Uncertainty-Oriented Textual Marker Selection for Extracting Relevant Terms from Job Offers

Albeiro Espinal, Yannis Haralambous, Dominique Bedart, John Puentes

Uncertainty-Oriented Textual Marker Selection for Extracting Relevant Terms from Job Offers

Albeiro Espinal[12] and Yannis Haralambous[1] and 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

Automated resume ranking aims at selecting and sorting pertinent resumes, among those sent to answer a given job offer. Most of the screening and elimination process relies on the resumes' content, marginally including information of the job offer. In this sense, currently available resume ranking approaches lack of accuracy in detecting relevant information in job offers, which is imperative to assure that selected resumes are pertinent. To improve the extraction of relevant terms that represent significant information in job offers, we study the uncertainty-oriented selection of 16 textual markers – 10 obtained by examining the behaviour of expert recruiters and 6 from the literature – according to 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 a state-of-the-art term extraction approach.

Keywords

Recruiter's Behavior Modeling, Relevant Term Extraction, Textual Relevance Marker Evaluation, Uncertainty Measure, Fuzzy Machine Learning.

1. 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 the candidates expressed in their CVs can be very complex. This is specially the case when recruiters receive dozens or hundreds 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. In part, this is due to three issues that have rarely been examined in the literature. First, the most relevant information in the JO is not optimally identified, generating irrelevant rankings with respect to the essential requirements [2]. Secondly, under-representation of the changing organizational context surrounding JOs tend 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.

An organization's context to define a set of relevant textual markers based on recruiters' strategies to select significant JOs' information, and estimated the consistency of those markers was already studied [6]. Nevertheless a question remains concerning the quantitative evaluation

David C. Wyld et al. (Eds): AIFZ, SIGPRO, NWCOM, CLSB, MLNLP, DTMN, BDIOT, CSITY - 2022
pp. 01-16, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121601

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. This article is organized as follows. Section 2 describes the related state of the art. We summarize some key aspects of our previous work in Sections 3 and 4. Section 5 describes the proposed uncertainty evaluation of textual markers. Experimental results are presented in Section 6. Discussion, conclusions and perspectives are presented in Sections 7 and 8.

2. State of the Art

CV ranking systems carry out three processing stages: CVs and JO pre-processing, representation, and automatic ranking of CVs in relation to the content of the JO. Those documents are pre-processed by extracting text from digital files (.pdf, .doc, .txt, among others). Then extracted texts can be standardized by eliminating noisy symbols, segmenting the documents, and making semantic annotations [1], as well as deleting stopwords [7]. Pre-processed documents can be represented based on n-gram models [1], bag-of-words [1], ontologies [8] and/or word embeddings [9]. From these representations, different approaches can be used to determine the most suitable CVs regarding a JO. They can rely on recruiters' feedback [1], neural architectures [9] and/or transformer models [10].

These methods, however, do not focus on extracting relevant information from the JO before ranking resumes. Some methods have proposed statistical and graph-based textual relevance markers for identifying significant terms in single documents [11] [12] [13].

Furthermore, uncertainty, a key concern of natural language processing, as automatic extraction of relevant information from individual documents [4], concerns the lack of information about an event. Three of the most studied approaches to determine uncertainty have been probability models [5], along with possibility theory and fuzzy logic models [14]. 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 as the 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, because of their possibilistic foundations at the crossroad of fuzzy sets and probability provide a simple and convenient setting for handling subjective tasks, as the automatic identification of the most relevant terms in JOs. Moreover, these types of models can be trained on small datasets to evaluate features relevance.

3. Representation of Job Offers

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

3.1. Organizational Context

The representation of societal contexts in machine learning models should be improved, allowing those models to become more adaptable to dynamic changes in organizations [15]. This is a critical aspect in our work, given that context influences strongly recruiters behaviors

[16]. 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, as specified in [17], by conducting an open dialogue with recruiters, specifying the entities and relationships that impact the JOs' life-cycle. As a result, the main entities, actors, processes, objectives, and organizational problems associated with JO management were identified. A pre-conceptual scheme was derived from this procedure and used for the construction of a mother-ontology, schematically described in the next section.

