# Fuzzy-Oriented Terminological Analysis to Extract Job Offer Information Relevant to Candidate Ranking

Albeiro Espinal[1,2], Yannis Haralambous[1], Dominique Bedart[2], and John Puentes[1]

[1]IMT Atlantique, Lab-STICC, CNRS UMR 6285, Brest, France
[2]DSI Global Services, Le Plessis Robinson, France

***Abstract*—*An automated resume ranking system selects and sorts relevant resumes from those sent in response to a job offer (JO). During the screening and elimination process, resume content is largely analyzed, while JO details are only marginally considered. In this sense, existing resume ranking approaches lack the accuracy necessary to detect relevant information in JOs, which is imperative to ensure that selected resumes are relevant to the JO. This study examines the uncertainty-based estimation to assess 16 textual markers applied to extract relevant terms in JOs—10 textual markers obtained by examining the behavior of expert recruiters and 6 from the literature—based on two approaches: fuzzy logistic regression and fuzzy decision trees. Results indicate that, globally, fuzzy decision trees improve the F1 and recall metrics by 27% and 53% respectively, compared to state-of-the-art term extraction techniques.*

***Keywords*—*Recruiter's Behavior Modeling, Textual Relevance Marker Assessment, Term Extraction, Uncertainty Measure, Fuzzy Machine Learning.***

## I. Introduction

Job offers (JOs) and curriculum vitaes (CVs) are the documents through which recruiters and candidates interact, as part of a recruiting process. An important stage carried out by recruiters is the "Screening Phase" that evaluates the CVs of candidates to identify those who are qualified for a job. Analyzing both the main requirements of a new JO and the skills of candidates expressed in their CVs can be very complex. This is even more the case when recruiters receive dozens, hundreds or even thousands of candidates resumes [1]. In order to reduce such complexity, multiple artificial intelligence models have been developed to analyze and rank CVs for a given JO.

Although several models have been proposed, the automatic ranking of CVs remains a difficult task. This is due, in part, to three issues that have rarely been examined in the literature. First, relevant information in the JO is not optimally identified, generating irrelevant rankings with respect to essential requirements [2]. Secondly, under-representation of the organizational context surrounding JOs tends to break this type of systems over time [2]. Thirdly, since writing JOs engages human cognition, the expressed information is highly susceptible to uncertainty phenomena like ambiguity [3], which could render AI models ineffective [4]. Being still an active research field [5], the study of uncertainty and its characterization, is fundamental to investigate the extraction of relevant terms from JOs.

Organizational context in order to define a set of relevant textual markers based on recruiters' strategies to select significant JOs' information, and estimation of the consistency of these markers has already been studied [6]. Nevertheless a question remains concerning the quantitative assessment of identified markers' uncertainty, which is the goal of this work. Our study intends to assess the pertinence of automatically identified relevant JO terms, applying two machine learning models—fuzzy logistic regression and fuzzy decision trees—focused on the quantification of uncertainty. The present article is an extension of a previous work [7] and is organized as follows. Section II describes the related state of the art. In Sections III and IV we summarize some key aspects of our previous work. Section V describes the proposed uncertainty assessment of textual markers. Experimental results are presented in Section VI. Discussion, conclusions and perspectives are presented in Sections VII and VIII.

## II. State of the Art

CV ranking systems carry out three processing stages: (a) CV and JO pre-processing, (b) CV and JO representation, and (c) automatic ranking of CVs with respect to the content of the corresponding JO. The underlying documents are pre-processed by extracting text from digital files (.pdf, .doc, .txt, among others). Then extracted texts are standardized by eliminating noisy symbols, segmenting the documents, and making semantic annotations [1], as well as restricting the vocabulary by stopword deletion [8]. Pre-processed documents can be represented based on n-gram models [1], the bag-of-words model [1], ontologies [9] and/or word embeddings [10]. From these representations, different approaches can be used to determine suitable CVs with respect to a given JO. They can rely on recruiters' feedback [1], neural architectures [10] and/or transformer models [11].

These methods, however, do not focus on extracting relevant information from the JO before ranking resumes. Some domain-independent methods have been proposed in order to identify the relevant information of an individual document, such as a JO. For instance, RAKE (Rapid Automatic Keyword


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Extraction) [12] and FRAKE (Fusional Real-Time Keyword Extraction) [13] model a document as a graph and extract pertinent terms based on centrality measures (such as node degree), frequency of words and textual characteristics. On the other hand, the approach of YAKE! (Yet Another Keyword Extraction) [14] algorithm extracts relevant terms in five basic stages going from simple-term extraction to the ranking of multi-word terms from highest to lowest relevance. In addition, recent studies have evaluated the feasibility of the popular deep learning BERT model for identifying relevant terms in individual documents, including job offers [15].

Furthermore, uncertainty, a key concern of natural language processing [4] concerns the lack of information about an event or situation. Among frequently studied approaches to determine uncertainty we have probability models [5], as well as possibility theory and fuzzy logic models [16]. Contrary to probability-oriented models, fuzzy models assume that probability distributions cannot be obtained for fuzzy data. In this regard, linear and non-linear fuzzy machine learning models have been proposed to deal with uncertainty. Linear models, such as fuzzy logistic regression, are utilized to deal with uncertainty as fuzziness and not as randomness [4]. Also, non-linear models as fuzzy decision trees have been studied, including ambiguity and vagueness metrics to estimate uncertainty [3].

We propose to evaluate the uncertainty of textual markers that indicate the relevance of information in JOs based on recruiters' knowledge. The proposed evaluation compares fuzzy linear and non-linear machine learning methods, which are appropriate to investigate the uncertainty question. Indeed, because of their possibilistic foundations at the crossroad of fuzzy sets and probability, they provide a simple and convenient setting for handling subjective tasks, as the automatic identification of relevant terms in JOs. Moreover, these types of models can be trained on small datasets to evaluate feature relevance.

## III. Representation of Job Offers

In order to evaluate the uncertainty of textual markers it is first necessary to specify the organizational context of JOs, to analyze what is relevant for recruiters in this type of document, and to extract textual markers that represent relevant information [6].

### A. *Organizational Context*

The representation of societal contexts in machine learning models can largely be improved, allowing those models to become more adaptable to dynamic changes in organizations [17]. This is a critical aspect in our work, given that context strongly influences recruiter behavior [18]. We began thus by representing the recruiters' context before analyzing their strategies related to information relevance in JOs. To this end, we used the UNC-method for representing organizational contexts. Originally used in the field of software development, the UNC-method integrates a set of components for representing the organizational context and identifying the main sources and solutions for specific problems. In our context, this method has been applied to the analysis of problems related to extracting the most relevant information from JOs.

As specified by this analytical methodology, we conducted an open dialogue with recruiters, specifying the entities and relationships that impact the JOs' life-cycle. The process involved the creation of traditional components, such as UML diagrams, and non-traditional ones, such as KAOS Objective Diagrams and pre-conceptual schemas. These elements were created as follows [19]:

- **Pre-conceptual schema:** It identifies and defines fundamental concepts related to the life-cycle of a JO from the perspective of the organization and recruiters. In our approach, the creation of this component is essential to adequately represent the JOs.
- **Domain model:** It identifies the main attributes and relationships of a JO. Basic elements that compose the JO can be identified through a linguistic examination, which allows for a more accurate representation. Based on these first two components, we can generate a mother ontology that represents JOs from the organization and recruiters' viewpoints. Then, new components are created to further enhance the context-driven representation.
- **Goals:** In particular, recruiters' goals are analyzed within the framework of a recruitment process, which is inherently associated with the life-cycle of JOs. In a *KAOS goals diagram*, the general goals are placed at the root of the hierarchy, while specific ones are at the bottom. By representing such goals, it is easier to identify aspects of JOs that are relevant.
- **Process diagrams:** Its purpose is to depict the organizational processes associated with the life-cycle of a JO. These types of diagrams are particularly useful for identifying relevant aspects of a JO from the organization and recruiters perspective.
- **Fishbone chart:** It is a visual representation of critical concerns associated with handling JOs. This diagram represents the problems associated with identifying essential information of the JO. In this way, it is possible to identify dynamics within an organization that are inconsistent or inconvenient if they are not reflected by machine learning systems that process JOs automatically.

