Resume Parsing with Named Entity Clustering Algorithm

Swapnil Sonar Bhagwan Bankar
[1]iamswapnilsonar@yahoo.in [2]bhagwan.bankar1@gmail.com
SVPM College of Engineering Baramati, Maharashtra, India

Abstract:-

The paper gives an outlook of an ongoing project on deploying information extraction techniques in the process of resume information extraction into compact and highly-structured data. This online tool has been able to reduce lots of burden on the shoulder of users of recruitment agency. The Resume Parser automatically segregates information on the basis of various fields and parameters like name, phone / mobile nos. etc. and huge volume of resumes is no problem for this system and all work is done automatically without any personal or human intervention.

The resume extraction process consists of four phases. In the first phase, a resume is segmented into blocks according to their information types. In the second phase, named entities are found by using special chunkers for each information type. In the third phase, found named entities are clustered according to their distance in text and information type. In the fourth phase, normalization methods are applied to the text.

I. INTRODUCTION

Large companies and recruitment agencies receive process and manage hundreds of resumes from job applicants. Besides, many people publish their resumes on the web. These resumes can be automatically retrieved and processed by a resume information extraction system. Extracted information such as name, phone / mobile nos., e-mails id., qualification, experience, skill sets etc. can be stored as a structured information in a database and then can be used in many different areas.

In contrast to many unstructured document types, information in resumes is in a semi-structured form where information is stored in blocks. Each block contains related information about a person's contact, education or work experience. Even if it is in a restricted domain and semi-structured form, resume documents are not easy to parse automatically. They tend to differ in information types, information order, containing full sentences or not, etc. Also, conversion from other document formats (e.g. pdf ,doc, docx etc.) to text yields unexpected layout of information. To parse resumes effectively, the system should be independent of the order and form of information in the document. We assumed that resumes have a three level hierarchical structure where top most level contains segments. Segments consist of blocks that contain related information. Each block can contain several chunks which are named entities.

II. PREVIOUS WORK

Recent advances in information technology such as Information Extraction (IE) provide dramatic improvements in conversion of the overflow of raw textual information into structured data which constitute the input for discovering more complex patterns in textual data collections.

Resume information extraction, also called resume parsing, enables extraction of relevant information from resumes which have relatively structured form. Although, there are many commercial products on resume information extraction, some of the commercial products include Sovren Resume/CV Parser [4], Akken Staffing , ALEX Resume parsing [5], ResumeGrabber Suite

and Daxtra CVX [6]. There are four types of methods used in resume information extraction: Named-entity-based,rule-based, statistical and learning-based methods. Usually a combination of these methods is used in many applications.

- Named-entity-based information extraction methods try to identify certain words, phrases and patterns usually using regular expressions or dictionaries. This is usually used as a second step after lexical analysis of a given document [2, 3]. Rule-based information extraction is based on grammars.
- Rule-based information extraction methods include a large number of grammatical rules to extract information from a given document [3].
- Statistical information extraction methods apply numerical models to identify structures in given documents [3].
- The learning-based methods employ classification algorithms to extract information from a document.

  Many resume information extraction systems employ a hybrid approach by using a combination of different methods.

III. PROPOSED SYSTEM

Information Extraction Process:-

The designed Information Extraction System consists of 4 phases: Text Segmentation, Named Entity Recognition, Named Entity Clustering and Text Normalization.

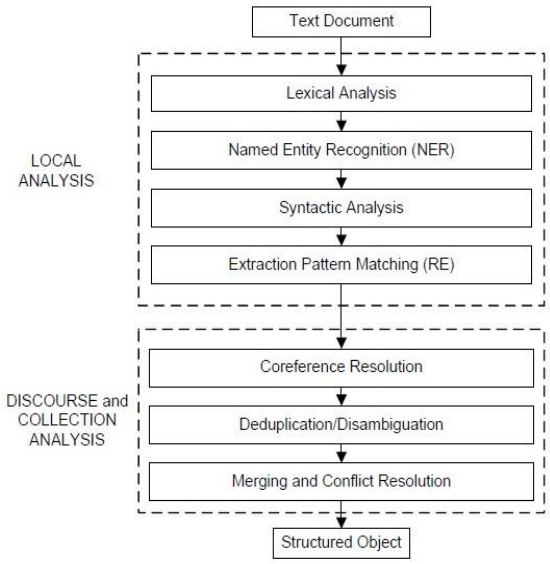

Figure1. Workflow of Resume Parser

1. Text Segmentation

Text Segmentation phase do work on the fact that each heading in a resume contains a block of related information following it. So in that case our resume will separate out into segments named as contact information, education information, professional details and personal information segment as shown in Figure 1.