3.2. Ontology Derivation

We define a mother-ontology as a large ontology of module specifications. A mother-ontology was used to represent the main concepts and relationships inherent to the recruiters' context and JOs. Additionally, existent ontologies related to the 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 [18]. The integration of these ontologies was achieved using a hybrid approach based on Bidirectional Encoder Representations from Transformers (BERT) [19], an analysis of terminological variation [20], and measures of ontology quality [23].

In this compound ontology, we specified also the structure of JOs in terms of concepts as sections, paragraphs, sentences, syntagms, terms, words, etc. Additionally, synonyms, meronyms and hyponyms were used to describe relationships 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 [25]. An upper-view of the ontology is presented in Figure 1.

3.3. Analysis of Recruiters Viewpoints

Based on the organizational context representation using the previous ontology, we analyzed recruiters' strategies related to the selection of the most essential information in JOs. During the annotation process they highlighted the most relevant terms. To represent the description of each recruiter's observed actions, the controlled language proposed by [17] was used. It allows to represent actions sequentially, as triplets of the form <subject, verb, predicate>.

We categorized those actions as explicit (eg, <recruiter, selects, a term>) or implicit (eg, <recruiter, avoids, a term> or <recruiter, avoids, a JO section>). Once the annotations were described in a controlled manner, the Apriori algorithm 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 3.2. Obtained rules represent textual relevance markers of information in JOs.

[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

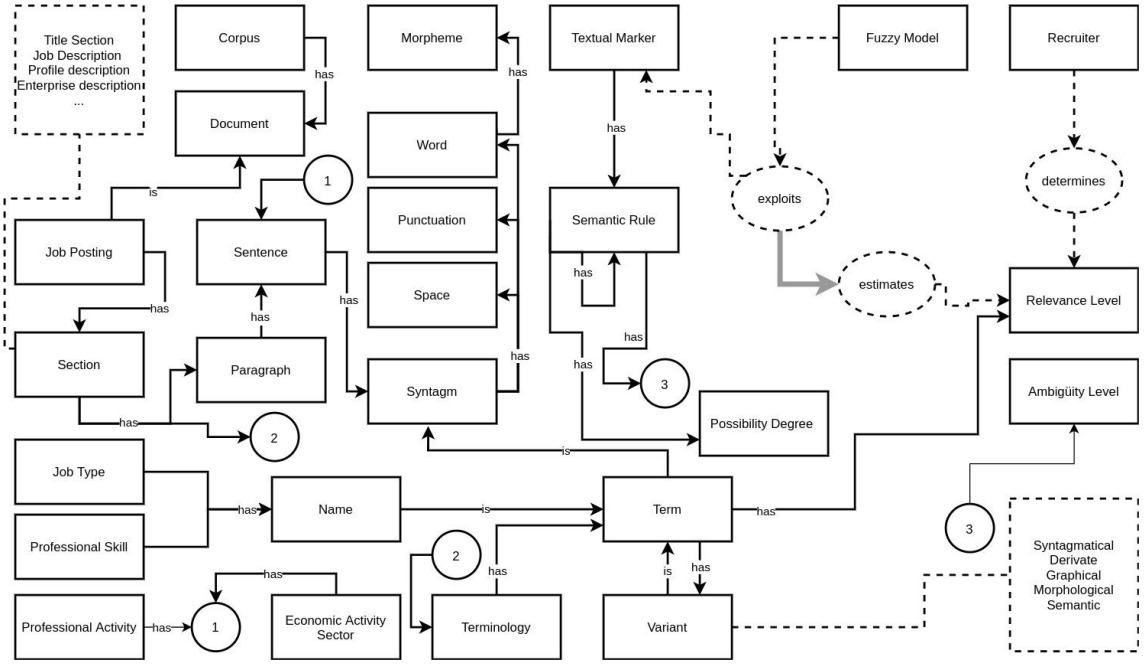

Figure 1. Upper-view of the mother-ontology created from the representation of the organisational context according to the principles of [17].

