Document Zone Identification and Classification

By
Nauman Shamim
2011-NUST-MS CS-03

Supervisor
Dr. Khalid Latif
Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters in Computer Sciences(MS CS)

In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(July 2015)

Approval

It is certified that the contents and form of the thesis entitled "**Document Zone Identification and Classification**" submitted by **Nauman Shamim** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Khalid Latif**
Signature: _____

Date: _____

Committee Member 1: **Dr. Peter Bloodsworth**

Signature: _____
Date: _____

Committee Member 2: **Dr. Hamid Mukhtar**

Signature: _____
Date: _____

Committee Member 3: **Dr. Asad Anwar Butt**

Signature: _____
Date: _____

Abstract

Automated data extraction from resume has a variety of applications such as online recruiting, human resource management. An efficient technique for resume zone identification and classification can help data extraction from resume and assist resume analysis against a job description. The segmentation of resume into zones is a challenging problem as the order and number of resume sections, their length and content representation is not according to a set model. The classification of resume segments is difficult as the classes for possible resume segments are not well defined in any previous work. Another issue is that the lengths of constituent segments are highly varied across different resumes. In comparison to text classification and segmentation the problem of resume zone identification and classification is relatively unexplored, the approaches already proposed have limitations in terms of order of resume sections and classes. Based on the fact that a resume consists of sections and each section is preceded by a section heading we proposed a technique to efficiently segment a resume into its constituent segments. The textual and structural features along with named entities of a section heading are used to detect section boundaries. To detect the content type of individual segments we trained a SVM classifier over word vectors. We further trained the classifier over word vectors of named entities and root words from contents of constituent resume segments. For training and testing we develop a data set of 1730 segments from 300 resume, presently there is no suitable data set available for research in this area. The work shows that the proposed technique segments resume with high precision (0.91) and recall (0.85).

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Nauman Shamim**
Signature: _____

Acknowledgment

I am thankful to Almighty Allah, whose blessings have always been enormous and who bestowed the skills, knowledge and strength upon me to complete this thesis.I owe my sincere gratitude to my father and family members, who have supported and encouraged me throughout my thesis work. This thesis would not have been completed without their help, support and encouragement.

I dedicate this thesis to my mother and father who were always around me whenever I needed them. It is their unconditional love that motivated me to complete this work.

Table of Contents

1 Introduction **1**
- 1.1 Introduction . . . 1
- 1.2 Motivation . . . 2
- 1.3 Organization of Thesis . . . 3

2 Literature Review **4**
- 2.1 Literature Review . . . 4
- 2.2 Segmentation and Classification Of Scanned Documents . . . 4
- 2.3 Worked Done on Text Documents . . . 7
- 2.4 Conclusion . . . 9

3 Design and Methodology **11**
- 3.1 Identifying Document Zones . . . 11
- 3.2 Segmentation . . . 13
  - 3.2.1 Zone Boundary . . . 14
  - 3.2.2 Boundary Markers . . . 14
  - 3.2.3 Boundary Marker Issues . . . 15
  - 3.2.4 Zone Uniqueness . . . 15
  - 3.2.5 Nesting of Zones . . . 17
  - 3.2.6 Zone Overlapping . . . 17
  - 3.2.7 Segmentation Approach . . . 18
  - 3.2.8 Segmentation Using Dictionary . . . 18
  - 3.2.9 Segmentation Using Structural Boundaries . . . 20
  - 3.2.10 Named Entities . . . 22
  - 3.2.11 Segmentation Using Combined Approach . . . 24
- 3.3 Classification . . . 25
  - 3.3.1 Text Preprocessing . . . 26
  - 3.3.2 Handling Numeric Data . . . 27
  - 3.3.3 Handling Nouns . . . 27
  - 3.3.4 Text to Word Vector Conversion . . . 27

4 Testing and Evaluation **29**
4.1 Data Set 29
4.2 Evaluation Criteria 30
4.3 Segmentation Results 30
4.4 Classification Results 31
4.5 Results Summary 33

5 Conclusion **34**
5.1 Future Directions 34

List of Figures

3.1 Resume layout-1 13
3.2 Resume layout-2 14
3.3 Boundary marker issues 16
3.4 Resume segmentation 18
3.5 Segmentation using dictionary 19
3.6 Detection of structural boundaries in resume 20
3.7 False heading detection example 21
3.8 Named entities in resume 22
3.9 Avoiding false section headings using named entities 23
3.10 Segmentation Combined Approach 24
3.11 Section heading detection 26

List of Tables

4.1 Classification Data Sets 30
4.2 Segmentation Results 31
4.3 Classification Results-1 31
4.4 Classification Results-2 32
4.5 Classification Results-3 32
4.6 Summary of Results 33

Chapter 1

Introduction

1.1 Introduction

Conventionally an organization in need of manpower advertises vacancies, collects job applications and perform selection after evaluation. This simple task becomes challenging for organizations dealing with large number of vacancies which are diversed and spreaded over longer distances. The online job portals such as [15] have made it easier to collect job applications, a candidate may apply for a job by either submitting a conventional resume [28] documents or via an online form. The job applications received through an online service is usually stored in an information system for further processing such as evaluation against a job description. Online forms provide structured information about an applicant where as resume documents contains unstructured information, this is why the extraction of information from resume documents is a challenging problem.

This work is about extraction of information zones from a resume document, this is a two step process, the first step is to identify and extract the zones, second step is to assign it a type according to its contents. This area of research has not been extensively explored yet however this work can benifit from research done in the field of information extraction from documents.

A resume is a document that contains various pieces of information about a person, specifically it is a summary of one's educational and professional experience [28] ; normally the information is placed in distinct sections with title, such as qualification, education and experience. The document structure of a resume is flexible, the number of sections, section titles, section placement order, contents of a section etc all are loosely defined or are not according to a standard model. Such nature of resumes makes it difficult to develope a system that can fully understand resumes and utilize the infor-

mation they contain A portion of document that is independently providing some information is considered as zone [31] such as the abstract or references in a scientific publication. A resume consistes of multiple zones, the number of zones their type and their order is not according to any specific format, efforts has been made to extract these zones and classify them according to their contents. This work can help classifying the document itself and also has applications in online information processing such as comparing quotations or tenders and removing advertisements from the web pages.

The generic field of study for this research is known as Information Reterival (IR) [18]. IR helps developing systems which can reterive information from a repository given a simple query. In this thesis IR and classification techniques have been used to develop a system that can extract desired information from resumes. This thesis proposes a four step process for zone extraction. These steps are briefly outlined here and are further explained in subsequent chapters.

1. **Pre-Processing**
   A resume document normally is not a plain text file. It usually contains formatting information which we intend to ignore,this formatting is removed.There are lots of stop words which are removed such as prepositions, articles,pronouns, common verbs etc.To improve the accuracy some substitutions are made such as durations, places names, addresses, can be represented in more usefull forms (discussed later).

2. **Segmentation**
   Dividing the documents into parts that contain information relevant to one topic such as "Education", all such segments are to be identified and same segments are to be merged into one.

3. **Classification**
   Assignment of section titles to each segment.

4. **Evaluation** Evaluation of performance and accuracy of the developed system against manually segmented and classified segments.

1.2 Motivation

This area of research has direct industrial application. Many job portals and Human Resource Management (HRM) systems offer online resume processing services to serve various organizational needs such as automated information repository building, candidate profiling, job profiling, and document

summary generation, e.g. automated recruiting softwares help in finding the best fit for a job. Such services help organizations save time and human resources, help avoid human error, provides up-to-date knowledge (in some cases) and adds reliability to the overall process. Some of such well-known softwares and companies are Alex Desktop, Daxtra [7], eConn, eGrabber [8] , RChilli, Sovren and TextKernal [13] (Recuriter.com).Information provided by such automated tools are much more updated,cost effective and error free. Online recruiter portals are providing services to hundreds of companies, companies such as Microsoft, Saudi Aramco, Covidien, Cineplex, and McKesson. Similarly social website, candidate recommending portals and job boards are using information retrieval methods in some form which signifies the importance of this work.

1.3 Organization of Thesis

The organization of rest of the theis is as, Chapter 2 provides literature review related to this work, Chapter 3 consists of design details of the proposed system, Chapter 4 is about the implementation details of proposed system, Chapter 5 discusses the performance, accuracy and other evaluations. And finally Chapter 6 concludes this word and presents an outlook on future direction. An annexture about related tools and litarature is also developed for interested reader.

Chapter 2

Literature Review

2.1 Literature Review

The work done so far can be categorized in two main streams, work done for (i) documents having text nature,(ii) document having image nature or scanned document. Both of these branches have been explored extensively, some of the techniques follow pure IR models while others are a hybrid of image processing and IR techniques. The initial task has always been to detect zones by defining a zone boundary. For document images usually a rectangular boundary is taken [30] , for text documents the general approach has been to use the textual features of contents. The algorithms used for document content analysis can be divided into three basic categories (i) Type specific detection, (ii) Page classification and (iii) Zone classification. [1] the first one is used to detect zones of specific type such as text or tables or figures, second type considers the whole document as one piece of information and classify the documents belonging to one of the several categories (the case of web) finally the last type of algorithms uses segmented documents and each segment is considered as one single piece of information, its contents are analyzed and a zone type is assigned. We first present zone classification in document images and later in text documents.

