduction

When Saussure told his students, in 1916, that "Language and writing are two distinct systems of signs; the second exists for the sole purpose of representing the first" [1, p. 23], he couldn't imagine that the written modality of language would become increasingly important at the second half of the 20th century and the beginning of the 21st. Even today, mainstream linguists are firm believers in the supremacy of oral vs. written language. In contrast, studies of *grapholinguistics* (i.e., linguistics specific to the written modality) [2] are nonetheless currently considered "exotic" and remain peripheral. On the other hand, computers store and process text using encodings that are descendants of the typewriter [3]. Unicode may provide codepoints for all languages of the world, but because of its inheritance from ASCII, EBCDIC (Extended Binary Coded Decimal Interchange Code), and the typewriter, one of its basic principles is that only "raw" text is encoded (no italics, bold, underlined, or otherwise formatted text [4]). Encoding raw text is like storing voice without prosody or other suprasegmental features: no phonologist will ever accept the principle of speech deprived of its prosody. The formatted text, being simultaneously ostracized by linguists and computer systems, is now only to be found in markup languages (HTML, XSL-FO, TEI) and the proprietary formats of word processor files.

Natural Language Processing (NLP) is located at the junction between Linguistics and Computing; as both avoid formatted text, it should be no surprise that NLP does the same. We decided to go against this practice. Our corpus, professional resumes (CVs), is heavily relying on text formatting: CVs are short, punchy documents of extreme importance for their authors, intended to impress their recipient (the recruiter). CVs have three essential features: (a) they have to show the unique individual values of the candidates, (b) they have to remain within the socially acceptable boundaries of seriousness and professionalism, and (c) they are strongly mimetic since the Web abounds of CV templates and online applications to produce them. This gives candidates a large spectrum of possibilities to format their CV, and the recruiter an insight into some of the candidate's personality and skills. It also provides us with the ideal playground to apply format-sensitive NLP methods and test their efficiency. In our particular case the research question focuses on how to extract, transform, and apply grapholinguistic content, for CV sections segmentation.

To investigate the effect that text formatting may have on automated CV analysis, in this study we propose a format-sensitive algorithm, to simplify and optimize automatic segmentation of modern resumes based on a small number of documents. As a first step, we will examine the segmentation problem. Analyzing organizational specificities from recruiters' perspectives, we construct a framework for format-sensitive resume representation. Following that, we infer knowledge from experts [5], which allows us to identify the most relevant CV's text-linguistic functions associated with an optimal segmentation process. As a final step, BERT-sequence classifiers are adjusted based on resume terminology and format-sensitive document markers to identify such functions, thereby improving the segmentation F1 score.

II. State of the Art

In general, four approaches have been designed to segment CVs automatically: rule-based heuristics, semantic annotation, word embeddings, and transformer models. Heuristic-based methods essentially identify the boundaries of CV sections. For instance, applying rules [6], or combining the rules with a Naive Bayes model to classify each text block into 5 types of pre-defined sections [7]. Ontologies have also been defined to identify CV sections [8] and to represent these sections as ontological concepts [9].

Recently, word embeddings and transformer models have been applied as alternative segmentation methods. Gradient boosting on decision trees was used to classify each line of the resume [10], while CNN (Convolution Neural Network) and Bi-LSTM (Bi-directional Long Short Term Memory) models, as well as conditional random fields, were combined to do the segmentation [11]. Also, automatic segmentation making use of BPNN (Back-Propagation Neural Network), CNN, Bi-LSTM, and CRNN (Convolutional Recurrent Neural Network) models was proposed [12]. Note that due to textual uncertainties [13] related to the term variants that candidates introduce when writing their resumes, voluminous CV corpora are often required [12], making deep learning CV segmentation techniques challenging and inappropriate for small-sized organizations with reduced datasets.

Simultaneously with the development of neural network models, other studies have used the BERT (Bidirectional Encoder Representations from Transformers) model to optimize tasks related to resume structuring, such as entity extraction [14]. In general, the use of the BERT model for performing various CV analysis tasks has increased in recent years since its language representation has proven particularly useful [15]. Yet, the BERT model is limited in some contexts due to the need for complementary information, like document structural and graphical attributes, in addition to the document's text. These limitations have led to the emergence of various approaches to integrate non-textual features into BERT. Some methods added layers of neural networks to represent document structure [16], whereas others incorporated additional features directly into the document's text [17]. Nevertheless, CV segmentation methods have not yet incorporated such representations as part of the structuring process.

Even if understanding of the BERT model to improve CV segmentation is incipient, we believe that the presumed human-like ability of BERT to represent surface features, syntactic dependencies, and semantic relationships of a text [18], provides a suitable basis for enriching CV texts with additional information, i.e., a description of its graphical format.

III. Proposed Approach

Our CV segmentation framework is divided into three axes. A first axis aims to construct a layer representation of the organizational context where CVs are processed. Using small datasets of this type, the second axis adapts BERT-based models to improve CV segmentation. A third axis aims to evaluate the pertinence of fine-tuned BERT models on new CV samples.

A. Introduction to Basic Concepts in Grapholinguistics

In this section, we will introduce some fundamental terms and concepts related to the study of written language, making it easier for non-expert readers to grasp the content.