4. TEXTUAL MARKERS

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

4.1. Initial Elements

Considering that terms are defined as functional classes of lexical units used in discourse [20], JOs' relevant terms were identified by the weirdness ratio that measures their termhood [20, 22]. Additionally, we extracted the JOs' terms by using the most frequent morphosyntactic rules of the language, identified on multiple corpora analysis [20], which are mostly nominal phrases. A JO is then represented by its terminology, and our approach aims to identify the most optimal set of textual markers.

Let d_i be a JO belonging to a corpus C and $T_{d_i} = \{t_1, t_2, ..., t_n\}$ 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 $A_{d_i} = \{a_1, a_2, ..., a_m\}$ 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 $E_{d_i} = \{e_1, e_2, ..., e_p\}$ be a set of qualifying adjectives and nouns that are linked to a subset of terms in T_{d_i} by syntax dependencies.

Let $O = \{o_1, o_2, ..., o_s\}$ be a set of ontologies (as the one presented in Section 3). Let $c_{o_s} = \{c_{s,1}, c_{s,2}, ..., c_{s,k}\}$ be the set of concepts of ontology o_s and $T_{c_j} = \{t_{j,1}, t_{j,2}, ..., t_{j,l}\}$ the set of terms lexically representing concept c_j in a given language.

4.2. 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's 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 [12], 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.

4.2.1. Title Sections (TM_1)

"Any term in the title that resembles 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 $t_{a1} = \{t_1, t_2, ..., t_u\}$ 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:

$$\forall t_k \exists c_j [c_j \in o_s \wedge t_k \in T_{c_j} \wedge t_k \in t_{a1}] \rightarrow t_k \in R_{d_i}, \quad (1)$$

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

4.2.2. 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:

$$\forall t_k \exists c_j ((t_k \in s_2 \vee t_k \in s_3) \wedge t_k \in T_{c_j}) \rightarrow t_k \in R_{d_i}, \quad (2)$$

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

4.2.3. 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:

$$\forall t_m \forall t_n (t_m \in T_{d_i} \wedge t_n \notin S) \rightarrow (P(t_m \in R_{d_i}) > P(t_n \in R_{d_i})), \quad (3)$$

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

4.2.4. Terms Dependent on Pertinence Expressions (TM_4)

"A relevant term is more likely to be one that bears a syntax dependency with a JO's syntagm."

- Let $t_k \in T_{d_i} \cap T_{c_j}$ for some c_j.
- We define a "pertinent expression" e_m as a syntagm that the recruiter employed in the JO (i.e., *excellent* C# skills, *good understanding* of Kubernets). 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:

$$\forall t_k \exists e_m (t_k \in T_{d_i} \wedge e_m \in E_{d_i} \wedge is\_dependent(t_k, e_m)) \rightarrow t_k \in R_{d_i}, \quad (4)$$

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

4.2.5. 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."

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:

$$\forall t_k (t_k \in T_{b_j} \wedge is\_object(t_k, b_j)) \rightarrow t_k \in R_{d_i}, \quad (5)$$

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

4.2.6. Terms Representing High Risk Professional Skills/Activities (TM_6)

In this marker, we aim to provide more relevance to terms that represent professional skills or activities on which an employee's mistake can adversely affect the company's economic performance. Value 0 indicates that a potential error will not significantly affect the economic activity, while value 1 indicates significant effects.

An ontology M describes the set of professional skills and activities of a given company. M contains a set of concepts $c_M = \{c_{M,1}, c_{M,2}, ..., c_{M,k}\}$. Recruiters manually assign a risk level $\epsilon_{c_{M,k}} \in [0, 1]$ to professional skills and activities.

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:

$$\forall s_j \exists c_{M,l} (s_j \in T_{d_i} \wedge c_{M,l} \in M \wedge s_j \in T_{c_{M,l}} \wedge is\_greater\_than(\epsilon_{c_{M,l}}, \beta_{c_{M,l}}) \rightarrow s_j \in R_{d_i}, \quad (6)$$

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

4.2.7. Actions Expressed in Management JOs (TM_7)

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:

$$\forall t_k \exists f_j (f_j \in d_i \wedge t_k \in f_j \wedge is\_management(d_i) \wedge is\_verb(t_k) \wedge is\_head\_of(t_k, f_j)) \rightarrow t_k \in R_{d_i}, \quad (7)$$

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

4.2.8. 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 [20].

