Automatic processing of CV: a linguistic approach

O.Nouali, S.Kirat, H.Meziani

Division of Theory and Engineering of the Information Processing Systems, CERIST.
Rue des 3 Frères Aissou Ben-Aknoun, Algiers, Algeria.

onouali@cerist.dz, sabahkirat@yahoo.fr, hadjer_meziani@yahoo.fr

ABSTRACT

This article describes an information extraction system to solve problems with which the companies and the private individuals are confronted and more particularly the professionals of recruitment and persons in charge for HR concerning the data capture and the selection of the applications. It uses a set of linguistic diagrams to extract relevant information and to organize them in templates. It makes it possible to absorb entering flows of applications resulting from various sources, to extract, standardize and qualify partly information useful for recruitment. The performances of the system depend closely on the quality of this linguistic knowledge. The acquisition of this knowledge needs a systematic search of the texts of the application field in order to accumulate the diagrams. This search is completed by a linguistic reflection, in order to release the textual regularities.

Keywords: *Analyze of applications, Linguistic approach, Statistic approach, Information extraction, Natural language, recruitment.*

1. INTRODUCTION

The data-processing phenomenon disrupts our companies since nearly 30 years. The revolution which the profileration of the data constitutes, and especially, the immediate availability of much information, varied and from great quality, brings up scientific, legal or ethical questions. Thus, the documentary explosion, observed since the end of 1960, leads to an quasi-exponential growth of the flow.

The large companies, which they are in active phase of recruitment or not, receive large annual volumes of applications (from 20.000 to 200.000 CV) distributed according to the web flow, email and mail paper.

The persons in charge for human resources (HR) must treat all these applications: at least to answer the whole of the candidates (public image of the company), to detect as soon as possible the profiles likely to answer expectations of the company, to contact the candidates, to meet them, etc.

Today, the responsible for recruitment and their assistants open the envelopes and e-mail, read the CV and applications for job, they capture the information in the candidates' database of the company, and answer the applicants: the reception of the applications is thus primarily manual.

This manual process is very expensive in time and money.

The consequences are in particular:

1. Too long delays: several weeks to send an acknowledgement of receipt.
2. Loss of profiles potentially interesting for the company: most of the time, HR keep trace only of candidates being able to correspond to a profile sought-after at a given moment.
3. Incomplétude of archived information : in charge ones of recruitment often have time to capture only one small part of the information sent by the candidate (contact for example), or not qualified information (global electronic CV, on which one will be able to make only free text research).

This growth of flow and its problems had led the actors of the information and communications sciences to develop new tools and techniques of research, treatment, diffusion of information and extraction of relevant textual information in a condensed form, it is here about the automation of the summarization activity. Those make it possible to the various users (those who seek information) to acquire relevant documents and to provide them information with strong added value [11].

The information extraction (IE) seems particularly attractive and useful. It was defined respectively by the program DARPA MUC (Message Understanding Conference, 92-98), like a linguistic discipline of engineering aiming to identify, gather and standardize relevant information for specific users or applications.

In this article we describe an information extraction system applied to the Curriculum Vitae. We use the techniques of information extraction, which result mainly from the natural language processing, to solve problems with which the companies and the private individuals are confronted. The CV are presented in the form of short texts and generally a whole of short sentences. The language used in this type of texts is more or less regular.

2. AUTOMATIC INFORMATION EXTRACTION

The goal of the information extraction is to identify and extract specific and well defined information from texts written in natural language. It requires a comprehension of the texts which is sufficient to answer the studied specific task. It was often focused on limited problems which aim at extracting information in the form of entities to fill automatically databases or well defined Templates which will be used as bases with the automatic generation of summaries, to make indexing or segmentation, to answer questions [2]. The idea to synthesize the information of a document in a database goes up at the beginning of the Fifties [6]. One of the first implementations of this idea was done years later for medical texts at the university of New York by Naomi Sagger [12]. Other projects have followed concerning for example the transformation of whole encyclopaedias in structured form.

An information extraction system consists in analyzing free text with the aim of extracting various types of specific information [3]. Such systems are conceived to analyze the totality of the text, but to extract only the parts of each document which contain relevant information. The relevance is determined by preset directives which must specify, with the most possible exactitude, which type of information the system must find. Although this task of extraction is limited, it remains complex, information can be distributed at various places of a document, expressed in various ways.