Applying [19], the various diagrams are unified by means of a **Process Explanatory Table**. By doing so, we are able to gain a comprehensive understanding of how to extract relevant information from JOs. As a result, the main entities, actors, processes, goals, and organizational issues associated with JO management were identified. Also, a mother ontology was derived, which is schematically described in the following section.

### B. *Ontology Derivation*

We define a mother ontology as a large ontology of module specifications. We used a mother ontology to represent the main concepts and relationships inherent to the recruiters' context and to JOs. Additionally, existing ontologies related to the


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

particular organizational context were integrated into it. This was the case of the internal professional skills ontology of DSI Group which contains the specification of more than 36,000 professional skills, the European ontology of professional skills ESCO,[1] the professional skills and job types frameworks of O*NET,[2] CIGREF,[3] and ROME,[4] based on text-to-RDF-triple transformations [20]. The integration of these ontologies has been achieved by the use of a hybrid approach based on Bidirectional Encoder Representations from Transformers (BERT) [21], an analysis of terminological variation [22] and measures of ontology quality [23].

The process of integrating external ontologies was conducted as follows. First, for each pair of concepts belonging to different ontologies, we defined three types of possible relationships:

- **Close Match:** When two concepts have a BERT similarity degree greater than a defined threshold $\alpha \in [0, 1]$, we consider that there is a close similarity between them.
- **Exact match:** As well as fulfilling the close match condition, at least one pair of concepts that are neighbors of the two main concepts, has the exact or close match relationship. Fig. 1 illustrates an example of Exact Match between two concepts, one belonging to the DSI Group ontology and the other belonging to the ESCO ontology.
- **There is no match:** If neither of the previous two relationships can be established, it is assumed that there is no evidence to conclude that there is a close meaning between the pair of concepts.

We highlight that we have extended the BERT method to manage these comparisons, by performing a terminological analysis of variants, in order to determine if terms of a concept correspond to variants of the second concept's terms.

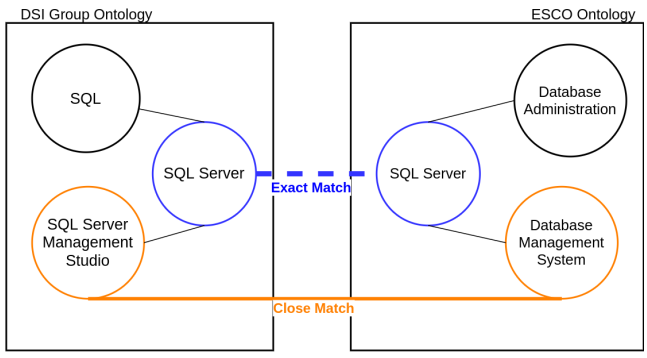


Fig. 1. Example of an exact match between two concepts belonging to different ontologies.

It is important to note that in this compound ontology we also specified the structure of JOs in terms of linguistic concepts such as sections, paragraphs, sentences, syntagms, terms, words, etc. Additionally, the usual relations of synonymy, meronymy and hyponymy were used, whenever appropriate, as relations between concepts. This enabled us to construct a more structured fuzzy model of the natural language contained in JOs by representing the basic constituents, as it has been suggested by [24]. An upper view of the ontology is presented in Fig. 2.

### C. *Analysis of Recruiters Viewpoints*

Based on the organizational context representation using the previous ontology, we conducted an experiment in order to analyze recruiters' strategies related to the selection of essential information in JOs. They were asked to annotate CVs by highlighting relevant terms. To represent the description of each recruiter's observed actions, the controlled language proposed by [19] was used. It allows us to represent actions sequentially, as triples of the form <subject, verb, predicate>.

We categorized those actions as active (e.g., <recruiter, selects, term>) or passive (e.g., <recruiter, avoids, term> or <recruiter, avoids, JO_section>). Once the annotations were described in a controlled manner, the Apriori algorithm [25] was used to identify action sub-sequences that the recruiter performed systematically. These sub-sequences of actions describe behavioral patterns, formalized as semantic rules, using the mother ontology described in section III.B. Obtained rules represent textual relevance markers in JOs. Fig. 3 illustrates an example of the analysis of recruiters viewpoints.

## IV. Textual Markers

In this section, we present briefly the evaluated textual markers and introduce the linguistic representation of JOs in our approach.

### A. *JO Terminology Extraction*

Considering that terms are defined as functional classes of lexical units used in discourse [22], JOs' relevant terms were identified by the weirdness ratio that measures their termhood (see below, as well as [22], [26]). We achieved this by exploiting common morphosyntactic patterns, as previously identified in multiple experimental studies [22]. Our patterns are mostly nominal phrases and we apply them through a syntactic analysis tailored for the French language and a parallel syntactic analysis tailored for the English language. Table I provides examples of patterns exploited for extracting the terminology of the JO.

Patterns are represented by regular expressions. Lemmatizing each JO's word is previously done using a part-of-speech tagger. Once morphosyntactic patterns are applied, JO terms are identified according to their weirdness ratio (WR(t)), which is defined by the following equation:

$$WR(t) = \frac{f_{norm}(t,\ C)}{f_{norm}(t, G)} \quad (1)$$

where $f_{norm}(t,\ C)$ stands for the relative frequency of the term t in a corpus of job offers and $f_{norm}(t, G)$ correspond to the relative frequency of the term in a general language corpus [22].

[1]https://esco.ec.europa.eu/en
[2]https://www.onetonline.org
[3]https://www.cigref.fr
[4]https://www.pole-emploi.fr/employeur/vos-recrutements/le-rome-et-les-fiches-metiers.html