Furthermore, we define $g(t_1, t_2)$ as the BERT semantic similarity between two terms. Using a SBERT [24] model pre-trained on Wikipedia corpus, complex terms were semantically analyzed. 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:

$$\forall t_1 \forall t_2 (t_1 \in R_{d_i} \wedge g(t_1, t_2) > 0) \rightarrow t_2 \in R_{d_i}, \quad (8)$$

with a possibility degree defined by the normalized equation:

$$\alpha_{t_2,8} = \|(1 - \alpha_{t_1}) * g(t_1, t_2) * f(t_2))\|. \quad (9)$$

4.2.9. 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:

$$\forall t_k (t_k \in T_{d_i} \wedge is\_sector\_requirement(t_k)) \rightarrow t_k \in R_{d_i}, \quad (10)$$

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.

4.2.10. Professional Skill Prerequisites (TM_{10})

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).

$$\forall t_1 \forall t_2 \exists c_1 \exists c_2 (c_1 \in o_i \wedge c_2 \in o_i \wedge t_1 \in T_{c_1} \wedge t_2 \in T_{c_2} \wedge is\_prerequisite(c_1, c_2) \wedge t_1 \in R_{d_i}) \cdot \quad (11)$$

with a possibility degree $\alpha_{t_{k,10}} \in [0,1]$ and $\alpha_{t_{k,10}}$ is equal to the possibility degree of $t_1 \in R_{d_i}$.

4.2.11. YAKE! Casing (TM_{11})

There is a tendency for upper-case terms to be more relevant. This YAKE! maker is defined as:

$$\forall t_k (t_k \in T_{d_i} \wedge is\_upper\_cased(t_k)) \rightarrow t_k \in R_{d_i} \quad (12)$$

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

$$\alpha_{t_{k,11}}(t_k) = \| \frac{max(TF(U(t_k)), TF(A(t_k)))}{ln(TF(t_k))} \|, \quad (13)$$

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

4.2.12. YAKE! Term Position (TM_{12})

In this marker, the hypothesis is that terms that appear at the beginning of the document tend to be more pertinent.

$$\forall t_k (t_k \in T_{d_i} \wedge is\_position\_marker\_activated(t_k)) \rightarrow t_k \in R_{d_i}, \quad (14)$$

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

$$\alpha_{t_{12}}(t_k) = \| ln(ln(3 + Median(Sent(t_k)))) \|, \quad (15)$$

$Sent(t_k)$ is the set of positions of the sentences containing t_k.

4.2.13. YAKE! Term Frequency Normalization (TM_{13})

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

$$\forall t_k (t_k \in T_{d_i} \wedge is\_frequency\_marker\_activated(t_k)) \rightarrow t_k \in R_{d_i}, \quad (16)$$

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

$$\alpha_{t_{k,13}}(t_k) = \| \frac{TF(t_k)}{MeanTF + \sigma} \|, \quad (17)$$

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

4.2.14. YAKE! Term Relatedness to Context (TM_{14})

This YAKE! marker is based on the following hypothesis: “The more terms co-occur on both sides of a candidate term t, the less significant that term is”:

$$\forall t_k (t_k \in T_{d_i} \wedge is\_relatednes\_activated(t_k)) \rightarrow t_k \in R_{d_i}, \quad (18)$$