Two principal families of methods are currently used. The first is based on statistical analyses of textual data. The second is based on complex linguistic methods and dependent on the language by combining sometimes certain statistical methods to them.

2.1. Information Extraction based on textual statistics

There is a panel of applications in the field of IE being based on statistical methods [15], [14], [1], [10], [5].

The statistical methods rest on the idea that there is a relationship between the contents conveyed by a text and the words used in this text, that this report/ratio is a function of the frequency of use of the words, and that there is a relation between the capacity of a word to be selected like relevant term and its frequency of employment. It was the progressive idea since 1957 by Hans Peter Luhn [8]. This one carries then the words of average frequencies, by eliminating those of high and low frequency, retained, on the other hand, by other researchers.

This approach was taken again by researchers like G. Salton [13], A.Bookstuin, Don R. Swanson [16] for whom the progress in automatic indexing passes by the refinement of the statistical models. With rough calculations of frequencies, one substitutes those of relative frequency calculated on the vocabulary of a document compared to the vocabulary of a corpus of texts of the same field. If necessary one attributes value to the results according to the assumption that a very frequent word in a document whereas it is not very frequent in a corpus of reference, is discriminating.

2.2. Information extraction based on linguistic analysis methods

The linguistic processing systems for the information extraction rely on various levels, lexical, syntactic and semantic. There are two large currents in the linguistic analysis. The first carries out a deep grammatical analysis by determining complex bonds between the entities of the statement (verb, subject, complement...) The analysis is complex and generates many possible choices. The second current uses a linguistic analysis of surface ("shallow parsing"). The analysis seeks only functional dependences "reliable" between two entities (subject-verb, name-adjective, verb-complement...). Many systems of Information Extraction based on this linguistic approach were set up. They are applied to documents coming from very diverse sources (army, press, finance, literature, history, biology...) [17], [7].

We will describe, below, our system based on the use of hierarchical linguistic diagrams.

3. GLOBAL ARCHITECTURE OF THE SUGGESTED APPROACH

The system is conceived in order to feed a database from the curriculum vitae.

The system is made up mainly of two modules, an analyzer and an extractor. (Figure 1).

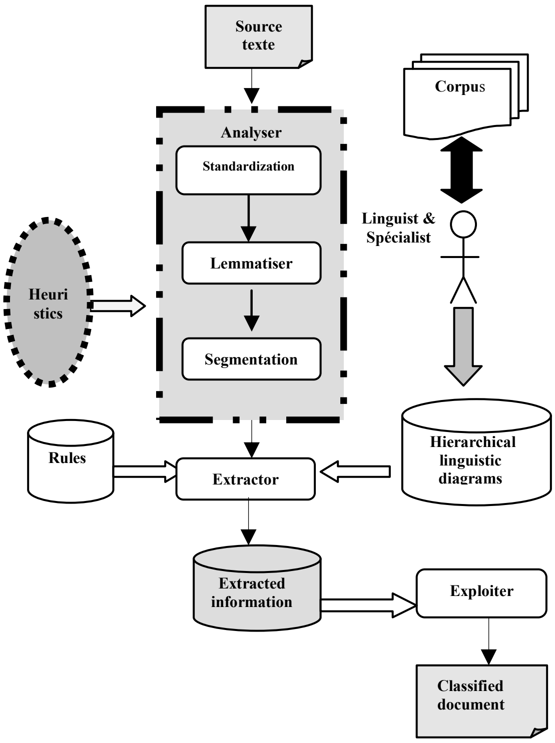

Figure 1: Global architecture of the approach.

3.1. ANALYZER

It permits to analyze a given text and delivers a first representation of the text.

The analyzer requires the following stages:

1. *Standardization :* is a necessary stage to the good development of the later phases (use of cuttings' rules of the text). This stage makes it possible to carry out a conversion of the source text, from a

given format to the format text (ASCII). It makes it possible to prepare the source text. Indeed, the texts are presented at the system under various formats (pdf, txt, word, htm...).