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

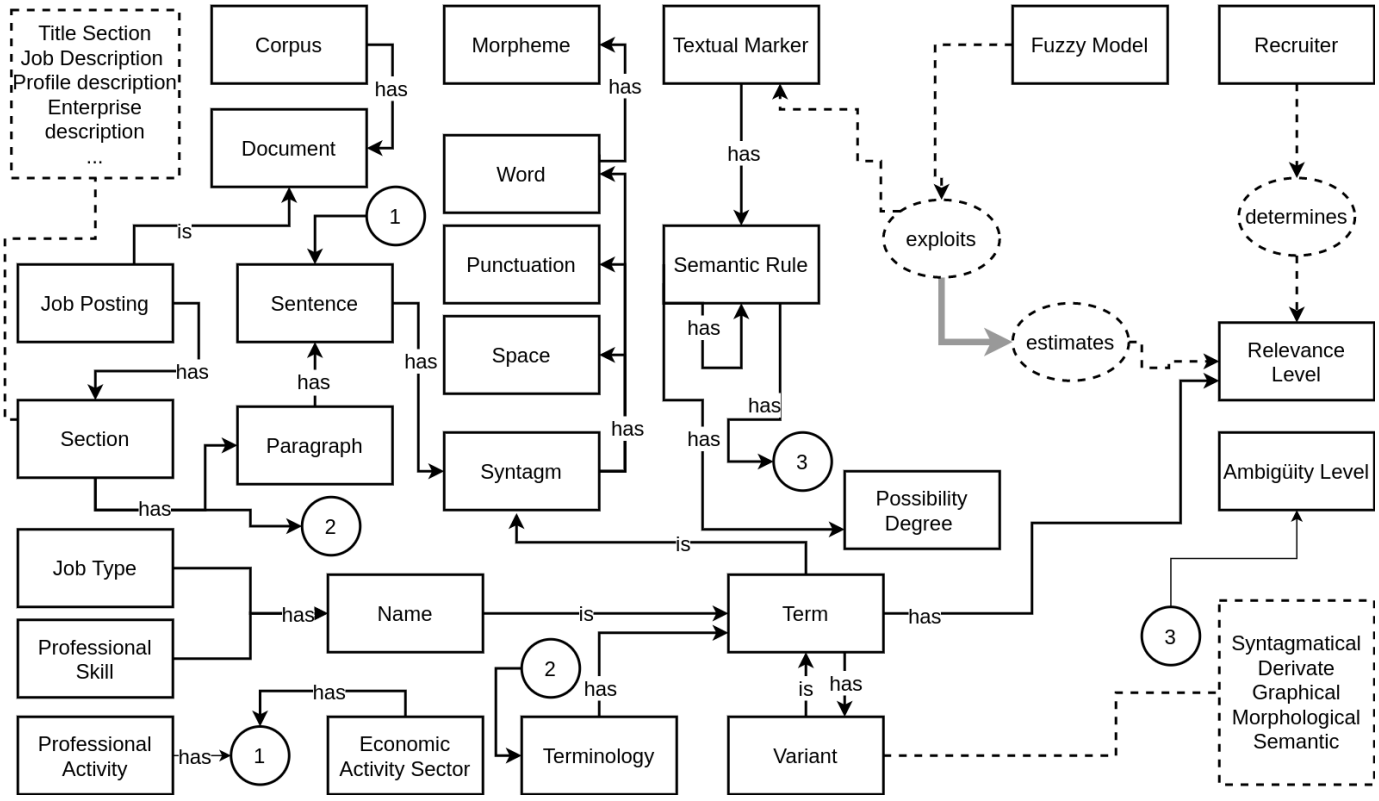


Fig. 2. Upper view of the mother-ontology created from the representation of the organizational context according to the principles of [19].

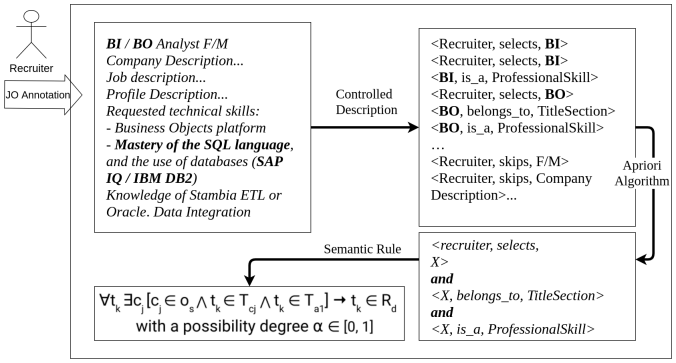


Fig. 3. Analysis process of recruiters' viewpoints, in the following order: the recruiter annotates a JO, the annotation is described in a controlled language, the Apriori algorithm is used to identify systematic behaviors, and semantic rules (textual markers) are derived.

TABLE I
EXAMPLES OF MORPHOSYNTACTIC PATTERNS AND RELATED TERMS. THE LETTER N STANDS FOR NOUN OR ABBREVIATION ACTING AS A NOUN, THE LETTER P FOR PREPOSITION, AND THE LETTER A STANDS FOR ADJECTIVE.

| JO Morphosyntactic Patterns | | |
|---|---|---|
| **#** | **Pattern** | **Example** |
| 1 | N | ETL |
| 2 | N N | Tableau Software |
| 3 | A N | Technical Specifications |
| 4 | N P N | Knowledge of Stambia |
| 5 | A N N | International Consulting Firm |
| 6 | N P N N | Knowledge of Stambia ETL |

The final step is to analyze the terms variants of the JO. Based on this analysis, we can describe the relationships between the simple JO terms and the more complex ones. Taking into account the experimental studies presented in [22], we identify four types of terminological variants:

- **Morphological variants:** Each simple term (a single word) is analyzed to determine whether it contains prefixes or suffixes.
- **Compound variants:** Analysis of complex terms is performed in order to identify their heads or hierarchical tree structures. As an example, the head of the multi-word term "Professional Experience" is "Experience".
- **Graphemic variants:** There is a systematic identification of terms differing as the result of spelling errors. For example, the term "Mstery of the SQL Language" is a graphemic variant of the term "Mastery of the SQL Language".
- **Semantic variants:** The BERT model allows the identification of JO terms with closely related meanings.

The analysis of terminological variants in this study enables us to reduce the diversity of terms in the JO and optimize the performance of the fuzzy machine learning models that are trained to assess textual markers.


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

### B. Definitions

Based on the previous elements, we provide the following preliminary definitions.

Let $d_i$ be a JO belonging to a corpus $C$ and $T_{d_i} = \{t_1, t_2, \ldots, t_n\}$ the set of terms of $d_i$. Let $R_{d_i} \subseteq T_{d_i}$ be the *set of most relevant terms* in $d_i$. Each term $t_i \in R_{d_i}$ is considered as relevant under a possibility degree $\alpha_{t_{k,i}} \in [0, 1]$.

Let $A_{d_i} = \{a_1, a_2, \ldots, a_m\}$ be the set of sections of $d_i$ (job description, profile details, etc). Each section $a_i$ can be represented by a subset of terms from $T_{d_i}$. A term can belong to multiple sections. Let $E_{d_i} = \{e_1, e_2, \ldots, e_p\}$ be a set of qualifying adjectives and nouns that are linked to a subset of terms in $T_{d_i}$ by syntax dependencies.

Let $O = \{o_1, o_2, \ldots, o_s\}$ be a set of ontologies (as the one presented in Section III). Let $c_{o_s} = \{c_{s,1}, c_{s,2}, \ldots, c_{s,k}\}$ be the set of concepts of ontology $o_s$ and $T_{c_j} = \{t_{j,1}, t_{j,2}, \ldots, t_{j,l}\}$ the set of terms lexically representing concept $c_j$ in a given language.

### C. Description of Textual Markers

In this section, we provide a summary of the derived textual markers [6] evaluated applying the proposed approach. Each marker provides a possibility degree for each JO term of becoming relevant. Textual markers $TM_1$ to $TM_{10}$ have been obtained from recruiters behaviors, while markers $TM_{11}$ to $TM_{16}$ correspond to those of the YAKE! (Yet Another Keyword Extraction) algorithm [14], found to be suitable, compared to other available algorithms in the literature. It is a domain-independent method applied in our case to identify potential relationships between textual markers and the context specificities of JOs.

We emphasize that some textual markers, such as $TM_8$, $TM_{11}$ and $TM_{12}$, provide possibility degrees that can be estimated using normalized equations. Other markers, such as $TM_1$, $TM_2$ and $TM_5$, provide a possibility degree of 1 if the semantic rule conditions are met, or 0 otherwise. Nevertheless, the maximum possibility degree of any textual marker is limited by its ambiguity level (more information about ambiguity can be found in Section V.C).

#### 1) Title Sections ($TM_1$):

| |
|---|
| Any term in the title that bears similarity to a term indicating professional skills or job types may potentially qualify as relevant. |

Let $a_1 \in A_{d_i}$ be the title section of $d_i$. Let $t_{a1} = \{t_1, t_2, \ldots, t_u\}$ be the set of terms contained in $a_1$. Lexically, $T_{c_j}$ is the set of terms that represent a professional skill or job type concept $c_j$ in the ontology $o_s$. Therefore:

$$\forall t_k \exists c_j [c_j \in o_s \wedge t_k \in T_{c_j} \wedge t_k \in t_{a1}] \rightarrow t_k \in R_{d_i} \quad (2)$$

with a possibility degree $\alpha_{t_{k,1}} \in [0, 1]$.