The most basic elements of written text are called *graphs* [2, p. 63]. The scientific field that investigates these elements is known as *graphetics*. When studying writing systems, we can identify *graphemes* as the smallest meaningful units, similar to how sounds are represented by phonemes in spoken language [2, p. 119]. The study of graphemes is referred to as *graphemics*.

A *1-dimensional graphemic sequence* (1-dim GS) is a series of graphemes arranged in a single line. This can be horizontal, as in English, or vertical, as commonly seen in East Asian languages. However, due to space constraints on a document's page, these sequences must be divided into smaller segments. This results in *2-dimensional graphemic sequences* (2-dim GS), which are organized into multiple lines on a page.

B. Resume Ontology

To develop more robust machine learning solutions, it is essential to obtain a representation of their societal/organizational context [19], [20]. Therefore, the first axis of our approach begins with the application of the UNC-method [21]. UNC stands for Universidad Nacional de Colombia, the place where the method was developed. We perform preliminary interviews with recruiters in order to identify the most convenient resume representation.

For this purpose, we construct various artifacts as a pre-conceptual scheme [21]. Moreover, recruiters' objectives concerning the CV life-cycle in recruitment processes are then identified and organized hierarchically. Process diagrams, as delineated by [21], illustrate enterprise process models related to CVs. A fishbone chart is also employed to derive the associations between organizational issues concerning CVs and their causative factors. Lastly, in accordance with [21], a Process Explanatory Table integrates all prior diagrams to consolidate the depiction of the organizational context.

The aforementioned process allows a direct extraction of a CV ontology tailored to the specificities of a given organization. This ontology is a fundamental tool for the rich and thorough depiction of each of the concepts that comprise the CV's text. Additionally, it provides the capability of performing further semantic annotations, which are usually applied after a segmentation process [15]. In Fig. 1, we show an upper view of the extracted ontology by using this approach.

C. Resume Terminology Extraction

We perform a format-sensitive extraction of the CV's text. Using a lazy parsing strategy, Unicode characters and their graphetic properties (coordinates and font styles) are identified in the PDF data of the CV. Then, we apply a layout algorithm to extract 1-dim and 2-dim GSs. After identifying them, we extract the CV's terminology using the approach proposed by [22]. Terms in the CV are identified by estimating their termhood [23] using weirdness ratio [22]. Furthermore, we identify syntagmatic compounds, i.e., morphological, graphemic, and semantic variants of terms [22].

D. Recruiter Annotation Analysis

Following the extraction of resume-specific terminology, the second axis of our approach commences with the analysis of the way recruiters manually segment CVs. A panel of recruiters manually segments a set of resumes $D = d_1, \ldots, d_M$. For each resume m, we obtain a tuple of annotated sections