Segment Type	Related Info under the Segment
Contact	Name
	Phone
	Email
	Web
Education	Degree
	Program
	Institution
Experience	Position
	Company
	Date Range

Table1. Segment containing extracted Information Types

A data-dictionary is used to store common headings in a resume which are definitely occurring in the resume. These headings are searched in a given resume to find segments of related information. All of the text information between the heading and the start of the next heading is accepted as a segment. One exception will possible or may occur is the first segment which contains the name of the person and

generally the contact information. It is found by extracting the text between the top of the document and the first heading. For each segment there is a group of named entity recognizers, called chunkers, that works only for that segment. This improves the performance and the simplicity of the system since a certain group of chunkers only works for a given segment. Segmentation is a crucial phase. If there is an error in the segmentation phase, chunkers will run on a wrong context. This will produce unexpected results.

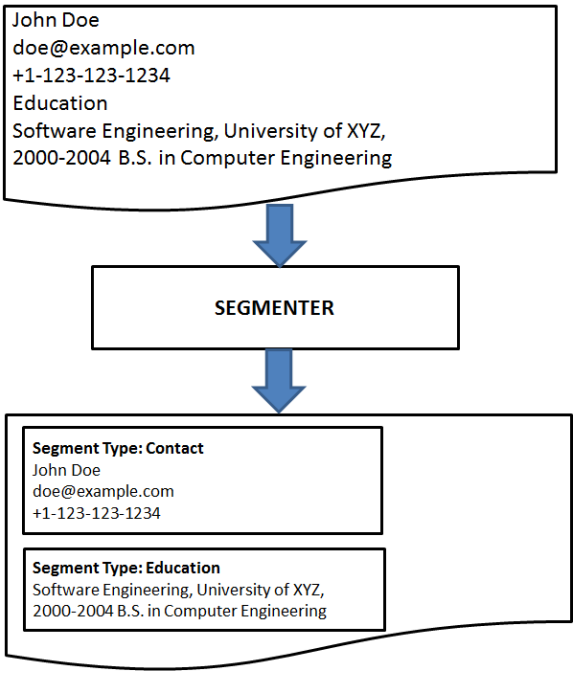

Figure2. Segmentation in the Resume by Segmenter

2. Named Entity Recognition

The tokenized text documents are fed to a Named Entity Recognizer. The term Named Entity refers to noun phrases (type of token) within a text document that have some predefined categories like Person, Location, Organization ,expressions of times, quantities, monetary values, percentages, etc. Numeric values like phone numbers and dates are also Named Entities.

Resumes consist of mostly named entities and some full sentences. Because of this nature of the resumes, the most important task is to recognize the named entities. For each type of information, there is specially designed chunker. Information types are shown in Table 1. Each chunker is run independently as shown in Figure 2.

Chunkers use four types of information to find named entities:

- Clue words: like prepositions (e.g. in the work experience information segment the word after "at" most probably a company name)
- Well Known or Famous names: Through data-dictionaries of well-known institutions, well known places, companies or organization, academic degrees, etc.
- From prefixes and suffixes of word: For institutions (e.g. University of, College etc.) and companies (e.g. Corp., Associates, etc.)
- Style of Writing Name of person: Generally the name of the person is written as First Letter capitalize then we will guess that this word possibly name of person.

Examples are "is-headquarter-of" between an organization and a location, "is-CEO-of" between a person and an organization, "has-phone number" between a person and phone number and "is-price-of" between a product name and a currency amount.

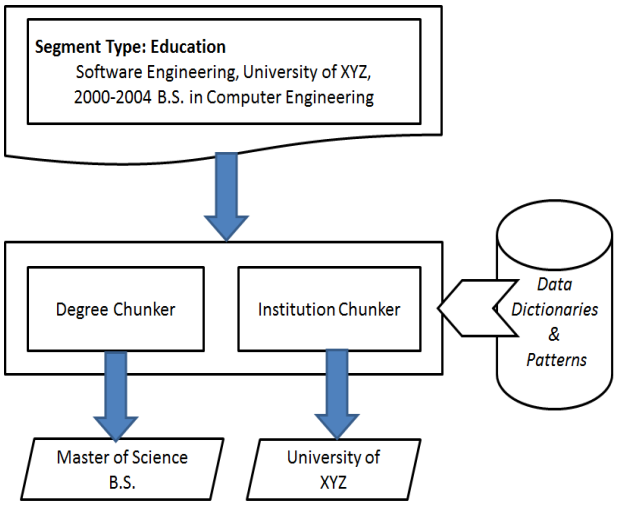

Figure3. Chunkers extracting from Segment Education

The chunkers produce an output that contains information about named entities as shown in Table2.