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

$$\alpha_{t_{k,14}} = \| 1 + (DL + DR \cdots) * \frac{TF(t_k)}{maxTF} \|, \quad (19)$$

where

$$DL[DR] = \frac{|A_{t,w}|}{\sum_{k \in A_{t,w}} CoOccur_{t,k}} \quad (20)$$

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

4.2.15 YAKE! Different Sentences (TM_{15})

"A term's relevance depends on how frequently it appears within different sentences", defined as:

$$\forall t_k (t_k \in T_{d_i} \wedge is\_sentences\_marker\_activated(t_k)) \rightarrow t_k \in R_{d_i}, \quad (21)$$

with a possibility degree obtained from the normalized equation:

$$\alpha_{t_{k,15}} = \| \frac{SF(t_k)}{\#Sentences} \|, \quad (22)$$

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

4.2.16. YAKE! Overall Score (TM_{16})

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

$$\forall t_k (t_k \in T_{d_i} \wedge is\_predicted\_by\_yake(t_k)) \rightarrow t_k \in R_{d_i}, \quad (23)$$

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

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

5.1. 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 existent literature. We denote them as $I(k) = \{TM_1, TM_2, ..., TM_k\}$. Therefore, each term t_m can be represented in the following form:

$$(x_{i0}, x_{i1}, ..., x_{ij}, \widetilde{Y}_i), \ 1 \leq i \leq m \quad (24)$$

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 μ_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 which represents the relevance levels of the relevant terms and C_2 which 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.

5.2. Linear Evaluation: Fuzzy Logistic Regression

Be $t = \{t_1, t_2, t_3 \ldots t_m\}$ 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 [14], 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 [14]:

$$\tilde{W}_i = ln\frac{\tilde{u}_i}{1-\tilde{u}_i} = A_0 + A_1 x_{i1} + \ldots + A_n x_{in}, i = 1,\ldots, m \tag{25}$$

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.

5.3. Non Linear Evaluation : Fuzzy Decision Trees

In order to train the fuzzy decision tree, we fuzzify each textual marker by applying a membership function μ_{TM_k}, built equivalently to μ_C, but taking into account the specific codomain of each marker TM_k. We define 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:

$$S(E_k, R_i) = \frac{M(E_k R_i)}{M(E_k)} = \frac{\sum_{t \in U} min(\mu_{E_k}(t), \mu_{R_i}(t))}{\sum_{t \in U} \mu_{E_k}(t)} \tag{26}$$

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

$$\pi(R_i|E_k) = \frac{S(E_k,R_i)}{max(S(E_k,R_1),S(E_k,R_2))} \quad (27)$$

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:

$$G(E_k) = g(\pi(R|E_k)) = \sum_{i=1}^{n} (\pi_i^* - \pi_{i+1}^*) ln(i) \quad (28)$$

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

Due to the fact that we evaluate ambiguity by considering whether a term is relevant (R_1) or not (R_2) based on TM_k, $n = 2$. 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 three, 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 evidences, the ambiguity is estimated using the partitioning approach [3].

6. Experimental Results

A test of our approach was conducted at DSI Group's recruitment department. In total, 5 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.

6.1. Procedure

As indicated in section 3.1, 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 3.2. 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 the most essential information in each job opening we derived relevance textual markers from his strategies. Generally, the annotated terms relate to professional skills, and to a lesser extent, 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, to analyze other CVs. 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 2501 terms was generated.
- To train the fuzzy models, a second dataset was generated using the random undersampling RUSBoost algorithm [22]. A dataset of 500 terms, with 35% relevant and 65% non relevant terms was obtained.
- 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).
- 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.

6.2. Example of an Annotated Job Offer

Below, we present a summary view of an example JO annotated (with relevant terms in bold) by recruiter B.

***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 1. *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 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 1 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.

6.3. Experimentation

Table 2 presents the results of our experiments. All tests were done applying the 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 [11] [13], 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.

Table 2. 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[5]	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**

Table 3 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. Evaluation 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.

5 Recall@N, Precision@N and F1-Score@N are equivalent at N.

Table 3. Individual uncertainty 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	

7. DISCUSSION

Uncertainty evaluation 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 6.2). Regarding the intercept value of the FLR by applying the extension principle [14], 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. Instead the intercept of CLR model gives a *probability* centered on 1%, indicating that even if all the regressor variables are zero, there is a level of uncertainty still not described, associated to 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 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 obtained a better F1-Score using only the context-driven markers, the context-independent markers, and both types of markers combined. Specifically, we observed that multiple decision rules obtained 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 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 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}.