#### 2) Terms Representing Professional Skills in a Job Description Section or Profile Description Section ($TM_2$):

| |
|---|
| Terms representing professional skills used in job descriptions or profile descriptions are more likely to be chosen as relevant terms. |

Let $s_2$ and $s_3$ be the sets of terms used in the job description section and the profile description section, respectively. Set $t_k \in T_{d_i}$. Let $T_{c_j}$ be the set of terms used to represent a professional skill concept $c_j$ in the ontology $o_s$. We request that:

$$\forall t_k \exists c_j ((t_k \in s_2 \vee t_k \in s_3) \wedge t_k \in T_{c_j}) \rightarrow t_k \in R_{d_i} \quad (3)$$

with a possibility degree $\alpha_{t_{k,2}} \in [0, 1]$.

#### 3) Relevance of Job Posting Sections ($TM_3$):

| |
|---|
| As a general rule, recruiters are more likely to select terms from the title, job description, and profile description sections, rather than from other sections (company description, contract details, etc.). |

As we don't require terms to be professional skills, this marker does not overlap with markers $TM_1$ and $TM_2$. Let $S = s_1 \cup s_2 \cup s_3 \subseteq T_{d_i}$, where: $s_1$ is the set of terms of the title section; $s_2$ is the set of terms of the job description section; and $s_3$ is the set of terms of the profile description section. Let $t_m \in T_{d_i} \cap S$. Then, we request that:

$$\forall t_m \forall t_n (t_m \in T_{d_i} \wedge t_n \notin S) \rightarrow (P(t_m \in R_{d_i}) > P(t_n \in R_{d_i})) \quad (4)$$

with a possibility degree $\alpha_{t_{k,3}} \in [0, 1]$. $P(t_* \in R_{d_i})$ denotes the possibility of $t_*$ being chosen as a relevant term.

#### 4) Terms Dependent on Pertinence Expressions ($TM_4$):

| |
|---|
| A relevant term is more likely to be one that bears a syntax dependency with a syntagm of the JO. |

- Let $t_k \in T_{d_i} \cap T_{c_j}$ for some $c_j$.
- We define a "relevant expression" $e_m$ as a syntagm that the recruiter employed in the JO (i.e., <u>excellent</u> C# skills, <u>good understanding</u> of Kubernetes). Assume that $e_m$ is syntactically dependent with $t_i$. Specifically, let $t_k$ be a qualifying adjective or a noun modifier directly dependent with $e_m$. Then:

$$\forall t_k \exists e_m (t_k \in T_{d_i} \wedge e_m \in E_{d_i} \wedge \text{is\_dependent}(t_k, e_m)) \rightarrow t_k \in R_{d_i} \quad (5)$$

with a possibility degree $\alpha_{t_{k,4}} \in [0, 1]$.

#### 5) Terms Used in Traces of Professional Activities Descriptions ($TM_5$):

| |
|---|
| If a JO explicitly describes an interaction with a professional concept, a term representing that concept is more likely to be considered relevant. |


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

In a JO, a trace of a professional activity is a sentence that describes an action performed by a worker. Be $b_j \in d_i$ a trace of a professional activity description described by the set of terms $T_{b_j}$. We request that $b_j$ contains at least one verb and one dependent object. As a result, the terms $t_k$ that represent these objects will have a higher chance of being selected as relevant. Thus:

$$\forall t_k(t_k \in T_{b_j} \wedge \text{is\_object}(t_k, b_j)) \rightarrow t_k \in R_{d_i} \quad (6)$$

with a possibility degree $\alpha_{t_{k,5}} \in [0, 1]$.

*6) **Terms Representing High Risk Professional Skills/Activities** ($TM_6$):*

| Terms representing professional skills or activities where an employee's mistake can negatively impact a company's performance tend to be more relevant. |
|---|

An ontology $M$ describes the set of professional skills and activities of a given company. $M$ contains a set of concepts $c_M = \{c_{M,1}, c_{M,2}, \ldots, c_{M,k}\}$. Recruiters manually assign a risk level $\epsilon_{c_{M,k}} \in [0, 1]$ to professional skills and activities. Value 0 indicates that a potential error will not significantly affect the economic activity, while value 1 indicates significant effects.

Let $s_j$ be a term in a JO $d_i$ representing a professional skill or activity in $M$. As one of the concepts associated to $s_j$, let $c_{M,l}$ be the one with the highest risk level. When this risk level exceeds a threshold $\beta_{c_{M,l}}$, then $s_j$ is selected as a pertinent term and:

$$\forall s_j \exists c_{M,l}(s_j \in T_{d_i} \wedge c_{M,l} \in M \wedge s_j \in T_{c_{M,l}} \wedge \text{is\_greater\_than}(\epsilon_{c_{M,l}}, \beta_{c_{M,l}}) \rightarrow s_j \in R_{d_i} \quad (7)$$

with possibility degree $\alpha_{s_{j,6}} \in [0, 1]$.

*7) **Actions Expressed in Management JOs** ($TM_7$):*

| A relevant term is more likely to be one that represents work actions in management job offers. |
|---|

The recruiter can identify what type of actions management JOs are required to perform. A management job might focus on team management, while another may involve accountability activities or even development tasks.

Be $d_i$ a management JO. Based on 14,000 curriculum vitae, a Latent Dirichlet Allocation model was trained to detect management JOs. Let $t_k$ be a verbal term of $d_i$. If $t_k$ is part of the trace of a professional activity $f_j$ and corresponds to the head of its syntactic tree, then this term may be relevant. We define it as follows:

$$\forall t_k \exists f_j(f_j \in d_i \wedge t_k \in f_j \wedge \text{is\_management}(d_i) \wedge \text{is\_verb}(t_k) \wedge \text{is\_head\_of}(t_k, f_j)) \rightarrow t_k \in R_{d_i} \quad (8)$$

with a possibility degree $\alpha_{t_{k,7}} \in [0, 1]$.

*8) **BERT Semantic Similarity of Professional Skills** ($TM_8$):*

| If a *specific term* that represents a professional skill is semantically close (in the sense of BERT) to already discovered relevant terms, then it will be considered relevant. |
|---|

Let $t_1 \in R_{d_i}$ and $t_2 \in T_{d_i}$. Let $f(t)$ be the specificity function of a term $t$ defined as its relative frequency in a specific corpus $C_s$, divided by its frequency in a multi-language corpus $C_L$ [22].

Furthermore, we define $g(t_1, t_2)$ as the BERT semantic similarity between two terms. Using a SBERT model [27] pre-trained on the Wikipedia corpus, we have semantically analyzed complex terms. As a result, this model was fine-tuned based on the following professional skill standards: CIGREF, e-CF, C2I, and ROME. We defined it as follows:

$$\forall t_1 \forall t_2(t_1 \in R_{d_i} \wedge g(t_1, t_2) > 0) \rightarrow t_2 \in R_{d_i} \quad (9)$$

with a possibility degree defined by the normalized equation:

$$\alpha_{t_2,8} = \|(1 - \alpha_{t_1}) * g(t_1, t_2) * f(t_2))\| \quad (10)$$

*9) **Relevance of the Economic Activity Sector** ($TM_9$):*

| Potentially relevant terms refer to the economic activities required by the job posting (e.g., finance, banks, aeronautics, etc.). |
|---|

This implies that:

$$\forall t_k(t_k \in T_{d_i} \wedge \text{is\_sector\_requirement}(t_k)) \rightarrow t_k \in R_{d_i} \quad (11)$$

with a possibility degree $\alpha_{t_{k,9}} \in [0, 1]$. In order to identify economic activity sectors, we aligned job posting terms and economic activity concept labels, provided by ESCO, O*NET, ROME, and CIGREF standards.