2. *Lemmatization* : this stage touches on the one of the great levels of linguistic processing which is the lexical level and that by making reduction of the word to its canonical form. This amounts to rewrite the standardized text in another text made up of its original units.
   The lemmatisation is an contents' analysis operation which operates by reduction of the words in an original entity (lemma), also called canonical form, which gathers the various variables of a word and its derivatives. This form is the infinitive for the verbs, the singular male form for the names, etc.
   This stage of lemmatisation is necessary; it permits to strongly decrease the number of linguistic diagrams by eliminating all the grammatical inflexions and derivations. This is carried out using a program called FLEMM [9]. It carries out the flexional morphological analysis of French texts labelled beforehand by BRILL [4].
   FLEMM calculates the lemma of each bent word (according to its label) and also provides its principal morphological features for example kind and number for the adjectives, number for the names, etc. The lemmatisor receives in input an unspecified text and provides in output a lemmatized text i.e. made up only of his canonical forms.
3. *Pre-Segmentation*it consists in applying a set of rules in order to cut out the text in processing units and to remove email addresses, URLs and abbreviations.
   The sentence is considered as the central unit of the natural language processing. One recognizes as phrase the series of the words situated between punctuation marks known as major such as the point, the exclamation mark, the question mark and of other which precede or follow its signs.
   The punctuation often constitutes a source of ambiguity for the cutting of the sentences and that causes problems as far as production of summaries. For example a point can be used to declare the end of a sentence but also to express an abbreviation or an acronym (ex: E.D.F or U.S.A), or even a decimal number, for example: 3.14 (writing anglosaxone).
   To mitigate this problem, the system uses heuristics permit to isolate the sentences, the addresses of electronic mails, the abbreviations and URLs.
4. *Segmentation:* We have studied the general organization of the CV. According to our analyses, the CV are organized according to a structure which contains various levels of information The whole of the textual units which cover the same subject forms a segment set of themes. For each topic, there is a limit which announces a change in the speech or the topic of the segment. To determine the limits of the thematic segments, we studied several indicating elements like the titles of the sections, the positions of the textual fragments and the linguistic expressions. Inside a thematic segment the sentences relate to the same subject.

3.2. EXTRACTOR

Constitutes the main module of the system. It consists in identifying and extracting relevant information from the text according to the linguistic diagrams' stored in a database. That while segmenting and classifying identity information of the individuals in heterogeneous cv, like all the information useful for the management of skills. The extractor identifies among other things, the name, the first name, the address, the age, the nationality, the sex, the marital status, the telephone number, the obtained diplomas, the professional experiences, the vocational trainings, the practised languages,…

We seek to directly exploit the textual organization of the authors' remark. The judgement of importance is founded primarily on what the author himself explicitly emphasized in his text. Several textual segments such as: vocational training…, trainings…, diplomas…, languages…etc, are as much relevant indicators used by the author to direct the attention of the reader towards certain information.

All this information will be stored in a database which its entries are presented as follows:

- Name, First name, Address, Age, Telephone number.
- Diploma.
- Training.
- Vocational training.
- Language.

Extracted information is expressed in the form of attribute-value and can thus be captured automatically.

Our methodology of extraction is based on a hierarchy of linguistic diagrams describing information to be extracted. A linguistic diagram is described by a set of rules of extraction. Each diagram takes part in the extraction.

One of the essential foundations of the system rests on the construction of a "base of linguistic diagrams" starting from a corpus.

The acquisition of these linguistic diagrams requires a systematic search of the texts in order to accumulate the diagrams. This search is done by a linguistic deliberation, in order to release the textual regularities.

To optimize the speed and the size of the base of the diagrams, the module of extraction integrates a linguistic base of the lemmatized diagrams. The latter are used by the module of the extraction according to user's needs'. It takes in entry a lemmatized text, carries out a morphological analysis and it retains only relevant information.

EXPLOITER

The exploitation of the extracted information is done according to a set of rules, these rules of extraction can be combined with the criteria of recruitment to extract only information having meaning for a given strategy of recruitment.

The system consists in identifying quite precise information of a text in natural language but also being able to represent it in structured form. For example, from a CV, our system will be able to identify, the name, the first name, the address, the age, the nationality, the sex, the marital status, the telephone number of the candidate as well as the obtained diplomas. The extraction is done by feeding a database. That makes it possible to classify the CV according to some criteria such as the geographical area, the activities' sectors or the skills fields. In addition, research on key words will be able from then on relating to all or part of the CV.

4. EVALUATION

An information extraction system is characterized by strong constraints relating to the processing time as well on the quality of extracted information. For the performance evaluation of our system we have tested it on a small corpus of CV. The latter are not annotated. For the candidate's marital status part (name, first name, civility, address, date of birth, email, telephone), the result without fault reaches on average 90%.