2.2 Segmentation and Classification Of Scanned Documents

A variety of methods have been proposed for segmentation and classification of scanned documents. Though our work is not related to scanned documents however to provide a contrast of work over text documents an

abridged overview of some segmentation and classification techniques of this domain is presented here.
For segmenting a document image the general approach used is to develop a contrast between segments that contains useful information and the areas which serves as document background such as white space.Various image processing techniques have been used to find relationship between pixels that belongs to segments or zones while same has been used to detect the document background. A weak relation between document background and segments considered as segment boundary. These techniques can be used to detect segments of rectangular as well as non-rectangular shapes. For example , the Connected Component Aggregation [4] follows a hierarchical clustering like approach and measures the closeness of each pixel to others in the document and then assigns labels. Another approach is to identify background space by using vertical and horizontal rectangles, combined with connected component it extracts background areas and use it as mask to extract the foreground areas.

Antonacopoulos [3] proposed a white tile approach to detect segments of rectangular as well as non-rectangular shapes. In this approach white space between lines and around the text is estimated by taking histogram of document image.The base of text lines is estimated and blank areas in the image are converted to white tiles. The tiles are connected to create a mask, this mask is used to uncover the foreground areas which actually are segments of the document image that contains text, image or any other graphical shapes. There are other such techniques which are based upon finding a relationship such as similarity between a pixel and its neighboring pixel. In next sections we will present different approaches adopted by various people working in this area.

For classification of segments of a scanned document,Support Vector Machine,Neural Networks,statistical methods,feature vector approach probabilistic and various clustering schemes have been used and have produced good results. Work has also been done to find best parameters to be used for these techniques. Following is a brief account of prominent work in this area.

Feature vector approach considers that each zone has certain features or key elements that can describe it and can be used to distinguish it from other zones. A group of such elements is termed as feature vector, a zone in this case is a collection of feature vectors, once the feature vectors are formed various similarity measures exists that can be used to perform classification of a zone. Sivarama Krishnanet [27] used zone mean run length, mean variance, spatial mean and zone size ratio with document for each zone. Yalin Wang [30] added two more features to this approach i.e. number of text glyphs in

zone and blank area of zone, they used 1600 document image, for optimal solution, binary decision trees and Viterbi algorithms are used, they also worked to measure the performance of their proposed method, they improved the accuracy of their own work i.e. from 97.4% to 98.7%.

Yaling et.al in 2002 [30] improved their own feature vector technique, this time a probabilistic model is used with same binary decision tree and Viterbi algorithm however the major difference between the two is that this time the feature set is reduced from 67 to 25 which greatly improved the performance, they experimented with 1600 technical document UWCDROM III with over 24177 zones and improved thecite accuracy of their previous work by 0.5 %. The zones were classified to one of nine classes. The basic technique was similar to that of or their previous work.

Yalin Wang et.al worked further on their technique and improved the classification of zone contents by proposing a new algorithm; they used 25 feature vectors and 9 classes[32]. In this work they introduced a new background analysis scheme and defined two background features,(i)the ratio of black pixels in a zone to its total number of pixels,this however is an established scheme in page segmentation to identify zone type, (ii)total area of large horizontal and large vertical block. For second feature they also defined a background analysis scheme. They extracted two types of features run length and spatial. Run length is the number of contiguous and similar (background or foreground) pixels in a given direction. They used background/foreground run length mean and variance of four directions. Spatial features measure the foreground pixel distribution information, each pixel is assigned a weight and spatial mean and variance were measured in all four directions. Along with classical spatial features background features were combined. The classification is once again done using decision trees and HMM. The experiments were performed on standard document database UWCDROM III. The accuracy achieved was relatively high then previous work.

Partial Least Square Method is another approach that worked well for classification of image segments.The basic idea is to perform a comparison of large set of commonly used features and include features that are known to produce good results. To perform a comparison a total of 8 features were extracted, 3 of them were selected based on their best performances in Content Based Image Retrieval systems while others are commonly used features [17], in short, mean and variance of run lengths in various directions, connected components, nearest neighbor features, image scaling, and texture features were used, in most cases a histogram was used. Standard database UWIII was used which contains about 1600 documents. Two classifiers were experimented i) k-nearest neighbor ii) log -linear classifier using maximum entropy criterion. The results were good and a healthy error rate of 1.5%

was achieved.

Wael et.al used SVM for classification, the basic idea is to first extract the low level features from zones and then perform Partial Least Squares (PLS) to reduce the feature space,features that are most dominant are extracted this way. These features are classified using a hybrid scheme which is based on binary SVM [1]. The technique uses background, foreground, spatial, structural, autocorrelation, local binary and texture feaures. Most common featuresinclude document run lengths, foreground, background fractions, binary features include rotation and gray scale invariance (it is said in literature that Local Binary Features are good measure of local distribution of binary textures). Instead of using one to one or one against many SVM they used a hybrid scheme and developed an indicator classifier, the basic binary SVM (one against other was constructed) the indicator classifier shows which classes a zone does not belongs ultimately leading to strong candidates. The standard data set UW was used containing 1690 documents with 24531 zones; an accuracy of 97.3% was achieved.

Zaidah et.al carried out a research for comparative study among BPNN, SOM, RNN and SVM. The classification was done for text, images, graphs and tables[14].A huge data set (which actually a mixture of multiple data sets or extraction from various sources) was used to carry out the comparison. Around 100 datasets for each category were cropped and about 400 data sets were created this way, 50% of these data sets were used for training and 50% for testing. The qualitative results shows that SVM perform better over all.

It is observed that in most of the literature zone classification is done using feature extraction and probabilistic models, classifiers that works best for feature vectors are most popular however in other classifiers SVM performs better than neural network, it is also found that feature reduction is mandatory in all cases

2.3 Worked Done on Text Documents

2.3.0.1 Subtopic Structuring

The work on document segmentation started with classical Information Retrieval methods; in this regards work of Salton and Buckly [25]is considered quite comprehensive. They worked on full length document structuring and used articles from encyclopedia and electronic mail. According to their work two paragraphs are same or similar if they are similar in length as well as in contents (sentence by sentence), however they also worked on finding the similarity between two paragraphs that are not of same length but contain

similar contents.

Hearst and Plaunt [12]presented a different approach for finding paragraphs having similarities, they imposed an even size block structure on the document, they defined block as collection of sentences and experimented with different block sizes (typical 2 to 3). Similarity among adjacent pair of blocks was measured using tf, idf. Blocks having high similarity were considered to be part of same paragraph or subtopic. They further extended their work for main topic / sub topic queries i.e. the user can specify the subtopic along with the context or the main topic for this the document terms were categories as belonging to main topic and subtopics, separate indexes were built for both of these. They left the task of measuring paragraph's similarity with main topic for future work.

2.3.0.2 Lexical Chains

Lexical chains have been used extensively for document segmentation;a chain is a sequence of words having a degree of cohesion among them. Morris presented the first lexical chain computation model[22](a classical one), many more were developed later.Barzilay and Elhadad(Resina Barzilay, 1997)have used lexical chains for text summarization and for finding topic boundaries.

2.3.0.3 Linear Text Segmentation

Choi presented an algorithm[6]that is 2 times more efficient and seven times faster than [24] which is considered state of the art in topic boundary detection. The basic idea is to construct stem frequency for each sentence and then compute similarity between two sentences using cosine similarity an image of similarity matrix is constructed where high similarity is represented by bright pixels. After computing similarity matrix ranking is performed, the ranking is done based on a new ranking scheme in which a rank matrix is calculated , in a 11 by 11 neighboring region of a cell, all lower similarity cells are counted this count is considered as rank of the cell. Finally clustering is performed to find the location of segment boundaries. Clustering uses three parameters, number of segments, area of segment and sum of ranks in a segment. Density is calculated as ratio of area and sum of rank in the area, a point which maximizes the density becomes the split point or boundary. For b boundaries b densities and b gradients are calculated, a significant decrease in gradient value suggests that optimal segmentation is achieved. By this method number of segments is calculated automatically, initially whole document is treated as one segment, divisive clustering is performed based on the densities and continued till there is a sharp decline in gradient of density.