*10) **Professional Skill Prerequisites** ($TM_{10}$):*

| Terms representing professional skills prerequisites tend to be more relevant. |
|---|

Assume there is a *prerequisite relation* between two professional skills $c_1$ and $c_2$ in an ontology $o_i$. Ontologies such as ESCO can be used to derive relations of this type. The possibility degree of $c_1$ will be inherited by $c_2$ if $c_2$ is a prerequisite of $c_1$ and $c_1$ is relevant (under a certain possibility degree).

$$\forall t_1 \forall t_2 \exists c_1 \exists c_2(c_1 \in o_i \wedge c_2 \in o_i \wedge t_1 \in T_{c_1} \wedge t_2 \in T_{c_2} \wedge \text{is\_prerequisite}(c_1, c_2) \wedge t_1 \in R_{d_i}) \rightarrow t_2 \in R_{d_i} \quad (12)$$

with a possibility degree $\alpha_{t_{k,10}} \in [0, 1]$ and $\alpha_{t_{k,10}}$ is equal to the possibility degree of $t_1 \in R_{d_i}$. As an example, a skill prerequisite relationship derived from the ESCO ontology can be the "Use of Functional Programming" ($c_2$) in order to master "Haskell" ($c_1$).


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

*11) **YAKE! Casing** ($TM_{11}$):*

Upper-cased terms tend to be more relevant.

This YAKE! marker is defined as:

$$\forall t_k(t_k \in T_{d_i} \wedge \text{is\_upper\_cased}(t_k)) \rightarrow t_k \in R_{d_i} \quad (13)$$

The normalized YAKE! equation is used to calculate the possibility degree as:

$$\alpha_{t_{k,11}}(t_k) = \left\| \frac{\max(\text{TF}(U(t_k), \text{TF}(A(t_k))}{\ln(\text{TF}(t_k))} \right\| \quad (14)$$

where $\text{TF}(U(t_k))$ is the number of times that $t_k$ appears uppercased, $\text{TF}(A(t_k))$ is the number of occurrences of $t_k$ as an acronym (for details see [14]) and $\text{TF}(t_k)$ is the term frequency.

*12) **YAKE! Term Position** ($TM_{12}$):*

Terms that appear at the beginning of the document tend to be more pertinent.

This marker is defined as:

$$\forall t_k(t_k \in T_{d_i} \wedge \text{is\_position\_marker\_activated}(t_k)) \rightarrow t_k \in R_{d_i} \quad (15)$$

with a possibility obtained from the following normalized YAKE! equation:

$$\alpha_{t_{12}}(t_k) = \| \ln(\ln(3 + \text{Median}(\text{Sent}(t_k)))) \| \quad (16)$$

$\text{Sent}(t_k)$ is the set of positions of the sentences containing $t_k$.

*13) **YAKE! Term Frequency Normalization** ($TM_{13}$):*

There is more relevance to the terms that are commonly used.

We define this marker as:

$$\forall t_k(t_k \in T_{d_i} \wedge \text{is\_frequency\_marker\_activated}(t_k)) \rightarrow t_k \in R_{d_i} \quad (17)$$

The possibility degree is calculated based on the following normalized equation proposed by YAKE!:

$$\alpha_{t_{k,13}}(t_k) = \left\| \frac{\text{TF}(t_k)}{\text{MeanTF} + \sigma} \right\| \quad (18)$$

where $\text{TF}(t_k)$ is the number of occurrences of $t_k$, which is balanced by the mean and standard deviation of frequency.

*14) **YAKE! Term Relatedness to Context** ($TM_{14}$):*

The more terms co-occur on both sides of a candidate term t, the less significant that term is.

Accordingly:

$$\forall t_k(t_k \in T_{d_i} \wedge \text{is\_relatednes\_activated}(t_k)) \rightarrow t_k \in R_{d_i} \quad (19)$$

with a possibility degree obtained from the normalized YAKE! equation:

$$\alpha_{t_{k,14}} = \left\| 1 + (DL + DR \cdots) * \frac{\text{TF}(t_k)}{\max \text{TF}} \right\| \quad (20)$$

where

$$DL[DR] = \frac{|A_{t,w}|}{\sum_{k \in A_{t,w}} CoOccur_{t,k}} \quad (21)$$

In a window of size w, $|A_{t,w}|$ corresponds to the number of different terms, and TF is the term frequency.

*15) **YAKE! Different Sentences** ($TM_{15}$):*

A term's relevance depends on how frequently it appears within different sentences.

Here, relevance is defined as:

$$\forall t_k(t_k \in T_{d_i} \wedge \text{is\_sentences\_marker\_activated}(t_k)) \rightarrow t_k \in R_{d_i} \quad (22)$$

with a possibility degree obtained from the normalized equation:

$$\alpha_{t_{k,15}} = \left\| \frac{SF(t_k)}{\#\text{Sentences}} \right\| \quad (23)$$

where $SF(t_k)$ is the number of sentences containing $t_k$ and $\#\text{Sentences}$ is the total number of sentences of $d_i$.

*16) **YAKE! Overall Score** ($TM_{16}$):*

Based on markers $TM_{11}, TM_{12}, TM_{13}, TM_{14}$ *and* $TM_{15}$ proposed by YAKE!, we include YAKE!'s global relevance score. Let $t_k \in d_i$. A term is considered as "possibly relevant" if it is predicted as such by the overall score:

$$\forall t_k(t_k \in T_{d_i} \wedge \text{is\_predicted\_by\_yake}(t_k)) \rightarrow t_k \in R_{d_i} \quad (24)$$

with a possibility degree $\alpha_{t_{k,16}} \in [0, 1]$.

## V. Evaluation of Textual Markers

Two factors should be considered regarding the recruiters' annotations of job offers. Firstly, it is a classification task, since it consists of determining whether or not each term of a JO is relevant to describe its essential content. Being a classification task, it can be understood as a rational action that an expert recruiter takes according to his/her knowledge [3]. Secondly, the act of annotating documents can be thought of as an inference process that recruiters undertake when reading the JO. Therefore, their annotations may be highly subject to cognitive uncertainties, which should be integrated


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

to natural language processing tasks [4]. In the following two sections, we present the two fuzzy-oriented models, applied to the evaluation of textual markers derived from recruiters' strategies.

### A. *Preliminary Definitions*

Let $U = t_1, t_2, ..., t_m$ be the set of terms of a JO, where $m$ represents the number of terms extracted. Each JO term $t_m$ can be described by a set of relevance textual markers ($TM_k$) derived from recruiters strategies and existing literature. We denote them as $I(k) = \{TM_1, TM_2, \ldots, TM_k\}$. Therefore, each term $t_m$ can be represented in the following form:

$$(x_{i0}, x_{i1}, ..., x_{ij}, \tilde{Y}_i), 1 \leq i \leq m \tag{25}$$

where $x_{ij}$ corresponds to a possibility degree obtained from textual marker $j$ for the term $i$ of being a relevant term. $\tilde{Y}_i$ represents the recruiter's annotation on this term which is inherently influenced by uncertainties (as such, we consider it an estimation $\tilde{Y}_i$ of the actual truth $Y_i$).

On the other hand, we define the fuzzy set $C$ that aims to model the relevance levels of the terms that the recruiters identify in the JOs. $C$ is composed of a membership function $\mu_C$ that allows to fuzzify the annotations made by the recruiters on the JOs. Furthermore, we define that the set $C$ is composed of two fuzzy subsets: $C_1$ that represents the relevance levels of the relevant terms and $C_2$ that represents the relevance levels of the non-relevant terms. These functions have been modeled using triangular functions whose support covers the range (0,1). In addition, we define the fuzzy set $R$ (resp. $R_1, R_2$), contained in $C$ (resp. $C_1$, $C_2$), and obtained after fuzzifying the annotations made by the recruiters. In the following sections, we present how the linear—fuzzy logic logistic regression—and non-linear—fuzzy decision tree—, approaches were applied to assess the uncertainty of relevant textual markers.