For the career's historical part (training, skill and Vocational training), the result reaches on average 60%.

5. CONCLUSION

Our system is an information extraction application using a linguistic approach. It carries out a linguistic processing of a text and store the information extracted into a database. The process takes into account hierarchical linguistic diagrams. The formats of texts supported by the system are HTML, PDF, DOC and TXT.

It allows the conversion of the applications (CV) into format TXT.

Our system is unilingual available in single-user version. It can be installed on any station equipped with a Java virtual machine. It has a flexible and modular structure, possibly enabling it to adapt to any extension and modification.

The system makes it possible to the experts to insert or remove linguistic diagrams from the database; as well to the responsible for recruitment to interrogate the database in order to obtain the application which answers its recruitment criteria.

Among future work, we plan to associate our system with a scanner and an OCR software for the processing of CV papers. To integrate a module of monitoring of the directories and emails accounts fed in applications' files thus allowing to the system a functioning in automatic mode.

To give to the system the possibility of being able to dynamically change its behavior to identify itself in which language the text is written (English, French).

REFERENCES

[1] ANDRADE M., VALANCIA A. (1997). Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. In : 5 international conference on Intelligent Systems for Molecular Biology (ISMB’97), Halkidiki (Greece), 25-32.

[2] AMINI M. (2001). Apprentissage automatique et recherche de l’information : application à l’extraction d’information de surface et au résumé automatique.

[3] APPELT D.E., ISRAEL D.J. (1999). Introduction to Information Extraction Technology. A tutorial prepared for IJCAI99.

[4] BRILL E. (1992). A Simple Rule-based Part of Speech Tagger. Proceedings of the Third Conference on Applied Natural Language Processing, ACL, 152-155.

[5] CARAVEN M., KUMLIEN J. (1999). Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In: Intelligent Systems for Molecular Biology (ISMB’99), AAAI Press.

[6] HARRIS Z. (1951). Structural Linguistics. The university of chicago press.

[7] HUMPHEREYS K., DEMETRIOU G., GAIZAUSKAS R. (2000). Two applications of Information Extraction to Biological science Journal Article: Enzyme Interactions and Protein Strustures. In: PacificSymposium on Biocomputing (PSB2000), Honolulu, 2000, vol.5, 502-513.

[8] LUHN H.P. (1961). The automatic derivation of information retrieval encodements from machine-readable texts. In: A.Kent edition, information Retrieval and machine translation, 1961, vol 3, Pt.2, 1021-1028.

[9] NAMER F. (2000). FLEMM : un analyseur flexionnel du français à base de règles, TAL, Vol. 41, n°2.

[10] OHTA Y., YAMAMOTO Y., ’KAZAKI T., UCHIYAMA I., TAKAGI T. (1997). Automatic construction of knowledge base from biological papers. In : 5 international conference on Intelligence Systems for Molecular Biology (ISMB’97), Halkidiki (Greece), 218-225.

[11] PILLET V. (2000). Méthodologie d’extraction automatique d’information à partir de la littérature scientifique en vue d’alimenter un nouveau système d’information.

[12] SAGGER N., FRIEDMAN C., LYMAN M. (1987). Medical Language Processing: Computer Management of Narrative Data. Addisson Wesley.

[13] SALTON G., WONG A., YANG C.S. (1990). Improving retreival performance by relevance feedback. Journal of the American Society for Information Science, vol 41, 288-297.

[14] SATOU K., ONO T., YAMARUMA Y., FURUICHI E., KUHARA S., TAKAGI T. (1997). Extraction of Substrutures of Protein to their Biological Functions by a Data Minig Technique. Intelligent Systems for Molecular Biology (ISMB97), vol 5, 254-7.

[15] STAPLEY B.J., BENOIT G. (2000). Biobibliometrics: Information Retrieval and Visualization from Co-occurencences of Gene Names in Medline Abstracts. Pacific symposium on Biocomputing (PSB 2000), Honolulu (Hawaï).

[16] SWANSON D.R., SMALHEISER N.R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. In : Artificial Intelligence, n°91, 183-203.

[17] THOMAS J., MILWARD D., OUZOUNIS C., PULMAN S., CARROLL M. (2000). Automatic Extraction of Protein Interactions from Scientific Abstract. In: Pacific Symposium on Biocomputing (PSB2000), Honolulu (Hawaï), vol 5, 538-549