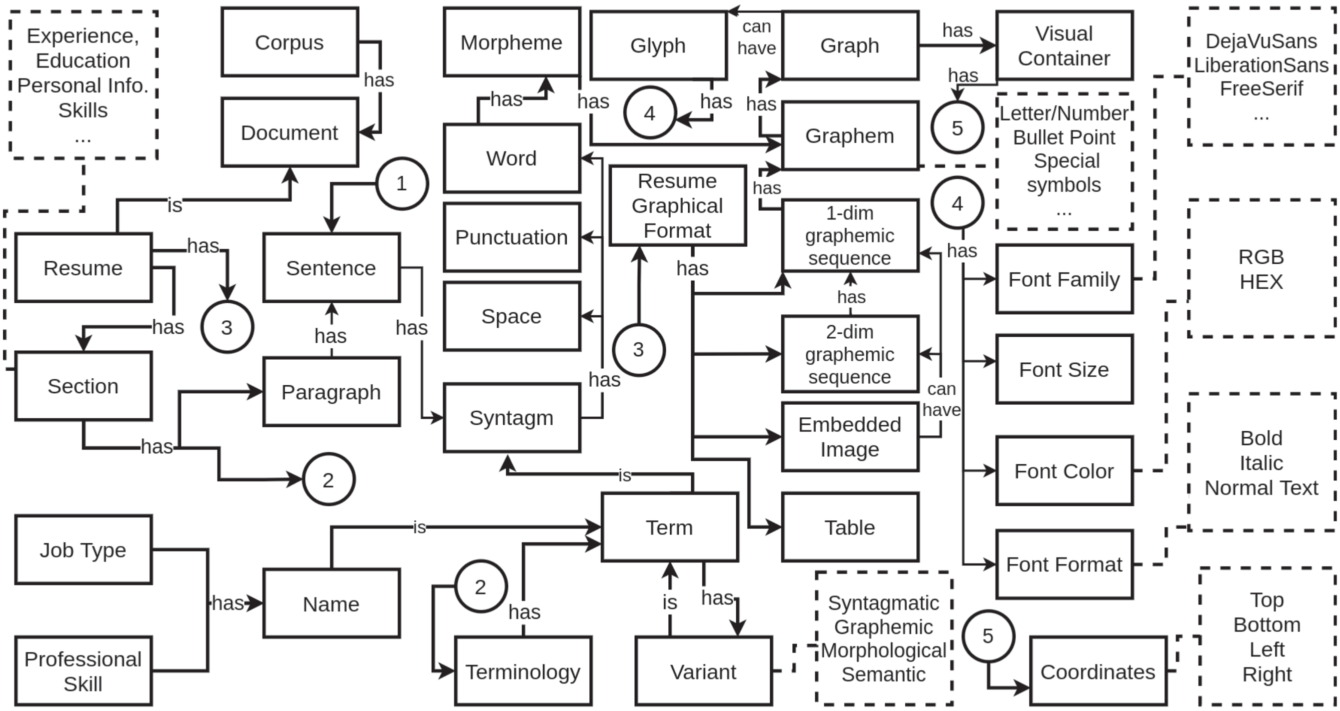


Fig. 1. Upper view of the ontology illustrating linguistic and graphical resume concepts. Circles with numbers (1, 2, 3, 4 and 5) are used to represent relationships between distant concepts in the diagram

$S_m = (s_{m,1}, \ldots, s_{m,l})$. Note that annotated sections can vary from CV to CV. We then generate automatic document segmentations that capture visual readability patterns [5] used by recruiters to identify CV sections.

To achieve this, we first represent every graphemic sequence (GS) in an annotated CV m in terms of its visual features, such as font family, font size, font color, and bold/italic/regular style. We cluster the GSs based on these features to obtain a set of clusters $C_m = C_{m,1}, C_{m,2}, \ldots, C_{m,N}$. For each cluster $C_{m,n}$, we obtain the spatial coordinates of its constituent sequences within the corresponding CV m and we utilize them as partitioning points to automatically segment the resume solely on the basis of its graphical or visual characteristics.

Thus, we obtain a tuple of hypothetical CV sections $H_{C_{m,n}} = (h_{C_{m,n},1}, h_{C_{m,n},2}, \ldots, h_{C_{m,n},k})$. Next, we apply a similarity metric designed to measure the level of resemblance between two segmentations, namely the S-similarity [24]. This measure enables us to compute the similarity between the recruiters' manual segmentation S_m and each hypothetical segmentation $H_{C_{m,n}}$, aiming to identify increasingly pertinent *format*-sensitive resume segmentations from the recruiters viewpoints.

The aforementioned process operates as an initial analysis that enables us to elucidate the connections between the graphical features of the CV and the manual segmentations conducted by recruiters. However, beyong this, it is of interest to identify the specific text-linguistic functions within the CV that are most strongly associated with these graphical features. Such knowledge could be leveraged to enhance the document segmentation process.

Therefore, after identifying the optimal segmentation cluster $C_{m,n}$ for each resume m with a visual perspective, we describe the GSs of such clusters in terms of text-linguistic functions (TLF), such as section title, section subtitle, list title, professional skill name, etc. We store this description in the form of triplets `<Cluster's Graphemic Sequence X, has, Linguistic Function Y>`, obtaining a set of descriptive triplets T_m for each resume. This description aims to unify the graphetic and linguistic properties of the CV, providing a deeper understanding of the relationships between its *grapholinguistic* content and the recruiters' manual segmentations.

Finally, we apply the Apriori algorithm [25] to the set T of all CVs triplets extracted, and we identify the set of the most frequent and relevant Text-Linguistic-Functions $\mathrm{TLF} = \{\mathrm{TLF}_1, \ldots, \mathrm{TLF}_R\}$, allowing the description of format-sensitive recruiter segmentations. These TLFs are a way to identify more optimal CVs segmentation coordinates.

E. Semi-supervised Construction of a Golden Corpus

At this point, we construct progressively a golden-corpus $\mathfrak{G}_r$ intended to represent the ground truth for each relevant $\mathrm{TLF}_r (r = 1, ..., R)$. From each resume m, we automatically extract 1-dim and 2-dim GSs corresponding to instances of TLF_r that are familiar to recruiters (e.g. CV section titles, subtitles, list names and so forth). Consequently, through a feature engineering process, we identify graphic/format markers GM_j and textual markers TM_i allowing to represent TLF_r such that $\mathrm{TLF}_r = \mathrm{GM}_{1,r}, \ldots, \mathrm{GM}_{j,r}, \mathrm{TM}_{1,r}, \ldots, \mathrm{TM}_{i,r}$. In other words, we identify the markers that best enable the modeling of text-linguistic functions most closely associated with the annotations made by recruiters. We evaluate the statistical significance of these markers through logistic regression models (optimized using maximum likelihood estimation). In our application case, logistic regression (LR) is advantageous due to its proven high explainability, its effectiveness at identifying relevant features (or markers) in relation to a predictor variable, and its robustness against over-fitting in small-datasets [26].

As an example, by applying our approach, we obtained the following pertinent markers related to the TLF "Section Title": **Font Size** (GM_1), i.e., usage of the given font size in the CV; **Font Family** (GM_2); **Color** (GM_3), i.e., the distance between the GS's color and the most frequent CV font color; **Bold** (GM_4); **Italic** (GM_5); **Capitalized** (TM_1); **Uppercased** (TM_2); **Term Title Variant** (TM_3), i.e., usage of term variants; **Frequency in resume titles** (TM_4), i.e., aggregated frequency of GS's words in the titles golden-corpus, being each word's frequency penalized by a factor $\sigma(-l)$, where l is where the word is located in the GS, and σ the sigmoid function; and **Frequency on resume common sentences** (TM_5), i.e., aggregated frequency of GS words in the negative samples of the golden-corpus. Fig. 2 illustrates different examples of the same section title in modern CVs.