8. 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. Those 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, B. Durette, 'Ranking résumés automatically using only résumés: A method free of job offers', *Expert Systems with Applications* 123, 91–107, 2019.

[2] J. Martinez-Gil, A. L. Paoletti, M. Pichler, 'A Novel Approach for Learning How to Automatically Match Job Offers and Candidate Profiles', *Information Systems Frontiers* 22(6), 1265–1274, 2020.

[3] Y. Yuan, M. J. Shaw, 'Induction of fuzzy decision trees', *Fuzzy Sets and Systems*, 69(2), 125–139, 1995.

[4] E. Pavlick και T. Kwiatkowski, 'Inherent Disagreements in Human Textual Inferences', *Transactions of the Association for Computational Linguistics* 7, 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, S. Nahavandi, 'A review of uncertainty quantification in deep learning: Techniques, applications and challenges', *Information Fusion 76*, 243–297, 2021.

[6] A. Espinal, Y. Haralambous, D. Bedart, J. Puentes, 'An Ontology-Based Possibilistic Framework for Extracting Relevant Terms from Job Advertisements', In *2022 International Conference on Fuzzy Computation Theory and Applications (FCTA), Accepted for Publication*, 2022.

[7] P. K. Roy, S. S. Chowdhary, R. Bhatia, 'A Machine Learning approach for automation of Resume Recommendation System', *Procedia Computer Science* 167, 2318–2327, 2020.

[8] D. Çelik, 'Towards a semantic-based information extraction system for matching résumés to job openings', *Turkish Journal of Electrical Engineering and Computer Sciences* 24(1), 141–159, 2016.

[9] C. Zhu, H. Zhu, F. Xie, P. Ding, H. Xiong, C. Ma, P. Li, 'Person-Job Fit: Adapting the Right Talent for the Right Job with Joint Representation Learning', *ACM Transactions on Management Information Systems* 9, 1–17, 2018.

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

[11] A. Zehtab-Salmasi, M.-R. Feizi-Derakhshi, και M.-A. Balafar, 'FRAKE: Fusional Real-time Automatic Keyword Extraction'. 2021.

[12] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, και A. Jatowt, 'YAKE! Collection-Independent Automatic Keyword Extractor', In *Advances in Information Retrieval*, 806–810, 2018.

[13] R. Dagli, A. M. Shaikh, H. Mahdi, και 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)*. 1–6, 2021.

[14] S. Pourahmad, S. M. T. Ayatollahi, S. M. Taheri, Z. H. Agahi, 'Fuzzy logistic regression based on the least squares approach with application in clinical studies', *Computers and Mathematics with Applications*, 62(9), 3353–3365, 2011.

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

[16] J. A. Breaugh, 'Employee Recruitment', *Annual Review of Psychology* 64, 389–416, 2013.

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

[18] M. Somodevilla García, D. Vilariño Ayala, I. Pineda, M. Somodevilla García, D. Vilariño Ayala, I. Pineda, 'An Overview of Ontology Learning Tasks', *Computación y Sistemas* 22(1), 137–146, 2018.

[19] S. Neutel, M. de Boer, 'Towards Automatic Ontology Alignment using BERT', In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, 2021.

[20] D. Cram, B. Daille, 'Terminology extraction with term variant detection', In *Proceedings of ACL-2016 system demonstrations*, 13–18, 2016.

[21] K. T. Frantzi, S. Ananiadou, J. Tsujii, 'The C-value/NC-value Method of Automatic Recognition for Multi-word Terms', Research and Advanced Technology for Digital Libraries 1513, 585–604, 2002.

[22] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, A. Napolitano, 'RUSBoost: A Hybrid Approach to Alleviating Class Imbalance', *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(1), 185–197, 2010.

[23] S. Mc Gurk, C. Abela, J. Debattista, 'Towards Ontology Quality Assessment'. 2017. Http://ceur-ws.org/Vol-1824/ldq paper 2.pdf.

[24] N. Reimers, I. Gurevych, 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks', In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, 3982–3992.

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