2.3.0.4 Statistical Model for Text Segmentation

Yamron, et al. proposed a statistical scheme for domain-dependent text segmentation where as Utiyama and Isahara[29] proposed a statistical method that finds the maximum-probability segmentation for domain independent text.Yamron, et al. trained a hidden Markov model (HMM). They used states of HMM to represent topics. Given a word sequence, their system assigns each word a topic so that the maximum-probability topic sequenceis obtained. Their model is similar to that explained in [19]. The work is based on classical Bayes probability theorem and some suppositions. It is supposed that different topics have varying distribution of words, topics which are not same are statistically independent and words within a topic are statistically independent. Two important probabilities used in their work are , probability of segmentation and probability that a word belongs to a certain segment. Probability of segmentation is a function of segment length for which certain assumptions are made however no prior probabilities are used. The second probability is calculated using word frequency in a segment, total number of words in a segment and total number of different words in segment. Once probabilities are calculated maximum Bayes probability is determined which is considered as segmentation cost. Finally an ordered graph is created where nodes represents word positions edges represents cost of segmentation, where segment consist of all the words between two word positions. To find optimal solution suitable graph algorithms that already existed or dynamic programming is used.Utiyama's scheme [29] do not require any training asthey calculated probabilities from given text, no prior probabilities are required, word frequencies/densities are used as probabilities which intuitively seems reasonable method as higher densities means that the word has been discussed in detailed. The work can be extended and experimented with training data.

2.3.0.5 Latent Dirichlet Allocation

The method is based upon statistical method developed by Utiyama [29]. The used LDA for defining probabilities. Their research showed [21] that LDA performs better than[29] and other unsupervised approaches.

2.4 Conclusion

The area has been addressed from various aspects by researchers from different domains, topic segmentation and text summarizations are two main areas that are further researched for different purposes. Classical work such

as lexical chains has reliability problems and is affected by common writing styles and language. Systems based on lexical chains might work well for one language but perform poorly on other. Statistical methods depends upon prior probabilities which might not be available for all data sets, also methods that do not use prior probabilities are based on assumptions, suitability of these assumptions for all data sets remains in question. It is observed that common methods of finding probabilities and probability densities are based upon word frequencies and word weighing schemes; also common similarity measure is cosine similarity or its modifications. Hidden Markove Model and graph theory has been used to model segmentation problem as graph and to find optimal solutions on these graphs,these methods are not scalable . More recently this area has been explored using data mining techniques such as clustering. However researchers are now merging techniques and developing hybrid techniques. As the information retrieval domains and requirements have changed, general purpose techniques can be modified and improved to solve well defined problems such as this more efficiently.

Chapter 3

Design and Methodology

A resume document consists of multiple sections and each contains information that is different from the other. The quest is to computationally understand and isolate these sections such that the information contained can be correctly used. The task can be completed in many possible ways however the one adopted here is explained in his chapter. The chapter proceeds as, first the document Zone Identification and Segmentation in resume is introduced and later the approaches developed to achieve this task are explained in details. Some closely related terms used here are defined to avoid any confusion.

Section: A part of the resume that contains information about one single topic such as "Education", "Experience" or Objective , can consist of single or multiple lines.

Segment: A part of resume that our system considers different from its neighbor in the document, can consist of single or multiple lines. Segments are pieces of a resume that may or may not be one complete section of resume.

Zone : A zone is same as resume section.

3.1 Identifying Document Zones

Identifying nature of a document or understanding the boundaries where a topic ends and a new topic starts is an ordinary daily life task, human brain performs all the necessary processing in such a smooth way that the task seems trivial and simple however it is not true in computation perspective. Following is an outline of activities that may be performed by the systems that intend to understand the contents of a document, that contains multiple pieces of information such as a news paper.

- Build a context

- Detect start / end
- Search for title/topic
- Understand the nature/type of document
- Ignore non-readable / irreverent items
- Analyze contents
- Build an opinion

A human brain working in a mysterious way can do all these and much more than that, however identifying a document as newspaper is not possible if this is a first encounter of a human brain with it.

For a computer system to identify zones in resume, it requires comprehensive background knowledge of each section that can be a part of resume. Resumes do not contain same number of sections also their order and contents are not same. Following is an outline of information that a computer system would need to develop a good knowledge base for understanding a resume.

- Number of distinct resume sections
- Order in which resume sections may occur
- Titles / Section headings
- Boundary condition of sections
- Type of information contained in each section
- Length of sections

Various sections of resumes need to be treated independently to obtain these pieces of information, one approach which is also adopted here is to obtain collection of resume sections and develop a technique to analyze and identify a resume section if presented independently. The over all approach is based on two distinct tasks

1. Segmentation
2. Identification

Following sections provide conceptual and methodological details of these two tasks.

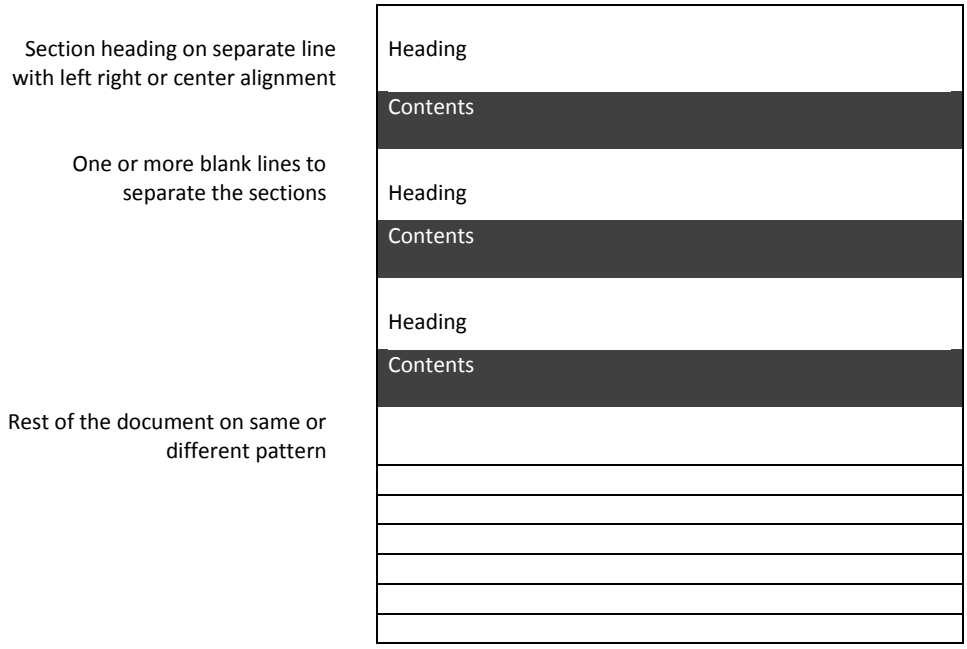


Figure 3.1: Resume layout-1

3.2 Segmentation

The task here is to mark the position in the document where a zone starts or ends. The problem seems similar to "topic boundary detection" [2] however the two are different in terms of nature and complexity. Following are certain aspects that are taken into consideration for segmentation, each of them is explained and analyzed in context of resume documents.

1. Zone boundary
2. Uniqueness of zones
3. Zone Order
4. Nested zones
5. Overlapping among zones

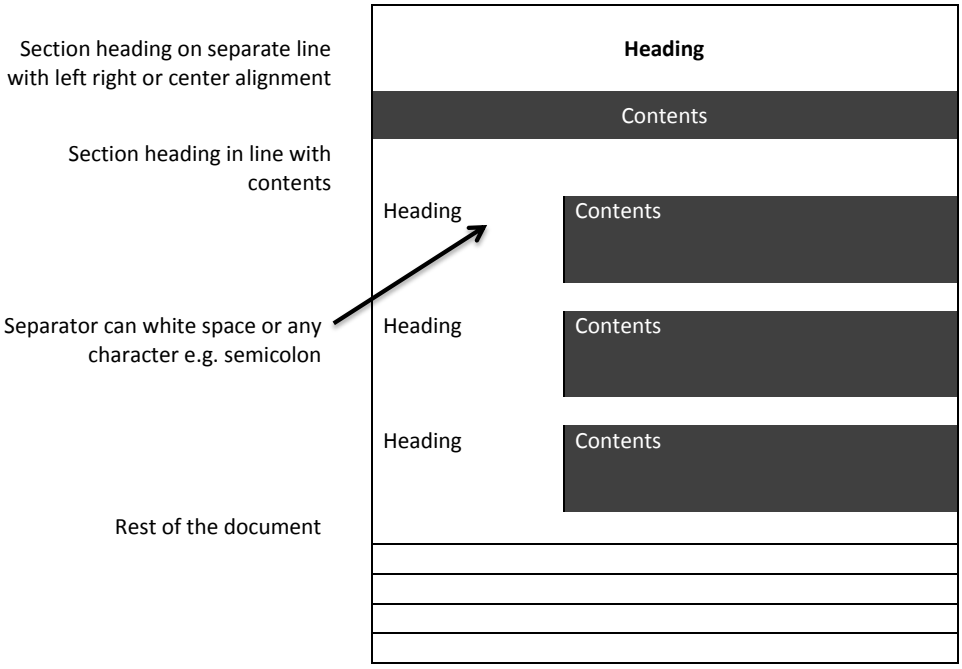


Figure 3.2: Resume layout-2

3.2.1 Zone Boundary

A zone in a resume is not a continuous paragraph, in usual cases it is a list of features regarding some qualification that is provided with a title, we call it a zone label here, usually this title is on a new line and is formatted differently than regular text.