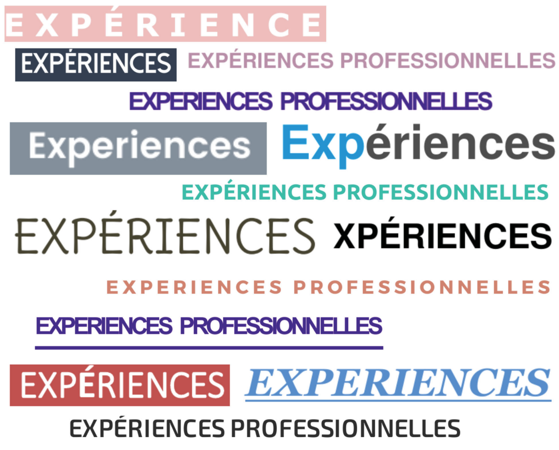

EXPÉRIENCE
EXPÉRIENCES
EXPÉRIENCES PROFESSIONNELLES
EXPERIENCES PROFESSIONNELLES
Experiences
Expériences
EXPÉRIENCES PROFESSIONNELLES
EXPÉRIENCES
XPÉRIENCES
EXPERIENCES PROFESSIONNELLES
EXPERIENCES PROFESSIONNELLES
EXPÉRIENCES
EXPERIENCES
EXPÉRIENCES PROFESSIONNELLES

Fig. 2. Examples of the section title "Experiences" extracted from contemporary French resumes. Color diversity, font family/size, bold font, and terms variations are some features that make a modern CV title stand out

In this manner, upon identifying the markers that optimally represent each TLF_r, every GS of a CV m is represented in terms of the TLF_r's related markers. From this representation, a second clustering process is performed per CV to automatically and exhaustively identify all the sequences related to TLF_r. We also estimate to which degree each resulting cluster represents a set of true instances of TLF_r, by computing an average possibility degree. To do this, we define a membership function f mapping tuples of TLF_r's markers to the interval [0,1], expressing to which possibility degree a given resume's GS can correspond to true TLF_r's instances. Clusters with an average possibility degree exceeding a threshold β_r are selected as more reliable TLF_r instances, and they become part of the golden-corpus $\mathfrak{G}_r$. Since those instances are concise, they can be validated manually to ensure the quality of the fine-tuning process. As a final step, using a clustering-based under-sampling approach [27], negative samples are also extracted from CVs to complement the corpus.

F. Sequences Formatting, BERT-based Models Fine-tuning, and Segmentation

Subsequently, each golden-corpus $\mathfrak{G}_r$'s GS is formatted in relation to TLF_r. This is done by concatenating [17] the GS's text, the fuzzified TLF_r's graphic/format, and textual markers. Markers are fuzzified in relation to five fuzzy categories (or linguistic variables) representing intervals of values, which were modeled with standard triangular functions. We use contrasting category names to improve the capability of BERT for interpreting the meaning of the fuzzified numerical markers (e.g., very small for 0 against very large for 1). By converting numeric markers into linguistic variables, fuzzification reduces the complexity of format-sensitive segmentation, which in our approach relies on recruiters' knowledge extraction. Extracted knowledge is inherently affected by phenomena like incomplete information and cognitive uncertainties [13] that fuzzification helps to overcome [28].

Based on the formatted TLF_r's golden-corpus, we use the distilled version of the multi-language BERT model [29], in order to fine-tune BERT-sequence classifiers for predicting resume's GSs as "Reliable Instances of TLF_r" or "Non reliable instances of TLF_r". Specifically, we use the *distilbert-base-multilingual-cased* version, as it is adapted for production environments and BERT classification tasks [29]. Fig. 3 provides the general architecture of our approach for representing and classifying resume's GSs.

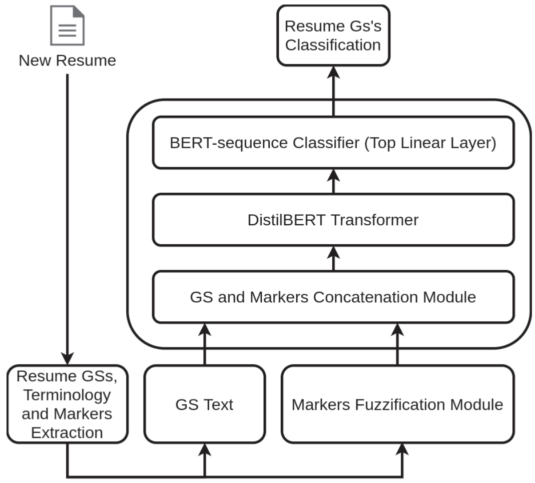

New Resume
Resume Gs's Classification
BERT-sequence Classifier (Top Linear Layer)
DistilBERT Transformer
GS and Markers Concatenation Module
Resume GSs, Terminology and Markers Extraction
GS Text
Markers Fuzzification Module

Fig. 3. General architecture to process GSs

After fine-tuning BERT-based models, we automate the task of segmenting CVs. Given a CV m, we extract its text and terminology to identify the GSs related to the most relevant TLF(s). The spatial coordinates of these GSs become reference coordinates for segmentation and serve as the initial coordinates for each section of the CV. Next, the remaining GSs on the CV are assigned to one of these sections based on three essential criteria. Firstly, we ensure that the GS is located at a minimum distance from the initial coordinate of the section. Secondly, we verify that the sequence is not spatially above the initial coordinate of the section. Finally, we confirm that the GS and the initial coordinate of the section are located in the same column of the CV. By following these steps, we can segment a CV into its various sections leveraging the most relevant TLFs that were identified throughout the proposed methodology. A workflow that illustrates this process is presented in Fig. 4.