### B. *Linear Evaluation: Fuzzy Logistic Regression*

Be $t = \{t_1, t_2, t_3, ..., t_m\}$ the set of terms of the JO. We assume that these terms can be represented as a linear combination of the set of textual markers $I(k)$. Applying the fuzzy logistic regression algorithm [16], let $\mu_i \in \{C_1(pertinent\ term), C_2(non\ pertinent\ term)\}$ be the recruiter's annotation on the ith term of a job posting. We estimate the parameter $\tilde{u_i}$ from the ratio $\frac{\tilde{\mu_i}}{1-\tilde{\mu_i}}$. In our context, $\frac{\tilde{\mu_i}}{1-\tilde{\mu_i}}$ can be interpreted as the possibility of a term of not being relevant in relation to the possibility of being relevant, or vice versa. Therefore, the model is [16]:

$$\tilde{W_i} = ln\frac{\tilde{u_i}}{1-\tilde{u_i}} = A_0 + A_1 x_{i1} + ... + A_n x_{in}, i = 1, ..., m \tag{26}$$

where $\tilde{W_i}$ is the estimated output that can be transformed back to $\tilde{u_i}$ by the extension principle and $A_i$=($a_i$, $s_i$) represents a triangular fuzzy and symmetrical number with center $a_i$ and spread $s_i$.

### C. *Non Linear Evaluation: Fuzzy Decision Trees*

In order to train the fuzzy decision tree, we fuzzify each textual marker by applying a membership function $\mu_{TM_k}$ built equivalently to $\mu_C$, but taking into account the specific codomain of each marker $TM_k$. We claim that this fuzzification represents an evidence $E_k$. From the fuzzification of each textual marker and recruiters' annotations, we estimate the possibility of representing the fuzzified recruiters' annotations $R$ in light of the evidence $E_k$. In particular, we evaluate how ambiguous the following implication is: If $E_k$ Then $R$. Multiple measures can be used to evaluate this implication [3]. We applied the subsethood measure to estimate how much the evidence $E_k$ implies the experts' classification $R$, according to:

$$S(E_k, R_i) = \frac{M(E_k, R_i)}{M(E_k)} = \frac{\sum_{t \in U} min(\mu_{E_k}(t), \mu_{R_i}(t))}{\sum_{t \in U} \mu_{E_k}(t)} \tag{27}$$

In relation to recruiters' strategies and viewpoints, we determine whether a term is relevant $R_1$ or not $R_2$ making use of:

$$\pi(R_i|E_k) = \frac{S(E_k, R_i)}{max(S(E_k, R_1), S(E_k, R_2))} \tag{28}$$

As possibility is intrinsically related to the concept of ambiguity [3], there is less ambiguity when we can clearly determine whether a term is relevant or not. From $\pi(R|E_k)$, we estimate the ambiguity level associated to marker $TM_k$ linked to the evidence $E_k$ as:

$$G(E_k) = g(\pi(R|E_k)) = \sum_{i=1}^{n} (\pi_i^* - \pi_{i+1}^*) ln(i) \tag{29}$$

where $\pi^* = \{\pi_1^*, \pi_2^*, \ldots, \pi_n^*\}$ is the possibility distribution $\pi(R|E_k)$ permuted and sorted so that $\pi_i^* \geq \pi_{i+1}^*$ for $i \in \{1, \ldots, n\}$ and $\pi_{n+1}^* = 0$.

Given that we evaluate ambiguity by considering whether a term is relevant ($R_1$) or not ($R_2$) based on $TM_k$, then $n = 2$. Subject to this ambiguity function, we can estimate the extent to which it can be clearly inferred that a term is pertinent or not, according to $I_k$. Therefore, $ln(n)$ indicates maximum ambiguity and 0 represents no ambiguity [3]. To train the fuzzy tree, our final step is to replace the classical information entropy measure with the previously presented ambiguity metric. In the case of complex evidences $E_k$ composed by subsets of evidence, the ambiguity is estimated using the partitioning approach [3].

## VI. Experimental Results

A test of our approach was conducted at DSI Group's recruitment department. In total, five recruiters participated in our experiment and we refer to them as A, B, C, D, and E. These recruiters had in-depth knowledge of the essential JOs' requirements they manipulated within the setting of this experimentation.


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

### A. *Procedure*

As indicated in section III.A, our experimentation began with the representation of the organizational context surrounding JOs, based on interviews with recruiters. From this procedure we derived the ontology illustrated in section III.B. Then we asked recruiter A, the director of the human resources department, to describe the most relevant requirements of five JO under his responsibility. In recruiting a candidate, relevant requirements are those that do not allow for any flexibility.

Using expert A's strategies for selecting essential information in each job opening we extracted relevance textual markers. Generally, the annotated terms related to professional skills, and to a lesser extent, to location and availability, among others. Once the textual markers were derived conforming to recruiter A findings, we invited the other four recruiters (B, C, D and E), to determine whether the strategy derived from recruiter A's behavior was valid or not. This evaluation process was executed as follows:

- Recruiters B, C, D, and E annotated JOs that they had managed. We obtained a total of 25 annotated documents. On average, each job posting contained 100 terms of interest, out of which between 4 to 10 terms were annotated as relevant. A first dataset of 2,501 terms was generated.
- To train the fuzzy models, a second dataset was generated using the random undersampling RUSBoost algorithm [28]. A dataset of 500 terms was obtained, out of which 35% were relevant and 65% irrelevant.
- Both the linear and non-linear fuzzy models were trained on 70% of the second dataset and tested on the remaining 30%. We used stratified sampling to guarantee the proportion of relevant and non relevant terms on each dataset. Additionally, we examined the reliability of the resulting models by using a stratified 10-fold cross-validation.
- Both fuzzy models were compared to a state-of-the-art term extraction approach. For each annotated JO, we assessed the suitability of each model, based on the precision@K, recall@K, and F1-score@K metrics (where N represents the number of terms annotated by the recruiter). Thus, we evaluate the top K predicted terms by each method that are relevant.
- Model evaluations were done with the remaining terms of the first dataset, after the terms of the second dataset used for training were excluded. The training procedure allowed to obtain the best model avoiding overfitting and guaranteeing a maximal variance of the training samples. Finally, the evaluation procedure for measuring the precision@K, recall@K, F1-Score@K metrics had as a goal to confront the trained models to a much more realistic setting with a significant amount of non relevant terms.

### B. *Example of an Annotated Job Offer*

Below (Example 1), we present a summary view of an example JO annotated (with relevant terms in bold) by recruiter B. Additionally, Table II shows an example of extracted terms from this JO using the terminological analysis.

TABLE II
EXAMPLE OF EXTRACTED TERMS FROM THE PREVIOUS JO (EXAMPLE 1) AFTER APPLYING THE TERMINOLOGICAL ANALYSIS.

| | JO Terminology | |
|---|---|---|
| **#** | **Term** | **Term Type** |
| 1 | BI | Simple |
| 2 | BO | Simple |
| 3 | Stambia ETL | Complex (Compound variant) |
| 4 | Knowledge of Stambia ETL | Complex (Compound variant) |
| 5 | SQL | Simple |
| 6 | Maîtrise du langage SQL | Complex (Compound variant) |

***Example 1.*** *JO annotated with relevant terms in bold:*
***BI / BO*** *Analyst M/W*
*Company Description...(it contains 121 words)*
*Job description... (it contains 89 words)*
*Profile Description... (it contains 69 words)*
*You hold a Computer Engineering degree. You have technical skills such as:*
*- Business Objects platform -* ***Mastery of the SQL language****, and the use of databases (****SAP IQ / IBM DB2****)*
*Knowledge of Stambia ETL or Oracle.*
*Data Integration would be appreciated*
*Good interpersonal skills, dynamism, spirit of synthesis, proactive,*
*and team spirit are qualities that characterize you.*
*Job experience: Minimum 2 years. Position location: Metz-57. Geolocatable: Yes.*

Table III presents the top N=5 terms predicted by the fuzzy logistic regression and decision tree models on the example JO, as well as the relevance scores of each term, with the associated intervals and ambiguity levels. Some predicted terms (like DSI and Enterprise Activity) are part of the company/job description sections. In this case, both syntactically and semantically, the decision tree model predicts closely terms that are annotated by recruiters.