3.2.2 Boundary Markers

Following structural and visual boundary markers are observed resume documents.

- A zone label placed at left most corner of a new line.
- A zone label followed by punctuation and text runs after this punctuation
- Use of multiple spaces or tab character after zone label while contents follow after on same line.

- A graphical separator such as a line or table, in some cases left most column is reserved for zone labels.
- Different formatting styles for zone label and contents
- Specific text formatting Following are the issues in using these boundary markers.

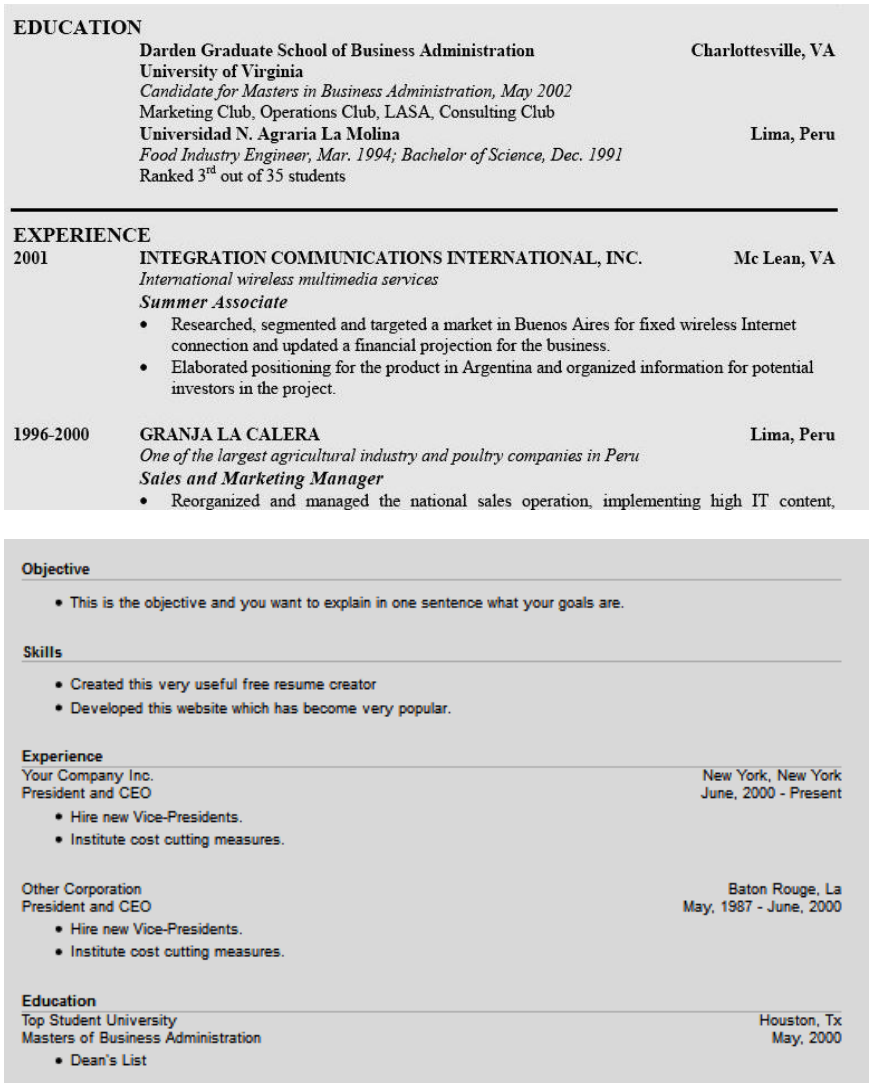
In most of the cases the zone labels are placed at the beginning of the line however right and center alignments are also observed. Similarly multiple line breaks and mix of different styles are also used to define a structural and visual boundary of sections or zones in resumes.

3.2.3 Boundary Marker Issues

- Zone labels are user defined and are not from a predefined set of labels, making it difficult to identify them as zone labels.
- Structural and formatting styles of zone labels can also be used with in zones making it difficult to establish a rule to distinguish labels from contents.
- Boundary markers may not be present at all
- Graphical boundary markers such as lines are hard to detect
- Graphical separators are difficult to identify and are not consistent
- False boundary markers may be present in the document

3.2.4 Zone Uniqueness

It is possible that same zone label may be repeated with relevant but different information or part of the zone may be placed independently with same or different label. A zone may or may not be present uniquely in terms of its label and contents. Ideally all pieces of information related to one aspect of certain qualification should be placed under one zone label, such as all details of education should be in Qualification section. However in practice this approach is rarely followed and contents that should belong to one zone distributed over multiple zones. In some cases such as education some details are placed in experience zone or training zone, this causes zone uniqueness problem. Coherence of information is contextual, a piece of information can follow any context i.e. can be placed in any zone however it changes its



EDUCATION

Darden Graduate School of Business Administration — **Charlottesville, VA**
University of Virginia
Candidate for Masters in Business Administration, May 2002
Marketing Club, Operations Club, LASA, Consulting Club
Universidad N. Agraria La Molina — **Lima, Peru**
Food Industry Engineer, Mar. 1994; Bachelor of Science, Dec. 1991
Ranked 3rd out of 35 students

EXPERIENCE

2001 **INTEGRATION COMMUNICATIONS INTERNATIONAL, INC.** — **Mc Lean, VA**
International wireless multimedia services
Summer Associate
- Researched, segmented and targeted a market in Buenos Aires for fixed wireless Internet connection and updated a financial projection for the business.
- Elaborated positioning for the product in Argentina and organized information for potential investors in the project.

1996-2000 **GRANJA LA CALERA** — **Lima, Peru**
One of the largest agricultural industry and poultry companies in Peru
Sales and Marketing Manager
- Reorganized and managed the national sales operation, implementing high IT content,

Style-01 Physical markers are used as section boundaries a long with section heading on new lines. Merger of two styles

Objective

- This is the objective and you want to explain in one sentence what your goals are.

Skills

- Created this very useful free resume creator
- Developed this website which has become very popular.

Experience

Your Company Inc. — New York, New York
President and CEO — June, 2000 - Present
- Hire new Vice-Presidents.
- Institute cost cutting measures.

Other Corporation — Baton Rouge, La
President and CEO — May, 1987 - June, 2000
- Hire new Vice-Presidents.
- Institute cost cutting measures.

Education

Top Student University — Houston, Tx
Masters of Business Administration — May, 2000
- Dean's List

Style-02 Physical markers are used as section boundaries in different way, along with section heading on new lines. Merger of two styles

Figure 3.3: Boundary marker issues

cohesiveness with surrounding information. Determining a correct context for any piece of information found in resume is not a simple task for certain cases it might be very easy such as education is not a suitable context for salary however "experience" and "education" both are correct and suitable contexts for "training".

3.2.5 Nesting of Zones

It is often the case that user presents information in a hierarchical fashion. Umbrella zones are created to present more abstract information and for specifics, sub-zones are created under respective umbrella zone. The concept of sub-zones causes problems in segmentation and information extraction. Following is a short account of some of these problems.

- Definition Problem ; How user defines umbrella zones, information seems abstract to one user may seems concrete.
- The zone identity problem ; Zone labels for umbrella zone in one resume can be normal zone labels in other resume.
- The boundary problem; boundary markers used for umbrella zones and sub-zone can be same.
- Nesting depth problem; There is no way to determine exactly how many levels of nesting is applied and how much details is placed at each level
- The layout problem; The order and layout of umbrella zones is user dependent and there is no single layout style that persists among resumes.

3.2.6 Zone Overlapping

Zone overlaping comes into play when two or more sections in a resume have information that is similar in context or is repeated. Large pieces of information or repeated resume sections can be considered as normal sections however the repeated inforamtion has to be detected later. However information that consist of one or two lines is difficult to deal with.To elaborate the overlaping problem lets take an example resume of an IT professional, it is possible that list of programming languages known to the person can be part of education and experience sections both. It may makes some sense to the user and might present information in more logical way but such overlapping makes it hard to distinguish between similar sections of resume.