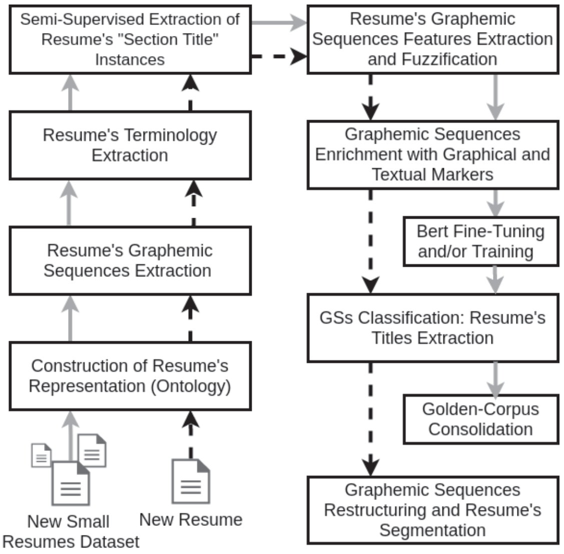

Semi-Supervised Extraction of Resume's "Section Title" Instances
Resume's Graphemic Sequences Features Extraction and Fuzzification
Resume's Terminology Extraction
Graphemic Sequences Enrichment with Graphical and Textual Markers
Bert Fine-Tuning and/or Training
Resume's Graphemic Sequences Extraction
GSs Classification: Resume's Titles Extraction
Construction of Resume's Representation (Ontology)
Golden-Corpus Consolidation
New Small Resumes Dataset
New Resume
Graphemic Sequences Restructuring and Resume's Segmentation

Fig. 4. A workflow derived by applying the proposed approach. Dotted black arrows illustrate how new resumes samples are processed based on a fine-tuned BERT-based model for extracting resume's titles. Solid gray arrows indicate how new small resumes' datasets can be exploited to fine-tune the BERT classifier

In the following section, we present the third and final axis of our approach, which is centered on evaluating the pertinence of the fine-tuned BERT models for new CV samples.

IV. Application

An assessment of the proposed framework was conducted in the setting of DSI Group's human resources department.

A. Experimentation Setting

Recruiters and their assistants annotated the 870 sections of a corpus of 150 resumes (dataset A), randomly selected from some hiring processes. These resumes belong to the French job market and contain a maximum of 2 pages with two or three columns each. We found that the most relevant linguistic function for describing their manual segmentations was the "Section Title" (average segmentation similarity of $S = 63\%$). We thus derived five formatted markers and five textual markers associated to those linguistic functions and assessed them by training LR models.

Next, we adapted six models to model the TLF: one LR classifier and five BERT-sequence classifiers. We used 70% of dataset A for training and the remaining 30% for evaluation. The results were validated using 10-fold stratified cross-validation. Afterward, with recruiters, we tested the validity of the adapted models using 153 newly annotated resumes, containing 923 sections (dataset B). Note that as our framework intends to exploit small datasets, we used a specific BERT fine-tune setting to avoid overfitting. First, the learning process is analyzed at the scale of steps instead of epochs, with early stopping. Second, we use the ADAM optimizer with a weight decay of 0.01. Third, each model is trained to a maximum of 3 epochs, with a learning rate of 2e-5, linear decay, and a batch size of 16. Lastly, due to known issues with instability in fine-tuning BERT models [30], each model is run 20 times and the average is reported.

B. Assessing Section Title Graphical and Textual Markers

As we realized that the graphical markers were not enough to represent the section titles in less-styled CVs, we complemented such markers with textual markers. Then, LR is used to estimate the statistical significance of markers in relation to the TLF "Section Title".

To evaluate the statistical significance, 30 resumes were initially selected from dataset A with random sampling, and each of the documents' GSs was represented by the TLF associated markers. This representation was used to train a LR optimized by 10-fold stratified cross-validation (35% GS titles, 65% GS not titles). An initial set of significant markers was identified.

Subsequently, a second LR was trained on the entirety of dataset A using an analogous procedure. A third LR was applied to dataset A filtering TLF instances, significantly reducing the number of GSs not corresponding to true titles. Table I shows the results of the second and third trained models.

C. Segmentation Evaluation

We assessed a total of 6 models for segmenting CVs by using different types of sequences enrichments. These models are presented and reported, as they represent various modeling approaches, ranging from a purely text-based analysis of the CV to a hybrid approach that incorporates both textual and graphical features for performing the segmentation process.