### C. *Experimentation*

Table IV presents the results of our experiments. All tests were done applying fuzzy logistic regression (FLR) and fuzzy decision tree (FDT) approaches. We trained each model using state-of-the-art textual markers [E], the proposed context-driven textual markers [R], and combining the two textual markers extraction procedures [R+E]. As indicated by the metrics, the fuzzy decision tree results are significantly better than the fuzzy logistic regression and the YAKE! algorithm. We also evaluated the algorithms proposed by [13] [15], which under-performed YAKE!. The fuzzy decision tree improved the best results of the state-of-the-art approach from 27% to 53%, being 78% for Recall@2N the highest performance. Note that the state-of-the-art textual markers were adapted to the specific context of JOs through the training process.


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

TABLE III
TOP N=5 TERMS PREDICTED BY THE FUZZY LOGISTIC REGRESSION AND DECISION TREE.

| | Fuzzy Logistic Regression | | | Fuzzy Decision Tree | | |
|---|---|---|---|---|---|---|
| **#** | **Term** | **Score** | **Interval** | **Term** | **Ambiguity %** | **Relevance Score** |
| 1 | DSI | 0.98 | $\pm 0.02$ | BI | 9 | 0.97 |
| 2 | Mastery of the SQL Language | 0.93 | $\pm 0.09$ | BO | 9 | 0.97 |
| 3 | Enterprise Activity | 0.91 | $\pm 0.15$ | Mastery of the SQL Language | 16 | 0.87 |
| 4 | BI | 0.87 | $\pm 0.16$ | SAP IQ | 28 | 0.71 |
| 5 | SAP IQ | 0.87 | $\pm 0.16$ | Technical Skill | 25 | 0.69 |

TABLE IV
PRECISION, RECALL, AND F1-SCORE RESULTS OF EACH METHOD TESTED ON 25 JOS. FLR: FUZZY LOGISTIC REGRESSION; FDT: FUZZY DECISION TREE; [E]: STATE-OF-THE-ART TEXTUAL MARKERS; [R]: PROPOSED CONTEXT-DRIVEN TEXTUAL MARKERS; [R+E]: COMBINATION OF STATE-OF-THE-ART AND PROPOSED CONTEXT-DRIVEN TEXTUAL MARKERS.

| **Metric/Model** | **YAKE!** | **FLR[E]** | **FDT[E]** | **FLR[R]** | **FDT[R]** | **FLR[R+E]** | **FDT[R+E]** |
|---|---|---|---|---|---|---|---|
| Precision@N, Recall@N and F1-Score@N [a] | 0.10 | 0.16 | 0.19 | 0.24 | 0.38 | 0.41 | **0.53** |
| Recall@2N | 0.25 | 0.33 | 0.40 | 0.42 | 0.57 | 0.62 | **0.78** |
| Precision@2N | 0.12 | 0.16 | 0.20 | 0.21 | 0.28 | 0.31 | **0.39** |
| F1-Score@2N | 0.16 | 0.22 | 0.27 | 0.28 | 0.37 | 0.41 | **0.52** |

[a]Recall@N, Precision@N and F1-Score@N are equivalent at N.

Table V presents the coefficient values for each of the textual markers, based on the obtained models. A classical logistic regression was also trained, to include a complementary well-known model. Assessment of the textual markers' ambiguity applying the fuzzy decision tree reveals interesting aspects of how relevant terms are identified. For instance, low ambiguity appears for indicators $TM_1$, $TM_{12}$, and $TM_{16}$, indicating that: recruiters tend to take into account relevant terms in job titles (according to $TM_1$); terms appearing at the beginning of the document tend to be relatively relevant (in agreement with $TM_{12}$'s), which could be due to the company description section appearing at the beginning in some JOs; because of YAKE! features, often highly irrelevant terms are predicted as relevant (as reported by $TM_{16}$), being an estimation of counter-relevance of terms in our context.

## VII. DISCUSSION

Uncertainty estimation is crucial to improve the identification of relevant terms extracted automatically from JOs. Our work proposes an analysis of possibility and uncertainty metrics, to assess the relevance of identified textual markers.

The classical logistic regression has a $R^2$ value of 0.64, which indicates a relative strong fit. This value was used as a convenient but not decisive indicator (because of the data uncertainty), revealing to which degree the introduction of the context-driven markers helped to better describe the recruiters viewpoints about what is relevant in JOs, from a statistical point of view. Moreover, our hypothesis that a probabilistic model of the recruiters' annotations was not sufficiently appropriate, is likely to be confirmed by the $p$-values of the classic logistic regression. According to the coefficients of the fuzzy logistic regression, recruiter-oriented indicators, $TM_2$, $TM_3$, $TM_4$, $TM_5$, and $TM_8$ seem to be the most pertinent contextual markers.

We noticed that marker $TM_8$ (similarity of terms with important skills) induces relevant terms corresponding to false-positives, strongly related to the JO's context (e.g. the term "Technical Skill" predicted in section VI.B). Regarding the intercept value of the FLR by applying the extension principle [16], the *possibility* of predicting a term as highly relevant is centered on 8% if all its textual markers values are zero, which is a more pertinent assumption due the uncertainty of recruiters viewpoints. The intercept of the CLR model gives a *probability* centered on 1% instead, indicating that even if all the regressor variables are zero, there is a level of uncertainty still not described, associated with the recruiters' viewpoints of information relevance.

The applied fuzzy models appear to be better suited to handle considerable uncertain information [4] communicated by recruiters. According to the obtained results, the fuzzy decision tree shows a better performance, implying its feasible alignment with recruiters' strategies. This is supported by the fact that the fuzzy decision tree F1-Score was better using only the context-driven markers, the context-independent markers, and both types of markers combined. Specifically, we observed that multiple decision rules produced after training the fuzzy decision tree match previously behaviors observed in recruiters. The following rule is an example:

> If it is highly possible that a term in the title represents a professional skill or job type ($TM_1$) and if it is highly possible that it represents a professional skill mentioned in the job or profile description sections ($TM_2$), then it is highly possible that such a term is relevant.