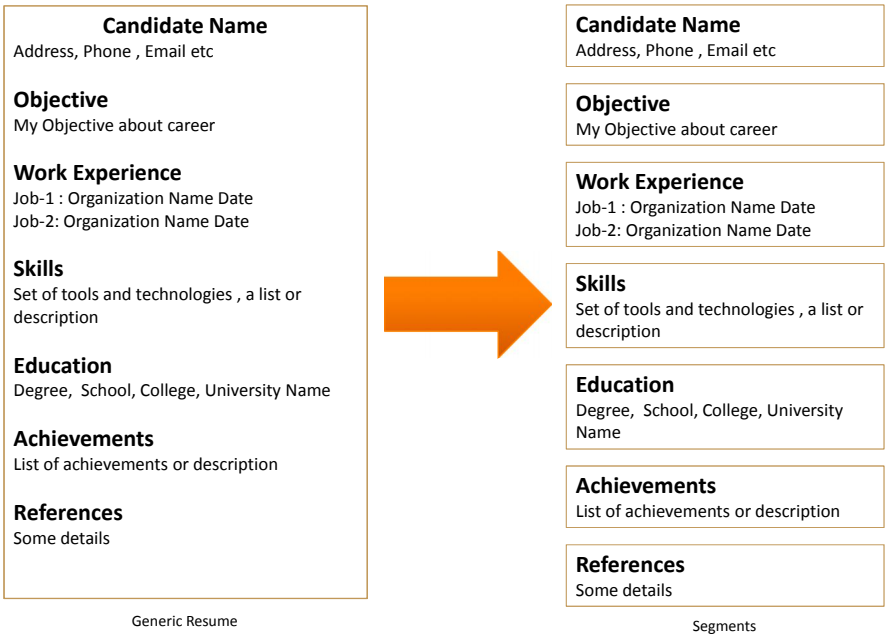


Figure 3.4: Resume segmentation

3.2.7 Segmentation Approach

Keeping in view the issues inherently present in resume documents several approaches were devised and experimented,following is some details of these approaches.

3.2.8 Segmentation Using Dictionary

This approach makes following assumptions

- A dictionary of all possible zone labels is present
- Each zone is preceded by a zone label
- Each zone label is placed on a new line with no text running after the zone label
- A zone label that is placed in line with paragraph is always the placed in the beginning of the paragraph

- A zone label that is placed in line with paragraph is always followed by punctuation
- A zone label must not repeat inside the zone itself

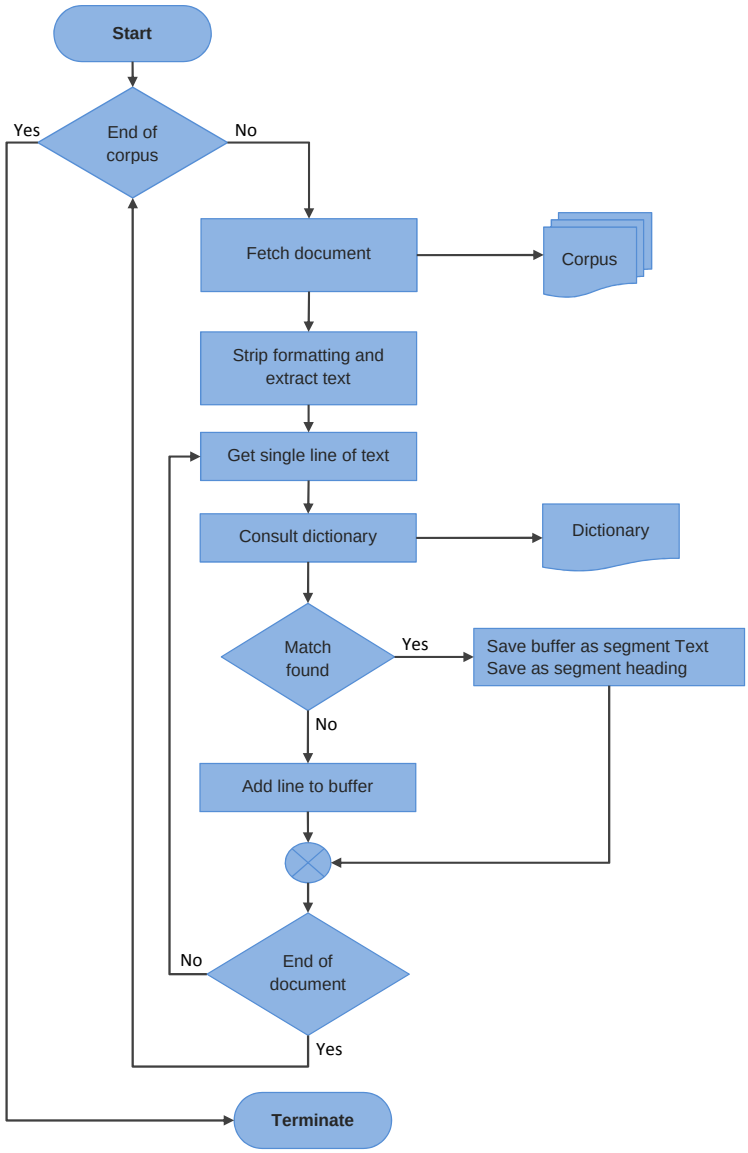


Figure 3.5: Segmentation using dictionary

The approach work well as long as the assumptions are being fulfilled however not all resume document have zone labels present on a single line and it is difficult to form a dictionary of all possible zone labels.

3.2.9 Segmentation Using Structural Boundaries

In this approach the structural information of zone boundary is used to mark the start and end of a zone segment. If certain structural pattern is present before zone or with zone label the place is considered as beginning of a new zone. After a analyzing sufficient number of resume documents a list of Key Structural Information was defined, the kSI are enlisted here according to their priority.

A sample execution of section heading detection algorithm, the algorithm marks the lines as candidate section headings, or indicates the the reason why the line is not considered as section heading candidate.

Candidate Name (Candidate)
1234 West Hickory St APT # 2, Denton, TX 768, (Numeric Data) (Multiple Conditions Failed)
Phone -54321-678-123, Email: mymail@exp.com (Numeric Data) (Multiple Conditions Failed)
(Empty Line)
OBJECTIVE (Candidate)
To build a career as an Electrical Engineer through an internship or fulltime employment. (Invalid Articles)
EDUCATION (Candidate)
(Split Line , First Part (Invalid Symbols)
MASTER OF SCIENCE (Electrical Engineering) GPA 3.66(2007)
University of North Texas, Denton, TX (Invalid Punctuation)
BACHELOR OF SCIENCE (Electrical Engineering) GPA 3.9 (2006)
Rajiv Gandhi University of Technology Bhopal, India (Invalid Punctuation)

SKILLS (Candidate)
Programming and Assembly languages: C/C++, VHDL, MATLAB, Visual Basic (First Part Candidate)
Stimulation tools: CADENCE, LTSPICE.XILINX, Modelsim, Precision … (First Part Candidate)
Platforms: MS-DOS, Mac OS, Windows XP/2000/NT, VISTA, UNIX, MAC (First Part Candidate)

PROJECTS (Candidate)
Implemented CMOS DOWNCONVERSION MIXER FOR GSM 1.92GHZ Receiver: Used double … (Numeric Data)
Digital Phase Locked Loop(DPLL) 715 MHZ: Used voltage controlled ring oscillator … (Numeric Data)

WORK EXPERIENCE (Candidate)
(Split Line , First Part Invalid Symbols)
Customer Service Representative (Lab Assistant) January 2008- Present
University Of North Texas, Denton, TX (Invalid Punctuation)
Responsible for managing a lab, handling a class and providing assistance to … (Invalid Punctuation)

Figure 3.6: Detection of structural boundaries in resume

- Single or multiple empty lines ; It is a common style to highlight the beginning of a zone
- Multiple words in the beginning of a line and rest of the line is empty, usually these multiple words do not conform to a sentence structure ; In case zone label is placed on a new line however there is no empty line before or after the label

- Multiple white spaces after few words, the words not conforming to a sentence structure ; It is the case where a zone label is in line with paragraph text and tab character or table is used as a separator

Figure 3.5 shows working of section heading detection using structure of line. The tags at the end of each line indicate the result of structure detection. If the structure of line matches the criteria of section heading structure it is labeled as [candidate], if a line is rejected the tag indicate the reason of rejection.There can be cases where first part of line is section heading and the section starts on the same line after some white space.For such lines the first part of the line is checked and if it matches the criteria it is tagged as [first part candidate]. The figure shows only few of the rejection reasons however there may be more as explained earlier. Figure 3.6 shows that some lines

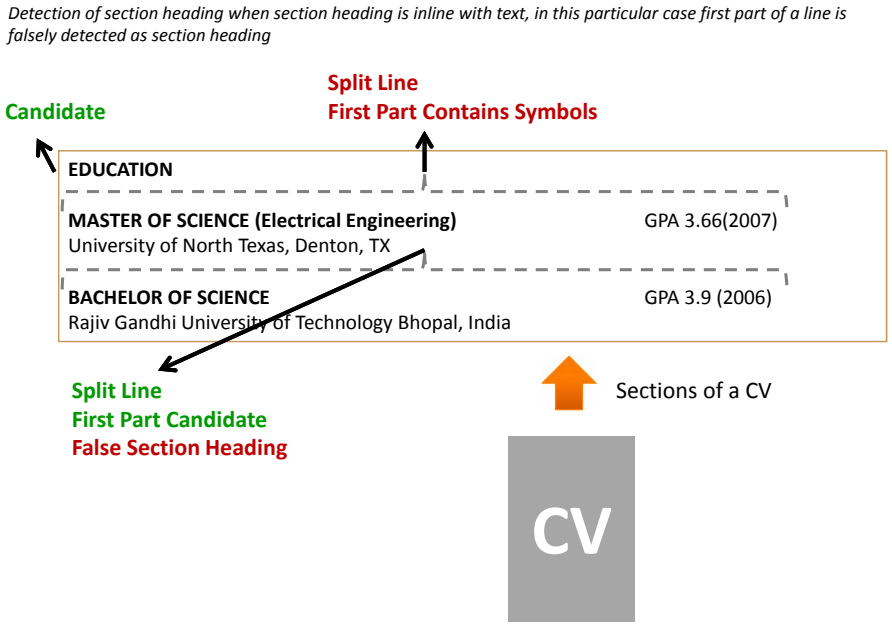


Figure 3.7: False heading detection example

which actually are not section heading are falsely taken as candidates. The problem here is that the first part of the line matches the criteria for being a candidate but the contents are actually not section heading, for this reason

the section heading structures alone cannot be trusted for segmentation,these KSI have their role, in next section we explain how this shortcoming is overcomed.