The base-line model is a LR which is useful for estimating the potential of the derived markers to describe the "Section Title" TLF. We also fine-tuned five BERT-sequence classifier models. The first model (**BERT** WM, without markers) receives as input the text of the GSs. The second model (**BERT+**AM, with all markers) is fine-tuned by enriching the GSs with all the derived markers. We also trained a model based on the GSs formatted with the most significant graphical markers (**BERT+**GM), and another one only enriching GSs with the most significant textual markers (**BERT+**TM). Then, we fine-tuned a model by enriching each GS with both the most significant graphical and textual markers (**BERT**

TABLE I. EVALUATION OF THE SIGNIFICANCE OF GRAPHICAL AND TEXTUAL MARKERS IN IDENTIFYING RESUME TITLES: LR WOF (LOGISTIC REGRESSION WITHOUT FILTERING) EVALUATED ON TITLE INSTANCES, NAMELY 17300 GSS WITH 870 GSS CORRESPONDING TO TRUE TITLES; LR WF (LOGISTIC REGRESSION WITH FILTERING) EVALUATED ON THE TITLE INSTANCES EXTRACTION APPROACH PROPOSED IN THE CURRENT STUDY FOR REDUCING NEGATIVE SAMPLES, SPECIFICALLY 2485 GSS CONTAINING 870 TRUE TITLES). P-VALUES WERE OBTAINED USING THE Z-TEST (WALD TEST)

		LR WOF			**LR WF**	
	Coefficients	**Std Err**	**p-value**	**Coefficients**	**Std Err**	**p-value**
GM_1 (Font Size)	4.61	0.68	<0.001	4.98	3.81	<0.001
GM_2 (Font Family)	−0.39	0.78	0.620	−1.18	0.85	0.210
GM_3 (Color)	1.86	0.34	<0.001	2.30	0.94	<0.001
GM_4 (Bold)	0.72	0.31	0.019	0.60	0.41	0.120
GM_5 Italic	−0.10	0.86	0.910	−0.22	0.40	0.770
TM_1 (Capitalized)	0.20	0.38	0.600	0.42	0.76	0.360
TM_2 (Uppercased)	1.52	0.34	<0.001	1.58	0.46	<0.001
TM_3 (Title Variants)	2.98	0.70	<0.001	4.11	0.42	<0.001
TM_4 (Freq. in CV titles)	5.63	0.59	<0.001	6.99	1.08	<0.001
TM_5 (Freq. in common sents.)	23.82	3.14	<0.001	24.80	0.88	<0.001
Intercept	−30.39	3.17	<0.001	−30.22	3.82	<0.001
R^2	0.72			0.76		

TM+GM).

We evaluated the performance of the fine-tuned models in the task of segmenting the unknown resumes of dataset B. We used the Recall, Precision, and F1-Score metrics to determine the models performance in segmenting CVs sections. Table II provides the results of our experiments. We show in Fig. 5.a how the best model (**BERT** TM+GM) evolved over the course of the evaluation on dataset B. In Fig. 5.b, we illustrate the validation and training loss of the corresponding model.

V. Discussion

The LR models applied to estimate the significance of markers, evidenced that markers GM_1 (Font size), GM_3 (Font color), GM_4 (Bold), TM_2 (Uppercased), TM_3 (Term Title Variant), TM_4 (Frequency in resume titles) and TM_5 (Frequency on resume common sentences), were the most significant with an average confidence level of 95%. It illustrates how potentially meaningful graphical and textual markers can be derived from recruiters' segmentations with a grapholinguistics perspective. This insight is supported by the R^2 value of both regressions, which reflects significant adjustment even using a less balanced data set (LR WOF). These results are also an evidence of the text-linguistic function "Section Title" potential, which is simpler to model and highly effective for identifying the structure of modern CVs.

In particular, the graphic markers GM_1, GM_3, along with GM_4 were identified as the most significant, reflecting the current frequent tendency of candidates to add titles with strong color contrasts (GM_3), larger sizes (GM_1), and special styles such as bold (GM_4) or, to a lesser extent, italics (GM_5). Even if a resume title can be written with a different font family (GM_2), this is not frequent, according to our results.

Regarding textual markers, marker TM_2 reveals a consistent applicant's tendency to write resume titles in uppercased letters. Examining marker TM_3 centered on the identification of title terminological variants, we found a statistical indication revealing that a portion of candidates provides ambiguous terms, which can highly degrade the performance of automated resume analysis methods. Consequently, a terminological analysis could be used to work on such sources of uncertainty. In addition, we observed that the TM_4 marker provides a clearer distinction between GSs corresponding to titles. In contrast, the TM_5 marker allows identifying those (more numerous) not corresponding to titles from the point of view of the CV textual content.