Named Entity	Start	End	Type
University of XYZ	22	39	Institution
B.S.	51	54	Degree

Table 2.Found Named Entities along with their Start & End position in the resume.

3. Named Entity Clustering

A loose definition of clustering could be "The process of organizing objects into groups whose members are similar in some way". So the cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Each segment (e.g. education information) contains a block of related information. For example, an education segment will have a number of blocks of information about educational institutions that a person attended. For example, an education information block can contain institution name, degree, major, and date information In the previous step we obtain many independent named entities as shown in Table 2.

The named entity are in the block of information are need to be grouped together to do the more process on it. Related information is defined as shown in Table 1.

Named entities (chunks) are grouped according to their proximity and type. The algorithm in Figure 3, tries to associate related entities into a group depending on their type and how much they are close to each other.

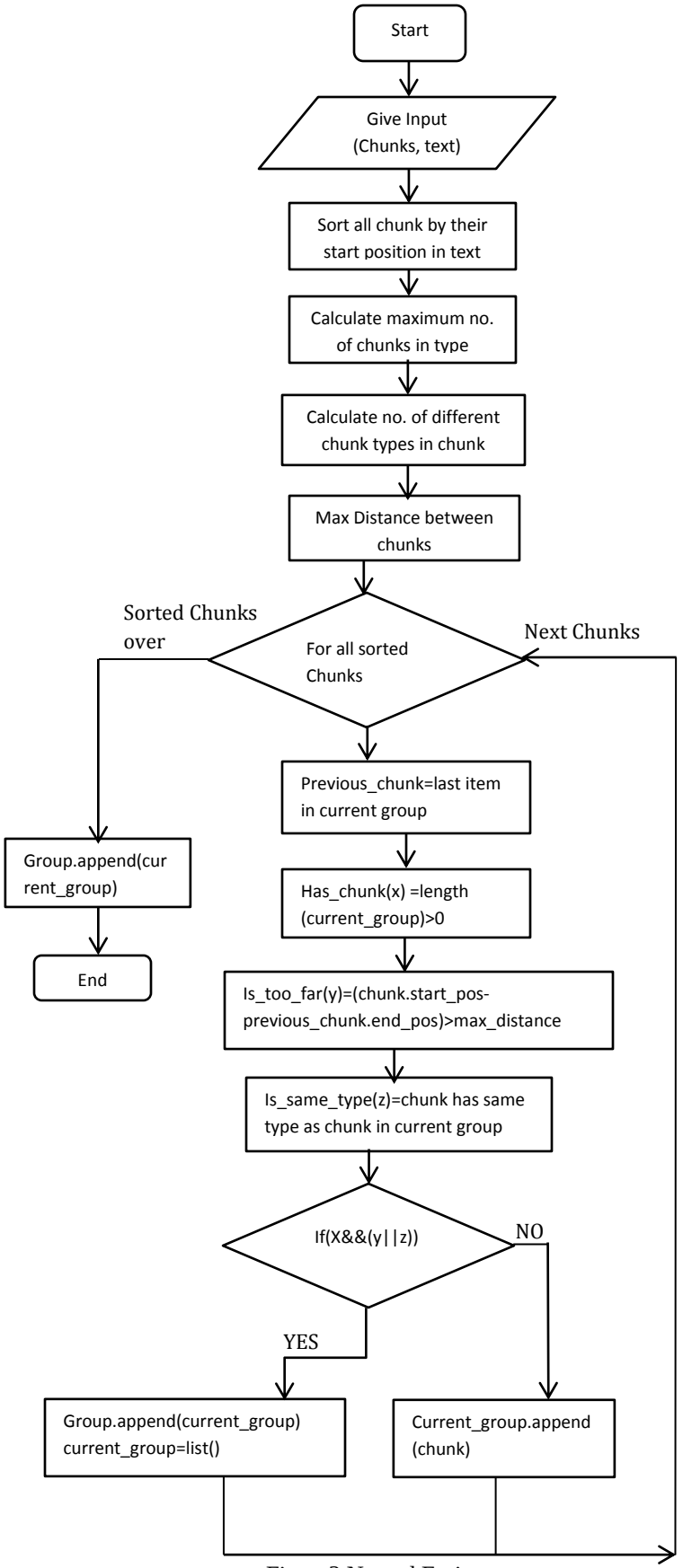

Figure3 Named Entity Clustering Algorithm

4. Text Normalization

In text normalization, some of the named entities are transformed to make it consistent. Table 3 shows some of the transformations performed on several text phrases.
In normalization phase, we expand some of abbreviations using dictionaries similar to the dictionary given in Table 4. For example, the abbreviation "B.S." is expanded as "Bachelor of Science". We also convert some of the text into a new form. For example, the first letters in a person's name is capitalized as shown in Table 3.