3.2.10 Named Entities

Named entities are general references to the names of predefined categories such as person, location, date, organization etc [26]. The concept was devised to develop a better understanding of a message and to develop better systems for Natual Language Processing [23] systems. Here we attempted to validate a candidate section heading based on the named entities.The lines detected as segment headings based on their structure may not be a section heading as any line of resume text can have such a structure. To verify it we used a two stage process. Firstly we determined the named entities present in a candidate section heading and checked type or category, later we checked the line against a list of acceptable named entities based on the results the line is accepted or rejected as section heading. To check the type or category of

MASTER OF SCIENCE (Electrical Engineering) GPA 3.66(2007)
University of North Texas, Denton, TX
Expected Graduation: August 2008
BACHELOR OF SCIENCE (Electrical Engineering) GPA 3.9 (2006)
Rajiv Gandhi University of Technology Bhopal, India

Text	**Entity Type**
Electrical Engineering	DegreeMajor
University of North Texas	Organization
August 2008	Date
2007	Date
Texas	Province
TX	Province
BACHELOR OF SCIENCE	Degree
MASTER OF SCIENCE	Degree
BACHELOR	Degree Title
Electrical Engineering	Skill
Denton	Person
University of North Texas	Organization
GPA 3.9	GPA
GPA 3.66	GPA
Denton	FirstPerson

Figure 3.8: Named entities in resume

the nouns such as organization names, places, degree titles etc we used Java

Annotation Pattern Engine grammer [9]. JAPE rules can easily be developed using GATE IDE [9], a default set of rules is also available in GATE. Using JAPE grammer type or category of various parts of english sentence can be identified to a great accuracy. Image 3.7 shows a sample of named entity detection Secondly we used a dictionary of section headings, the lines exactly matched in the dictionary are accepted rest are rejected. By this three phase process we achieve better detection of section headings. The entity types can

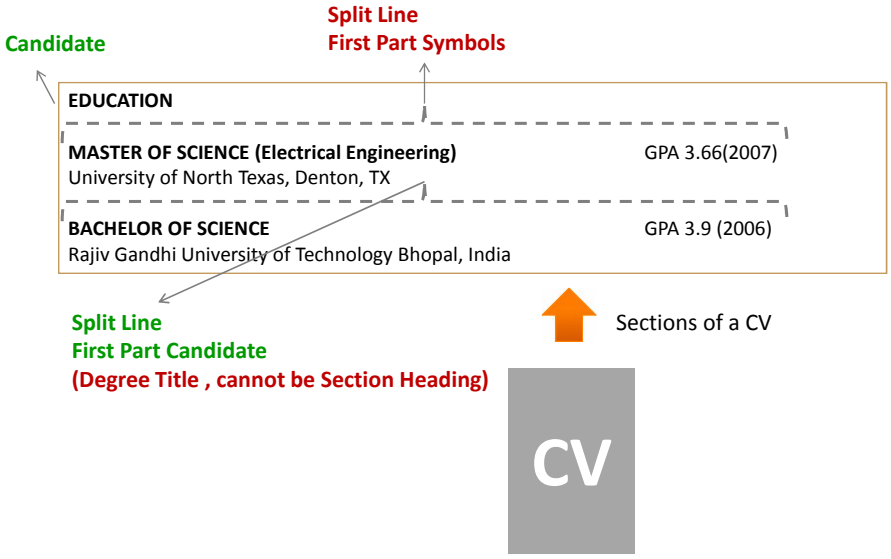


Figure 3.9: Avoiding false section headings using named entities

be checked against list of permissible entities in section headings, as shown in figure 3.8 the degree title which previously was taken as section heading is now corrected because a section heading do not starts with a degree title same is the case with nouns which are names of places, or occupations. After experimenting we developed a list of entities which do not occur in section headings. All candidate lines that have impermissible entities are rejected.

3.2.11 Segmentation Using Combined Approach

In this approach structure of each resume line is checked to see if it follows the structure of section headings,all positively matched lines are considered as candidate section heading.

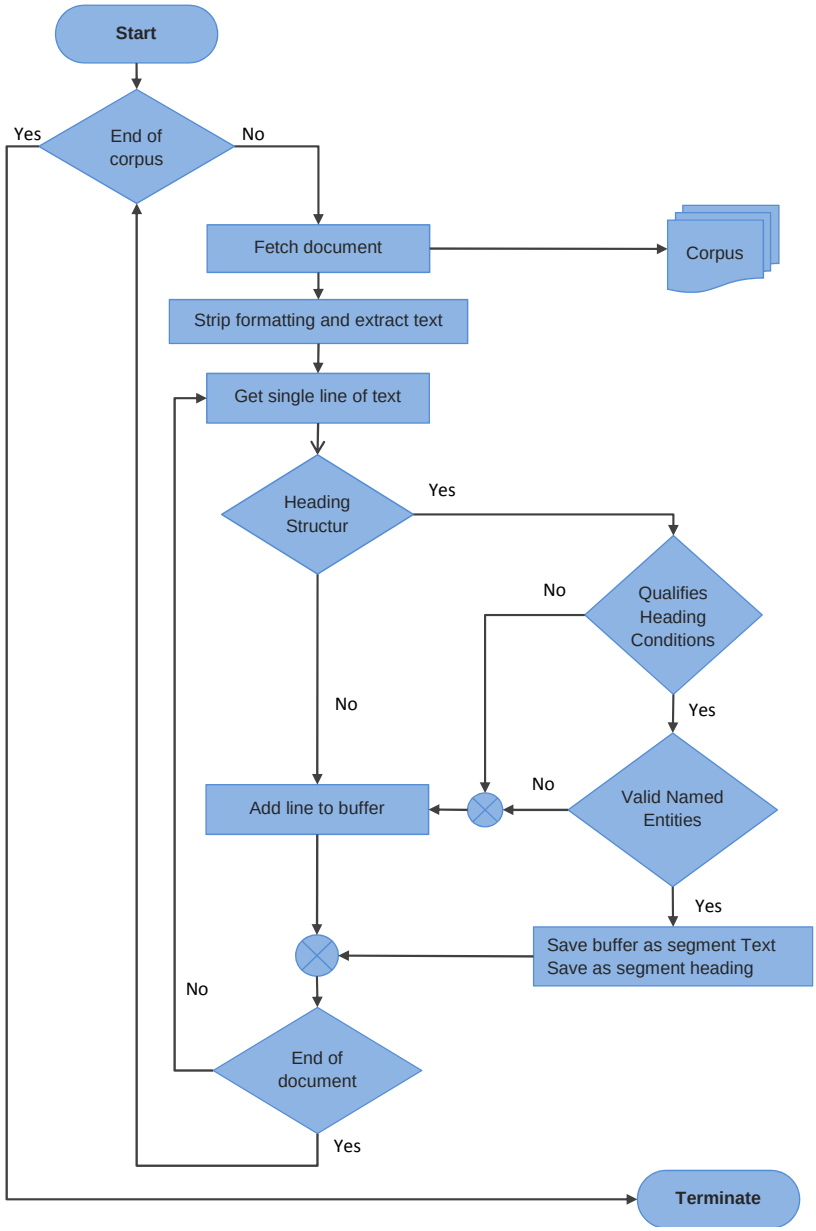


Figure 3.10: Segmentation Combined Approach

The candidate section headings are checked against a list of conditions

that a section heading should must conform. If a candidate section heading passes all of the conditions it is taken as section heading in case one or more conditions are failed the line is discarded from list of candidate section headings. Following are some of the conditions from the list of conditions defined for line to be a section heading. The conditions are prioritiez and have an order, when checked, if a condition is failed rest of the conditions are not checked.