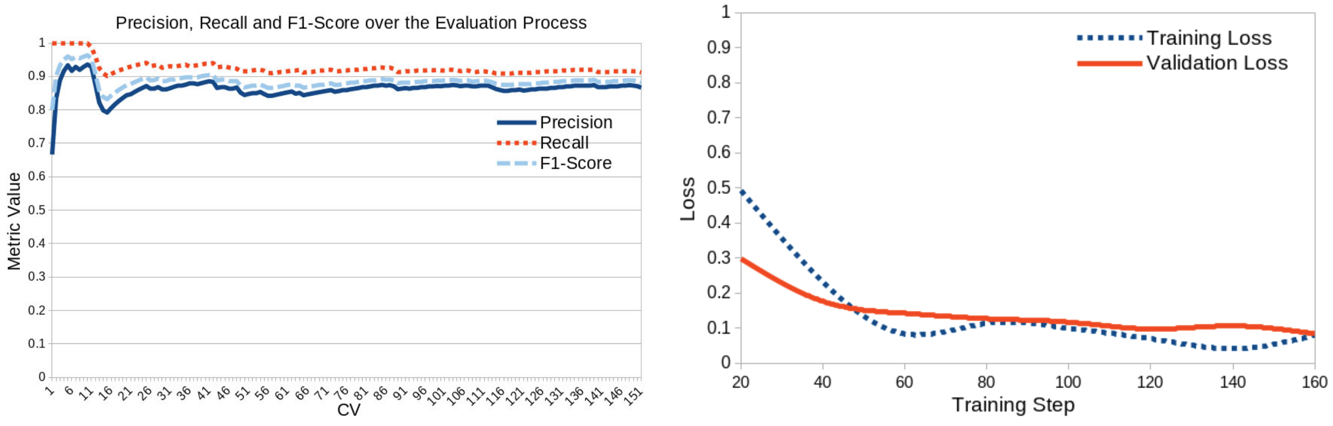
Concerning the models' performance to segment CVs, the LR showed relatively stable behavior in dataset B of CVs. However, it did not perform as well as the BERT-based models. This illustrates the robustness of the derived graphical and textual markers associated with the text-linguistic function "Section Title." In the case of the BERT-based models, we first point out that it is possible to train them minimizing phenomena as overfitting or underfitting, which occur when there are small data sets or non-optimal training parameters. Second, we found that the model **BERT** WM, which uses only the text of the GSs, can fit the training dataset much better, which could be explained by the small size of the CV titles, representing a reduced complexity. However, this model's performance drops when evaluated on dataset B of unknown CVs. We observed that due to the large terminological variability of titles from resume to resume, several new and unknown titles were not adequately interpreted by such a model.

With respect to the models whose GSs were formatted with the markers, we found that textual enrichment with the graphical markers (**BERT+**GM) made the BERT-based model slightly more robust for unknown CVs, allowing to produce correct predictions on very rare titles not well managed by the only-text model. However, when less styled resumes are processed, the model introduces new types of mistakes because attention focuses on graphical features that aren't relevant to these types of CVs.

On the other hand, better results were obtained in relation to the non-formatted and graphical model, by the **BERT** TM+GM model, which combines the most significant graphical and tex-

TABLE II: PRECISION, RECALL AND F1-SCORE RESULTS FOR EACH MODEL EVALUATED ON THE TESTING SAMPLES OF DATASET A (45 CVS) AND THE SAMPLES OF DATASET B (153 CVS)

Model	**LR**	**BERT** WM	**BERT+**AM	**BERT+**TM	**BERT+**GM	**BERT** TM+GM
Training Dataset						
Recall	0.88	**0.97**	0.93	0.94	0.92	0.95
Precision	0.88	0.92	0.91	0.91	0.92	**0.93**
F1-Score	0.88	**0.95**	0.92	0.93	0.92	0.94
New Dataset						
Recall	0.86	0.88	0.76	0.85	0.88	**0.91**
Precision	0.84	0.85	0.73	0.82	0.86	**0.87**
F1-Score	0.85	0.86	0.74	0.84	0.87	**0.89**



(a) Best Model (**BERT** TM+GM) performance evolution on dataset B

(b) Best **BERT** TM+GM model training and validation loss

Fig. 5. Best **BERT** TM+GM model assessment during the evaluation process

tual markers. This model captures more pertinently intrinsic patterns between the titles text, graphical markers, and textual markers, achieving the best performance. Finally, the model **BERT+**AM obtained good results on the training dataset but not on dataset B, as too many markers are embedded into each GS's text, making the classification task much more complicated, and the dataset size not enough to perform a general fine-tuning-learning process.

VI. Conclusions and Perspectives

The results of this study suggest that the proposed framework makes it possible to optimize the segmentation of CVs through the format-sensitive approach. Starting from the analysis of the organizational context in which CVs are processed, we constructed a grapholinguistic representation of resumes, to identify and evaluate graphical and textual markers associated with relevant linguistic functions, improving the segmentation process from the recruiters perspective. Moreover, the process for adapting the BERT language model is likely to enhance CV segmentation, without requiring additional complex layers or architectures.

It's worth emphasizing that the attainment of such outcomes was significantly influenced by the preliminary steps of the proposed methodological approach. This approach allocates a pivotal role to the examination of knowledge and language utilized by domain experts who analyze CVs. Consequently, this facilitated the derivation and consolidation of a visual and written CV representation and segmentation procedure, customized and tailored with the distinct requirements of the organizational context in which the documents are processed.

To conclude, the defined experiments showed both statistically significant and understandable findings, providing a foundation for further research directions. Firstly, we aim to extend our approach to resumes with more graphetic elements and heterogeneous content. Secondly, we will also examine how logical relationships between markers, revealed by BERT model learning, can influence the decision-making processes in resume analysis. Finally, we will investigate the optimal integration of other types of term variants in the segmentation process, such as derivational ones, in order to reduce the uncertainty caused by term variability.