Some of the phrases are also converted to its most common form. For example, "University of ABC" is converted to "ABC University" as shown in Table 3.

Input	Output	Type
B.S.	Bachelor of Science	Degree
JOHN DOE	John Doe	Name
University of ABC	ABC University	Institution

Table 3.Applying Text Normalization using Data-Dictionary

Term (Full Form/word)	Abbreviation
Bachelor of Science	B.S.,BS ,BSc
Master of Science	M.S.,MS ,MSc
Bachelor of Arts	B.A., BA
Doctor of Philosophy	Ph.D., PhD
Doctor of Medicine	Medicine Doctor, M.D.
Bachelor of Computer Application	BCA,B.C.A

Table 4.Sample Data-Dictionary of the degree chunker

IV. Performance Evaluation

The focus in Information Retrieval research lays on text classification systems which make binary decisions for text document as either relevant or non-relevant with respect to a user's information need. Capturing the user information need is not a trivial task. We used precision, recall and F-measure metrics for performance evaluation [7].

- Precision measures the number of relevant items retrieved as a percentage of the total number of items retrieved.

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})}$$

- Recall measures the number of relevant items retrieved as a percentage of the number of relevant items in collection.

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})}$$

- The F-measure is the harmonic mean of precision and recall.

$$\text{F-measure} = 2 * \left[\frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \right]$$

The Precision, Recall and F-measure rates for segments will be predicted as follows. Segments will recognize well by the system. F-measure for segments is between 95% and 100%. This is because of the fact that the segment data-dictionary includes almost all of the common headers used in the resumes. The identification rates of named entities such as name, e-mail and education program will in acceptable ranges or any tolerated error includes in it so because these named entities has specific or changed format .

V. ADVANTAGES

This online tool has been able to reduce lots of burden on the shoulder of Applicant and HR Managers as well. The Resume Parser automatically segregates information on the basis of various fields and parameters like name, phone / mobile nos., e-mails id., qualification, experience, skill sets etc. and huge volume of resumes is no problem for this system and all work is done automatically without any personal or human intervention.

So, in a Nutshell, The Resume Parser will behave as a Suite, which will provide

1. A precise and Auto Profile fill-up Utility,
2. Extraction of predefined types of information from unstructured documents based on IE (Information Extraction) Technology,
3. As Recruitment Assistant,
4. Concise Information about Person.

VI. DISADVANTAGES

Complete system is dependent on web, computer and modern facilities like internet so this not useful in rural areas but now days with increasing globalization this disadvantage easily minimized.

While used the auto profile fill-up utility some of the error may occur during filling information or someplace problem in extracting information.

VII. APPLICATIONS

This online tool has been able to reduce lots of burden on the shoulder of jobseeker in Online Recruitment System.

Maintain the basic information of employees in the Company/Organization.

VIII. CONCLUSION

We presented a resume information extraction system based on Named Entity Clustering Algorithm. The developing system has a flexible and modular structure that can be extended easily. This system is use and test in Online Recruitment Agency project.

The resume extraction process consists of 4 phases. In the first step, a resume is segmented into blocks according to their information types. In the second step, named entities are found using special chunkers for each information type. In the third step, the found named entities are clustered into groups according to their distance in text and information type. In the fourth step, normalization methods are applied to the text.

After trying to work with Resume Parser it is found that it reduces the time consumed for generating a profile of user with extracted information by almost 50%.

IX. FUTURE SCOPE AND ENHANCEMENT

As a future work, we will expand the resume data collection set and improve the performance of the proposed system. After success of proposed system we also plan to extend it for other languages too.

REFERENCES

[1]. R.Grishman,"Information Extraction: Techniques and Challenges", Lecture Notes In Computer Science, vol. 1299, Springer-Verlag, London, 1997, pp. 10-27.

[2]. C. Siefkes and P. Siniakov, "An Overview and Classification of Adaptive Approaches to Information Extraction", Journal on Data Semantics, vol. 4, Springer, 2005, pp. 172–212.

[3]. S.Sarawagi,"Information Extraction", Foundations and Trends in Databases, vol. 1, 2008, pp 261- 377.

[4]. Sovren Resume/CV Parser, http://www. sovren.com/ (Accessed on Feb 2, 2012).

[5]. ALEX Resume Parsing, http://www.hireability.com/ ALEX/ (Accessed on Feb 2, 2012).

[6]. Daxtra CVX, http://www.daxtra.com/ (Accessed on Feb 2, 2012).

[7]. Ertuğ Karamatlı, Selim Akyokuş "Resume Information Extraction with Named Entity Clustering based on Relationships"