- It is either a new line or a split line
- Should have word characters only
- Should not have numeric data
- Should not have symbols e.g.
- Should not have semicolons , commas , sign of exclamation
- Not all punctuations and articles are allowed
- Usually no verb is allowed

Figure 3.9 shows final execution of our section heading detection algorithm. All candidate lines are detected in first phase, in second and third phase lines are verified to be section headings, each line is labled which indicate strength of verification. All candidate section headings that were on a saperate line are labled as [Main Line],candidate lines that are first part of any line are labled as [Inline],if candidate line is not found in dictionary the lable is appended with [Orphan Heading] otherwise it is appended with [Section Heading], all candidates rejected are appended with [False Heading]. The candidate lines that passes all three phases eventually have lable [Mainline Section Heading], all lines that become candidate and are not rejected in phase 2 or 3 are considered as section headings.

3.3 Classification

Once the resume is segmented and zones are extracted, these zones are to be assigned a correct label that would indicate what type of information is contained in it. For classification text of each segment is converted to word vectors has been adopted here but extended further to meet the challenges of our problem and get improved results, following is an outline of our approach

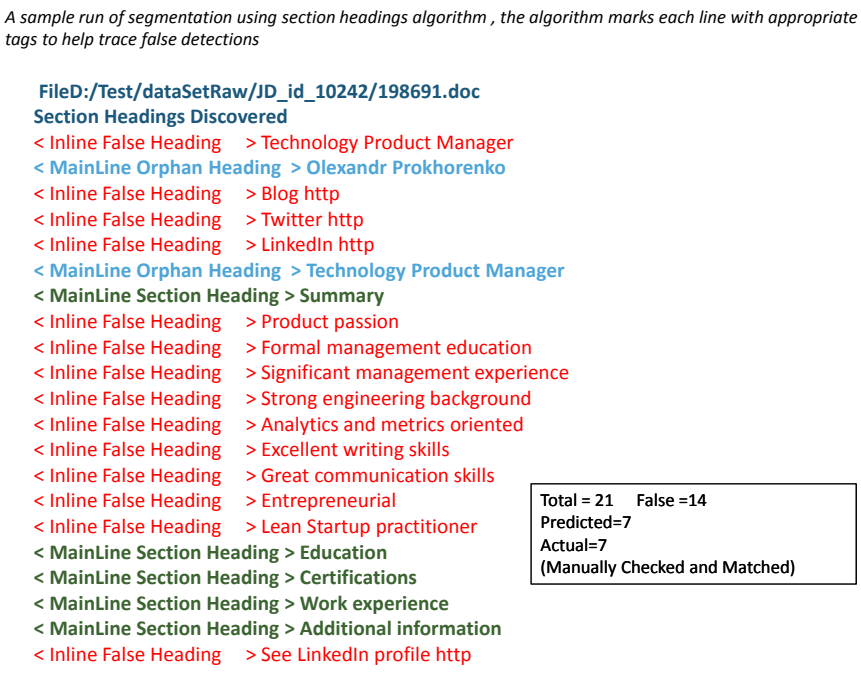


Figure 3.11: Section heading detection

- Text Preprocessing
- Text to word vector conversion
- Term Frequency/Inverse Document Frequency score calculation
- Classification with LibSvm [5]

3.3.1 Text Preprocessing

Preprocessing in our case was the removal of graphical formatting such as fonts, text styles and colors, tables, graphical lines , bullets etc. The words were converted to stems so that we get better word frequency count, in parallel any approach we developed for classification was also tested without doing text stemming or any replacement.

3.3.2 Handling Numeric Data

The very first problem we encountered was to identify the category or type of numerical data, such as salary , date of birth, job duration, number of publication or projects etc. This numerical data was though valid but was not very helpful in classification using Term Frequency / Inverse Document Frequency (TF/IDF) [18] . Numerical data belonging to one type with different value or style would be taken as a new term which misrepresents the TF/IDF values. Following measures were adopted to cope this problem

- Replace numeric data with appropriate sematic label such as a specific date value 12-10-2010 was replaced with named entity "date", numeric quantities such as number of publications or salary etc. were substituted with term "number" , amount etc.
- Some of the numeric data such as page numbers , numbered bullets were considered noise and were removed
- Alphanumeric data appearing mostly in addresses such as 10/A was left unchanged

3.3.3 Handling Nouns

The second problem was to deal with nouns which specifically name of an organization, person or place etc. Nouns poses same problem in calculation of TF/IDF score as were posed by numerical data. To deal with nouns the idea was to detect nouns, and place them with suitable lable that represent the category of these nouns. It was not possible to detect all nouns and perform a substitution however we did as much substitution as was possible using WordNet library [20]. Nouns were replaced with their type such as name of a country was substituted as "country". We experimented with nouns by replacing all types of nouns to only a few keeping in view the improvements in classification results. A second approach to deal with nouns was to replace them with named entities for that we used General Architecture for Text Engineering (GATE) tool and Java Annotation Pattern Engine (JAPE) grammar , not all but many of the nouns were substituted with their entity type.

3.3.4 Text to Word Vector Conversion

For text classification, word vector representation of text and Support Vector Machine are used this model is very well suited for text classification in

terms of efficiency high dimensionality and results [16].In this case data of a segment or zone was converted to a word vector using Weka library [10], a word vector is a representation of text according to Vector Space Model or Term Vector Model [18] where each term is represented by its weight, the term weight can be calculated in many ways , we used Term Frequency / Inverse Document Frequency here. Each zone / segment is taken as one single document, for classification each document was converted to word vector and passed to SVM classifier. The parameters of the classifier were adjusted after experimentation and final results were obtained.

Chapter 4

Testing and Evaluation

The techniques developed to identify and extract document zones in resumes were subjected to experiments to evaluate and compare their effectiveness. In this chapter we explain the experimental setup, data set, outcome of experiments, shortcomings identified, improvements made and the final results.

4.1 Data Set

Both classification and identificaion/segmentation of zones in resume required data set that is built from resumes segments. No such data set was available, and one has to be developed for this work. Following requirements are identified for the data set.

- The data set should consist of resume segments
- The segments should have been assigned the class lables.
- The nature of contents under a class label should be consistent
- The classes defined for segments should be present in majority of the resume
- The classes defined should cover all possible segments in the resume.

For classification, initially a small data set from 130 resume was developed consisting of 400 segments, 11 classes were defined for these segments. Later the data set was improved and a larger data set from 300 resumes was formed. The larger data set has around 1692 segments.However the number of instances were not same for different classes, not all classes were present in all resumes, some classes are more common than other such as class Work

for Work Experience is present in almost all of the resumes while class Pulbication is relatively rare. The unqual number of class instance is itself a problem. For this work we defined 17 classes that in our opnion represent all of the classes that can be present in a resume. The number of classes were defined after experimentation.Initially we defined a class for every zone or section of resume, that resulted in bad classification accuracy as some resume sections are overlaping and some contain contents that are similar in nature. We later grouped the similar classes and formed more generic classes. Experiments showed that that a moderate grouping of sections can produce good classification results alongwith flexibility of representing much better number of zones or resume sections.

Table 4.1: Classification Data Sets

Data Set	Resumes	Segments	Classes	Data Type
1	130	828	11	Plain Text
2	300	1692	18	Plain Text
3	300	1692	15	Plain Text

4.2 Evaluation Criteria

To evaluate the segmentation approaches resumes are processed one by one, the section headings are detected and compared with actual section headings present in the resume. The process could not be automated hence the comparisions are done manually. In all approaches the section headings are treated as boundary markers of zones or sections. Once the section headings are detected correctly the resume can be splitted into segments using these boundary markers. Two parameters precision and recall [11] are calculated for each approach. As there is no bench mark data set available, the results could not be compared also due to manual evaluation a small set of resumes is evaluated. However care is taken while selecting resumes, and it is attempted that different types of resumes are used for the evaluation.

4.3 Segmentation Results

The simple dictionary approach served as starting point, the problem with dictionary approach is that it detects only those segments for which a section heading is present in the dictionary. The detection process collapsed at the point where a section heading is not found in the dictionary, in such cases all

subsequent sections are detected as part of the last segment. The precision of dictionary approach much depends upon the completeness of the dictionary, as it is not possible to construct a dictionary that contain all possible section headings of resume the approach can not be generalized. Approach 1 is

Table 4.2: Segmentation Results

Approach	Resumes	Detected	Correct	Actual	Precision	Recall
1	30	247	144	165	0.63	0.86
2	30	196	149	165	0.81	0.91

detection of section headings using hurestics and dictionary, the high recall indicates that actual section headings are detected well however low precision shows that there is large number of false detection. As the approach detect section headings by their structure, all lines in the document that follow structure of section heading are detected, this include actual headings as well as sub-headings. After experimenting with named entites of section headings it is determined that certain named entites cannot be present in main section headings but can be present in sub-headings such as job titles, degree titles etc. Approach 2 uses named entites to mitigate the false detection, initially all candidate section headings are detected, later a list of named entities is consulted to discard false section headings. The approach worked well and both precision and recall were improved significantly.

