SELECTING ANSWERS FROM CV TEXT

Nasr S.
Information technologies in automatized systems department,
The Belarusian State university of informatics and radioelectronics
Minsk, Republic of belarus
E-mail: sara.nasr@gmail.com

The problem of extracting answer(s) from text block is considered. The text represents a Curriculum Vitae (CV) divided into semantic blocks. The problem is to find in CV the corresponding block containing applicant data to be extracted and inserted in some predefined form (PF) for automatic processing. So, the required semantic block should be found and processed appropriately to deliver necessary information for PF. The method is outlined to solve this problem.

Introduction

One of the interesting applied problems is automatic CVs processing. The final goal is to select the best candidate for some position(s) from those who sent his/her e-mail. Typically, CV contains a block of private and professional data of the candidate including age, education, participating in real projects, availability of published papers or research activities, possessing modern programming languages and technologies and so on. The problem of selection of the best candidate requires to estimate the integral choice function usually represented in the next form [1]

$$CF = \sum w_i \cdot uf_i.$$

Here, w_i stands for the weight of criterion i, and uf_i represents the value of utility function of criterion i. One of the basic problems is connected to extracting data from CV to be used in evaluation of the functions uf_i. We shall restrict our consideration by the demonstration of a common approach to solving the above problem. Namely, we suppose that CV is previously divided into semantic blocks in a way that to get answer to some question like «what is your age?» or «when you were born?» is possible by means of some text processing procedure outlined later on in this report. Of cource, to divide text intosemantic blocks is some serious applied problem as well. One of its possible solution is based on computing the correlation co-efficients between keywords in CV text and define semantic blocks as those, containing the subset of keywords mostly correlated to each other. We, however, leave this question without detail explanation. So, the main idea of the report consists in preparing some list of questions to CV text needed for filling PF. Clearly, the specificity of this process is its extreme undefiniteness. The rest of this short report contains the details.

I. Demonstration of the approach

As was said earlier, we require to prepare a list of questions to get the desired information. As we consider here an example with applicant age definition, let the corresponding list to consists of the next questions:

Q1. What is your age?
Q2. How old are you?
Q3. What is your date of birth?
Q4. When were you born?
Q5. What is the date of your borning?
Q6. How much years do you have?

The question is asked if and only if the previous question failed to be answered. This means that CV text does not contain the text block identified as suitable for (answering) the question. In order to be more concrete, let us consider the following CV text block in the field of our interests: «I am 25 years old and was born 1n 1996». This block contains even extra information for our needs. Now (leaving explanation for next section) the first question Q1 remains unanswered. But Q2, and Q4 with Q6 are success. It is enough to find data delivered by Q2 what prevents asking Q4. Now let us explain the idea of the method.

II. Method definition

We have to explain the method of CV-text processing. The entire process is divided into four stages in general [2]. The first stage is to transform Word document or a pdf-file with CV to a plain text. This can be done with a Tika system which extracts a raw text and deletes unnecessary control information such as colors, fonts, and the like. The next step is to get keywords of the text. The idea is to consider practically all text words as keywords due to not big size of CV. It is required that each sentence in CV contains at minimum one keyword. Besides, this simplifies the algorithm as it not requires to define each keyword score and test that each sentence is covered at least by one keyword.The prepositions, conjunctions, pronouns, auxiliary and modal verbs (such as can, have, may etc.) are excluded. For an example above, the set of keywords contains the words years, old, born. The keywords are selected in such a way that the similar words are identified as the same. For instance, programming and programmer are considered as one keyword. By thus, each keyword labels one or more semantic text blocks. To realize the keywords selection we use the Dice metrics (measure) given by the formula

$$q = \frac{2|X \cap Y|}{|X| + |Y|}.$$

This formula estimates similarity of two sets X and Y, |X| denotes the number of elements in X; $X \cap Y$ denotes two sets intersection. From practical viewpoint, the value of q should be not less than 0.5 (however, the experiments are required since this co-efficient depends of the compared sets sizes quite essentially). In the above formula, X stands for the set of keywords of some semantic block, and Y represents a question keywords.Again, pay attention to the fact that CV is short text, so practically each word of CV may be regarded a keyword.

The next stage consists of building two directories. Each directory represents a collection of the pairs <key, value>. The first directory contains the pairs of items and each pair represents a record with key standing for a keyword and a value representing the set of sentences (text blocks) numbers labelled with the keyword key. The second dictionary consists of the pairs representing the sentences numbers and their texts. Now let us explain how to extract an answer to the question represented by a set of words (some of words are keywords and some – not (these latter words do not belong to CV text)). For each keyword k_i in the question the set of sentences numbers N_i =$n_{i1}, n_{i2}, ..., n_{iz}$ is defined. Then for all N_i the most frequently encountered number(s) nw is (are) defined. This number nw defines a sentence (block) to be considered as an answer. If there are more than one candidate to be an answer then all candidates are considered.In the example the only keyword «old» is common in question and text block in resume («I am 25 years old and was born 1n 1996»). This gives Dice estimation equals to 0.67. By this, the text block B with the required data is defined.

The next step consists in processing B to extract the required data. Of cource, the question requires some commonly adopted form of the answer as, for example, «I am 25 years old» or «I am twenty five years old».The main feature of these two sentences is availability of the word «old». To extract age of the applicant from the text block defined, the grammar parsing technique is used, associated with the text of the answer. The text of the question («How old are you») supposes the following possible answer structures:

A1. I am ? years.
A2. I am ? years old.
A3. I am ?.
A4. Me is ? years.
A5. Me is ? years old.
A6. Me is ?.
A7. am ? years.
A8. am ? years old
etc.

Here, «?» stands for the data we are looking for. So, each answer pattern should be checked. Again, the Dice metrics is applied to evaluate degree of closeness of the resume text of block textitB and Ai. To identify the the numeric value of the applicant age is a simple string processing task.

III. Conclusion

Let us summarize the general approach to extracting data from CV,

1. The sets of possible questions to define each data item in resume should be prepared in advance. This means, that some set of questions should rekate to age, some to education, some to professional skills *etc.*
2. Each question is associated with some set of keywords (the keywords are directly defined from the question).
3. The system finds the textblock (semantic block) in CV which is close to the question with respect to Dice metrics (obviously, there are other metrics, besides Dice metrics, as well. However, selection of the metrics is not crucial for the system work).
4. After a semantic block is found, the data should be extracted. This is done by means of the previously created number of answer patterns. Each pattern should be considered with respect to it closeness to the semantic block. By means of the Dice metrics we are in position to define the required pattern and extract data accordingly to this pattern.

The technique outlined in this report is used as an essential part of the CV-processing system. It admits availability of the text mistakes and inaccuracies as well.

One of the important issues concerns definition of the optimal set of features to be analized in CV. This question is also of our research interests.

IV. References

1. Saati, T.L.Decision making with the analytic hierarhy process / T. L. Saati// Int. J. Services Sciences, Vol. 1, No. 1, 2008 .
2. German, Yu. Information extraction method from resume / Yu. O. German, O. V. German, S. Nasr// Proceedings of BSTU Scientific journal. (Minsk, Belarus).2019, №1(218), p.p.64–69.