We also observed that some domain-independent markers are correlated to the context of JOs. For instance, the $TM_{11}$ marker is associated with the behavior of recruiters who capitalize terms representing professional skills, which are generally relevant to JOs. Despite its importance, such a


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

TABLE V
INDIVIDUAL FUZZY-ORIENTED EVALUATION OF THE 16 EXTRACTED TEXTUAL MARKERS APPLYING CLASSIC LOGISTIC REGRESSION (CLR), FUZZY LOGISTIC REGRESSION (FLR), AND FUZZY DECISION TREE (FDT). COEF.: CLR COEFFICIENTS, SE: CLR STANDARD ERRORS, COEF. A: CENTER OF THE TRIANGULAR FUZZY NUMBER, COEF. S: SPREAD OF THE TRIANGULAR FUZZY NUMBER.

| | **CLR** | | | **FLR** | | **FDT** |
|---|---|---|---|---|---|---|
| Textual Marker | Coef. | SE | *p*-value | Coef. A | Coef. S | Ambiguity % |
| $TM_1$ | 1.18 | 0.67 | 0.078 | 0.33 | <0.001 | 12 |
| $TM_2$ | 4.02 | 0.52 | $< 0.001$ | 3.40 | <0.001 | 40 |
| $TM_3$ | 2.66 | 0.81 | $< 0.001$ | 1.23 | <0.001 | 26 |
| $TM_4$ | 1.66 | 0.52 | 0.002 | 1.00 | <0.001 | 17 |
| $TM_5$ | 2.30 | 0.56 | $< 0.001$ | 1.61 | <0.001 | 18 |
| $TM_6$ | 1.48 | 0.65 | 0.023 | 0.03 | <0.001 | 9 |
| $TM_7$ | -0.41 | 0.63 | 0.512 | 0.63 | <0.001 | 8 |
| $TM_8$ | 1.81 | 0.53 | $< 0.001$ | 1.08 | <0.001 | 13 |
| $TM_9$ | -0.30 | 0.66 | 0.647 | 0.71 | <0.001 | 8 |
| $TM_{10}$ | 1.02 | 0.68 | 0.132 | 0.26 | <0.001 | 8 |
| $TM_{11}$ | 1.09 | 0.45 | 0.015 | 0.81 | <0.001 | 39 |
| $TM_{12}$ | -0.56 | 0.26 | 0.029 | -0.85 | <0.001 | 19 |
| $TM_{13}$ | -0.27 | 0.63 | -0.436 | 0.68 | <0.001 | 31 |
| $TM_{14}$ | 0.12 | 0.10 | 0.246 | -0.02 | <0.001 | 20 |
| $TM_{15}$ | 3.87 | 2.73 | 0.160 | 1.71 | <0.001 | 35 |
| $TM_{16}$ | 1.86 | 0.91 | 0.041 | 0.41 | <0.001 | 5 |
| Intercept | -4.51 | 0.86 | $< 0.001$ | -2.48 | 0.730 | |

marker could also be ambiguous (39%), which is consistent because capitalization does not necessarily imply importance. Globally, our results indicate that the most pertinent textual markers are $TM_2$, $TM_3$, $TM_4$, $TM_5$, $TM_8$, $TM_{11}$ and $TM_{12}$.

## VIII. CONCLUSIONS AND PERSPECTIVES

In this study, we evaluated two fuzzy models—linear and non-linear—for assessing the uncertainty of textual markers in terms of ambiguity, with respect to recruiters' knowledge. These textual markers serve to extract automatically relevant terms that are appropriate to model the information in JOs. It is therefore likely that reliable textual markers can be identified according to ambiguity. Possibility intervals and ambiguity scores provide flexibility to the evaluation process centered on uncertain information provided by experts, within a specific organizational context, with the potential of being adapted to other JOs' organizational contexts. In general, textual markers derived from recruiters' strategies were more pertinent than those extracted from the literature, although results improved significantly when both were combined.

These results provide further support to the suggestion that machine learning systems should systematically include an organizational context layer representation, which in our case certainly improved the evaluation of textual markers. The scope of this study was mainly limited in terms of the corpus size and the modeled aspects of the organizational context. Further research is therefore still required. It will be necessary to examine a larger corpus in order to determine whether the selected textual markers can be applied to different organizational contexts. Additionally, a question remains about the suitability of uncertainty measures to particularities of different organizations and the impact of organizational changes in the evaluation of textual relevance markers.

## REFERENCES


[1] L. A. Cabrera-Diego, M. El-Béze, J. M. Torres-Moreno, and B. Durette, "Ranking résumés automatically using only résumés: A method free of job offers," *Expert Systems with Applications*, vol. 123, pp. 91–107, jun 2019.

[2] J. Martinez-Gil, A. L. Paoletti, and M. Pichler, "A Novel Approach for Learning How to Automatically Match Job Offers and Candidate Profiles," *Information Systems Frontiers*, vol. 22, no. 6, pp. 1265–1274, dec 2020.

[3] Y. Yuan and M. J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets and Systems*, vol. 69, no. 2, pp. 125–139, 1995.

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

[5] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, dec 2021.

[6] A. Espinal, Y. Haralambous, D. Bedart, and J. Puentes, "An ontology-based possibilistic framework for extracting relevant terms from job advertisements," in *Proceedings of the 14th International Joint Conference on Computational Intelligence - FCTA*, INSTICC. SciTePress, 2022, pp. 163–174.

[7] A. Espinal, Y. Haralambous, D. Bedart, and J. Puentes, "Uncertainty-oriented textual marker selection for extracting relevant terms from job offers," in *Proceedings of the 8th International Conference on Artificial Intelligence and Fuzzy Logic System - AIFZ*, vol. 12, no. 16. AIRCC Publishing Corporation, 2022, pp. 1–16.

[8] P. K. Roy, S. S. Chowdhary, and R. Bhatia, "A Machine Learning approach for automation of Resume Recommendation system," *Procedia Computer Science*, vol. 167, pp. 2318–2327, 2020.

[9] 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.

[10] C. Zhu, H. Zhu, F. Xie, P. Ding, H. Xiong, C. Ma, and P. Li, "Person-Job Fit: Adapting the Right Talent for the Right Job with Joint Representation Learning," *ACM Transactions on Management Information Systems*, vol. 9, pp. 1–17, 2018. [Online]. Available: https://doi.org/10.1145/3234465

[11] X. Wang, Z. Jiang, and L. Peng, "A Deep-Learning-Inspired Person-Job Matching Model Based on Sentence Vectors and Subject-Term Graphs," *Complexity*, vol. 2021, pp. 1–11, 2021.

[12] S. Rose, D. Engel, N. Cramer, and W. Cowley, *Automatic Keyword Extraction from Individual Documents*. John Wiley



 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

and Sons, 2010, ch. 1, pp. 1–20. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470689646.ch1

[13] A. Zehtab-Salmasi, M.-R. Feizi-Derakhshi, and M.-A. Balafar, "FRAKE: Fusional Real-time Automatic Keyword Extraction," 2021, arXiv 2104.04830.

[14] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, "YAKE! Collection-Independent Automatic Keyword Extractor," in *Advances in Information Retrieval*. Springer, 2018, pp. 806–810.

[15] R. Dagli, A. M. Shaikh, H. Mahdi, and S. Nanivadekar, "Job Descriptions Keyword Extraction using Attention based Deep Learning Models with BERT," in *3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, jun 2021, pp. 1–6.

[16] S. Pourahmad, S. M. T. Ayatollahi, S. M. Taheri, and Z. H. Agahi, "Fuzzy logistic regression based on the least squares approach with application in clinical studies," *Computers and Mathematics with Applications*, vol. 62, no. 9, pp. 3353–3365, nov 2011.

[17] 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.

[18] J. A. Breaugh, "Employee Recruitment," *Annual Review of Psychology*, vol. 64, pp. 389–416, 2013.

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

[20] M. Somodevilla García, D. Vilariño Ayala, I. Pineda, M. Somodevilla García, D. Vilariño Ayala, and I. Pineda, "An Overview of Ontology Learning Tasks," *Computación y Sistemas*, vol. 22, no. 1, pp. 137–146, 2018.

[21] S. Neutel and M. de Boer, "Towards automatic ontology alignment using bert," in *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, 2021.

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

[23] S. Mc Gurk, C. Abela, and J. Debattista, "Towards Ontology Quality Assessment," 2017, http://ceur-ws.org/Vol-1824/ldq_paper_2.pdf.

[24] V. Novák, "Fuzzy logic in natural language processing," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017, pp. 1–6.

[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] 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.

[27] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019, pp. 3982–3992.

[28] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: Improving classification performance when training data is skewed," in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.


 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731