4.4 Classification Results

Table 4.3: Classification Results-1

Class	Precision	Recall	class	precision	recall
Achievements	0.20	0.06	Patents	0.75	0.30
Activities	0.61	0.51	Projects	0.89	0.43
Associations	0.64	0.71	Publications	1.00	0.14
Awards	0.78	0.76	References	0.88	0.86
Basic Info	0.88	0.97	Related Courses	0.71	0.80
Education	0.89	0.97	Skills	0.87	0.88
Internships	0.00	0.00	Summary	0.70	0.56
Languages	0.78	0.92	Trainings	0.62	0.60
Objective	0.66	0.88	Work Experience	0.90	0.91
Weighted Average	0.809	0.819			

Initially we experimented with 18 classes and based on the results we modified the number of classes by merging similar classes.The results showed that some of the classes are very poorly classified, such as internship.It is discovered that some closely related classes are often misclassified to more abstract class. We examined the misclassification and merged the classes. The results were improved as shown below. The same results were improved slightly

Table 4.4: Classification Results-2

Class	Precision	Recall	class	precision	recall
Activities	0.581	0.478	Publications	1.00	0.286
Associations	0.691	0.792	References	0.923	0.857
Awards	0.692	0.574	Related Courses	0.80	0.821
Basic Info	0.955	0.982	Skills	0.875	0.883
Education	0.942	0.966	Summary	0.670	0.616
Languages	0.648	0.979	Trainings	0.640	0.604
Objective	0.758	0.818	Work Experience	0.931	0.917
Patents	1.0	0.50			
Weighted Average	0.849	0.849			

when the Term Frequency parameter was normalized that is weighted average of precision and recall changed to 0.850 and 0.853 respectively.Using wordnet API we attempted to assign generic type to the words representing names, numeric data, places, organization etc. We attempted replacement, concatination and conjustion of these words with their root words, classification results however did not improved much,see table .

Table 4.5: Classification Results-3

Class	Precision	Recall	class	precision	recall
Activities	0.537	0.477	Publications	0.897	0.833
Associations	0.555	0.520	References	0.714	0.641
Awards	0.622	0.558	Related Courses	0.811	0.796
Basic Info	0.851	0.950	Skills	0.875	0.883
Education	0.895	0.955	Summary	0.639	0.555
Languages	0.679	0.818	Trainings	0.487	0.377
Objective	0.714	0.50	Work Experience	0.914	0.889
Patents	0.4	0.142			
Weighted Average	0.78	0.79			

4.5 Results Summary

Table 4.6: Summary of Results

Approach	Precision	Recall
Segmentation Approach 1	0.63	0.86
Segmentation Approach 2	0.81	0.91
Classification 18 Classes	0.80	0.81
Classification 15 Classes	0.84	0.84
Classification 15 Classes Normalized TF	0.85	0.85
Classification 15 Classes with Wordnet	0.78	0.79

From the results table it is evident that the method of segmentation adopted here produced good results. Though the section headings do not follow any model still their structural information is very much effective in detecting segment boundaries. The classification results shows a slight improvement when number of classes were reduced, this indicates that a certain level of overlapping among data classes can be overcome using more abstract classes. It can be seen that classification involving Wordnet did not produce good result, there can be many reasons for this few of them can be incomplete generalization, reduction in uniqueness of test examples, or over all generalization itself. Over all the approaches developed demonstrated good results and can be improved further.

Chapter 5

Conclusion

In this research we explore the problem of zone identification and classification in resume documents. Not much prior work is available in this area, the techniques developed for resume documents assume presence of specific zones or zone order, we attempted to develop a generic approach for the same. We developed a novel approach to identify zones in resume using section headings and dictionary, alongwith an approach to classify the resume zones. We used section headings as boundary marker for resume section and treated each section as zone, we developed algorith to successfully identify section headings using structural details and named entities. For classification we developed a data set of resume zones that was not previoulsy available , we experimented classification of zones with the help of wordnet API and named entities. We discovered that section headings can be used to identify resume zones with high precision and recall.

5.1 Future Directions

The area of work is relatively new in text processing, lack of bench mark data set makes it difficult to compare and test the approaches developed. The segmentation approach presented in this work is simple, novel and effective. The approach can be further improved by developing more effective techniques for section heaading detection, also a richer dictionary of section heading can improve the performance and accuracy. The classification work can be improved further by using much larger, diverse and richer data set so that a good contrast among the classes can be available.

Bibliography

[1] Abd-Almageed, W., Agrawal, M., Seo, W., and Doermann, D. Document-zone classification using partial least squares. In *19th International Conference on Pattern Recognition* (2008), IEEE, pp. 1–4.

[2] Allan, J. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer international series on information retrieval. Springer US, 2002.

[3] Antonacopoulos, and Apostolos. Page segmentation using the description of the background. *Computer Vision and Image Understanding* (1998), 350–369.

[4] Baird, H. S. Background structure in document images. In *Advances in Structural and Syntactic Pattern Recongnition* (1992), World Scientific.

[5] Chang, C.-C., and Lin, C.-J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. 2*, 3 (May 2011), 27:1–27:27.

[6] Choi, F. Y. Y. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference* (2000), NAACL 2000, Association for Computational Linguistics, pp. 26–33.

[7] Daxtra.com. Intelligent recuritment solutions. http://www.daxtra.com//, Nov 2014.

[8] Egrabber.com. egrabber - automated list building, business lead generation, resume parser, email list completion, data capture software for b2b sales and recruiting. http://www.egrabber.com/, Nov 2014.

[9] Gate.ac.uk. Gate.ac.uk - index.html, 2015.

[10] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The weka data mining software: An update. *SIGKDD Explor. Newsl. 11*, 1 (Nov. 2009), 10–18.

[11] Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.

[12] Hearst, M. A., and Plaunt, C. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1993), SIGIR '93, ACM, pp. 59–68.

[13] HR, T. Textkernel - resume parsing (cv parsing), semantic searching and matching software. http://www.textkernel.com/, Nov 2014.

[14] Ibrahim, Z., Isa, D., Rajkumar, R., and Kendall, G. Document zone content classification for technical document images using artificial neural networks and support vector machines. In *Second International Conference on the Applications of Digital Information and Web Technologies. ICADIWT '09.* (2009), IEEE, pp. 345–350.

[15] Indeed.com. Job search — one search. all jobs. indeed.com. http://www.indeed.com, April 2015.

[16] Joachims, T. Text categorization with support vector machines: Learning with many relevant features, 1998.

[17] Keysers, D., Shafait, F., and Breuel, T. M. Document image zone classification-a simple high performance approach. In *2nd Int. Conf. on Computer Vision Theory and Applications* (2007), pp. 44–51.

[18] Manning, C., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[19] Manning, C. D., and Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[20] Miller, G. A. Wordnet: A lexical database for english. *Commun. ACM 38*, 11 (Nov. 1995), 39–41.

[21] Misra, H., Yvon, F., Jose, J. M., and Cappe, O. Text segmentation via topic modeling: An analytical study. In *Proc of the 18th ACM Conference on Information and Knowledge Management* (2009), CIKM '09, ACM, pp. 1553–1556.

[22] Moris, J., and Hirst, G. Lexzical cohesion computed by thesaural relations as an indicator of the structure of the text. *Computahtional Lingusitics* (1991), 21–48.

[23] Nlpu.com. Robert dilts nlp home page, 2015.

[24] Reynar, J. C. *Topic Segmentation: Algorithms and Applications*. PhD thesis, 1998.

[25] Salton, G., and Buckley, C. Automatic text structuring and retrieval-experiments in automatic encyclopedia searching. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1991), SIGIR '91, ACM, pp. 21–30.

[26] Sekine, S., and Ranchhod, E. *Named Entities: Recognition, Classification, and Use*. Benjamins current topics. John Benjamins Publishing Company, 2009.

[27] Sivaramakrishnan, R., Phillips, I., and Subramanium, J. S. Zone classification in a document using the method of feature vector generation. In *Proc of Third International Conference on Document Analysis and Recognition (ICDAR)* (1995), IEEE.

[28] Srivastava, S. *Career Counseling*. Atlantic, 2007.

[29] Utiyama, M., and Isahara, H. A statistical model for domain-independent text segmentation. 499–506.

[30] Wang, Y., Phillips, I. T., and Haralick, R. M. A method for document zone content classification. In *Proc of 16th International Confernece on Pattern Recognition* (2002), IEEE, pp. 196–199.

[31] Wang, Y., Phillips, I. T., and Haralick, R. M. A study on the document zone content classification problem. In *In Fifth IAPR International Workshop on Document Analysis Systems* (2002), pp. 212–223.

[32] Wang, Y., Phillips, I. T., M., R., and Haralick. Document zone content classification and its performance evaluation. *IEEE* (2006).