References

[1] F. de Saussure, *Course in General Linguistics*. New York: Columbia University Press, 1959, translated by Wade Baskin.

[2] D. Meletis and C. Dürscheid, *Writing Systems and Their Use. An Overview of Grapholinguistics*. Berlin/Boston: De Gruyter, 2022.

[3] Y. Haralambous, *Fonts & Encodings. From Advanced Typography to Unicode and Everything in Between*. Sebastopol, CA: O'Reilly, 2007.

[4] Y. Haralambous and M. Dürst, "Unicode from a linguistic point of view," in *Proceedings of Graphemics in the 21st Century, Brest 2018*, Y. Haralambous, Ed. Brest: Fluxus Editions, 2019, pp. 167–183.

[5] L. Rello, M. Pielot, and M.-C. Marcos, "Make it big! the effect of font size and line spacing on online readability," in *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 3637–3648.

[6] V. Bhatia, P. Rawat, A. Kumar, and R. R. Shah, "End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT," 2019, arXiv:1910.03089.

[7] H. Sajid, J. Kanwal, S. U. R. Bhatti, S. A. Qureshi, A. Basharat, S. Hussain, and K. U. Khan, "Resume parsing framework for e-recruitment," in *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2022, pp. 1–8.

[8] D. Çelik, "Towards a semantic-based information extraction system for matching résumés to job openings," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 24, no. 1, pp. 141–159, 2016.

[9] V. S. Kumaran and A. Sankar, "Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping expert," *International Journal of Metadata, Semantics and Ontologies*, vol. 8, pp. 56–64, 05 2013.

[10] M. Tikhonova and A. Gavrishchuk, "NLP methods for automatic candidate's CV segmentation," in *2019 International Conference on Engineering and Telecommunication (EnT)*, 2019, pp. 1–5.

[11] C. Ayishathahira, C. Sreejith, and C. Raseek, "Combination of neural networks and conditional random fields for efficient resume parsing," in *2018 International CET Conference on Control, Communication, and Computing*, 2018, pp. 388–393.

[12] J. Liu, Y. Shen, Y. Zhang, and S. krishnamoorthy, "Resume parsing based on multi-label classification using neural network models," in *Proceedings of the 6th International Conference on Big Data and Computing*, ser. ICBDC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 177–185.

[13] E. Pavlick and T. Kwiatkowski, "Inherent disagreements in human textual inferences," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 677–694, 2019.

[14] X. Li, H. Shu, Y. Zhai, and Z. Lin, "A method for resume information extraction using bert-bilstm-crf," in *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, 2021, pp. 1437–1442.

[15] A. Barducci, S. Iannaccone, V. La Gatta, V. Moscato, G. Sperlì, and S. Zavota, "An end-to-end framework for information extraction from italian resumes," *Expert Systems with Applications*, vol. 210, p. 118487, 2022.

[20] D. Hovy and D. Yang, "The importance of modeling social factors of language: Theory and practice," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 588–602.

[16] M. Trabelsi, Z. Chen, S. Zhang, B. D. Davison, and J. Heflin, "Strubert: Structure-aware bert for table search and matching," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 442–451.

[17] K. Gu and A. Budhkar, "A package for learning on tabular and text data with transformers," in *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*. Mexico City, Mexico: Association for Computational Linguistics, 2021, pp. 69–73.

[18] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3651–3657. [Online]. Available: https://aclanthology.org/P19-1356

[19] D. Martin Jr., V. Prabhakaran, J. Kuhlberg, A. Smart, and W. S. Isaac, "Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context," 2020, arXiv:2006.09663.

[21] C. M. Zapata and F. Arango, "The UNC-method: a problem-based software development method," *Ingeniería e Investigación*, vol. 29, pp. 69–75, 2009.

[22] D. Cram and B. Daille, "Terminology extraction with term variant detection," in *Proceedings of ACL-2016 system demonstrations*, 2016, pp. 13–18.

[23] K. T. Frantzi, S. Ananiadou, and J. Tsujii, "The C-value/NC-value Method of Automatic Recognition for Multi-word Terms," *Research and Advanced Technology for Digital Libraries*, vol. 1513, pp. 585 – 604, 03 2002.

[24] C. Fournier and D. Inkpen, "Segmentation similarity and agreement," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, Jun. 2012, pp. 152–161.

[25] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94, vol. 1215. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, p. 487–499.

[26] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, pp. 673–705, Nov 2019. [Online]. Available: https://doi.org/10.1007/s10458-019-09408-y

[27] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409-410, pp. 17–26, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025517307235

[28] S. Thaker and V. Nagori, "Analysis of fuzzification process in fuzzy expert system," *Procedia Computer Science*, vol. 132, pp. 1308–1316, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918307798

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," in *NeurIPS EMC^2 Workshop*, 2019, https://arxiv.org/abs/1910.01108.

[30] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi, "Revisiting Few-sample BERT Fine-tuning," in *International Conference on Learning Representations*, 2021, https://arxiv.org/abs/2006.05987.