



TEXT ZONING IN A SEMI-STRUCTURED TEXT

FIELD OF THE INVENTION

The present invention relates to a system for text zoning in a semi-structured document.

- 5 More particularly, the present invention relates to a system and method to perform text zoning in a resume text or other recruitment related text for information extraction.

BACKGROUND ART

- 10 In recruitment activities, resume is still a preferred medium of information exchange for job seekers and HR managers alike, despite the common usage of structured web form. There are a number of reasons for that. For the job seekers, a resume is the only standard medium available, which they can use for applying job in many different corporations and recruitment firms. For the HR managers' end, a resume text is
15 preferred as it offers a flexible and expressive medium for the job applicant to express themselves. This is valuable for some aspect such as skill level, work experience, personality and strength, which are open-ended by nature and thus not possible to be described well through structured web forms.

- 20 Unfortunately, the inflow of a large volume of resume text in turn brings about a heavy burden on the employer or recruitment firm. Typically, a manual data entry process is needed to transfer the information from the resume text to the company's structured database system.

- 25 The common convention used in composing a resume is to divide the information into sections and zones, each describing a particular category of information such as education, personal particulars and work experience. Sub-sections or sub zones are used to describe separate events in a job seeker's education qualification, for instance different working stints in different companies and separate academic courses.

30

The composition of a resume text is only loosely guided by convention, not rigidly



confined to structural and formatting rules. It is neither a completely structured format suitable for populating a database nor a free text fully comprising grammatical sentences. Furthermore, extra attention should be given to the fact that a real-world resume text will be prepared by persons of disparate writing skills and styles. As such,
5 the semi-structured nature of a real-world resume text requires special treatment not currently provided by the information extraction systems designed for a completely structured or completely free form text.

Therefore, there is a need for a technology that is able to replace or at least reduce the
10 burden on the manual effort needed in the task of extracting specific information from a resume text for populating a job applicants database.

SUMMARY OF THE INVENTION

15 Accordingly the present invention provides a computer-assisted method for extracting domain-specific information from a semi-structured text comprising the steps of annotating the boundaries for the zones and sub zones in the semi-structured text as the training text; inducing a decision tree classifier for zoning and another for sub zoning by
20 running a decision tree induction algorithm on a pre-classified training text collection. Then the decision tree classifiers are utilized to segment the text in a new resume text into individual text zones and sub zones containing domain-specific information whereby the domain-specific information is able to be extracted thereafter by any text extraction algorithm.

25 The present invention consists of certain novel features and a combination of parts hereinafter fully described and illustrated in the accompanying drawings and particularly pointed out in the appended claims; it being understood that various changes in the details may be without departing from the scope of the invention or
30 sacrificing any of the advantages of the present invention.

BRIEF DESCRIPTION OF THE DRAWINGS

For the purpose of facilitating an understanding of the invention, there is illustrated in the accompanying drawings the preferred embodiments from an inspection of which
5 when considered in connection with the following description, the invention, its construction and operation and many of its advantages would be readily understood and appreciated.

FIG 1 is an overall view of the text zoning and sub zoning process.

10

FIG 2 is the two-step process of the text zoning and sub zoning process.

FIG 3 is the system training phase to create a decision tree classifier.

15 FIG 4 is a decision tree classifier for Text Zoning

FIG 5 is the main Text Zoning process.

FIG 6 is the classification process.

20

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention relates to a system for text zoning in a semi-structured document. More particularly, the present invention relates to a system and method to perform text
25 zoning in a resume text or other recruitment related text for information extraction. Hereinafter, the Text Zoning in a Semi-Structured Text shall be described according to the preferred embodiments of the present invention and by referring to the accompanying description and drawings. However, it is to be understood that limiting the description to the preferred embodiments of the invention and to the drawings is
30 merely to facilitate discussion of the present invention and it is envisioned that those skilled in the art may devise various modifications without departing from the scope of

the appended claim.

As mentioned earlier, a resume document used for job application is a semi-structured text. A semi-structured text contains information in both structured and free-text
5 formats. In a resume text, information is categorised into personal particulars, education and work experience zones. Each of these zones might contain further sub-zones, such as working stints in different companies or particular courses in education. Some form of pattern is used to separate zones and sub-zones from each other. However, due to the highly individual writing styles and skills of job applicants, the pattern can take many
10 different forms. Some may use a company name followed by a date range, some others may use a bullet symbol followed by a position title. To process and extract needed information, these zones and sub-zones are to be identified and segmented by the embodiment.

15 Zones like work experience and education is typically presented chronologically, with a particular segment (sub zone) allocated to a single event. The event can be a working stint in a particular company during a certain period of time. An event can also be an academic course undergone in an educational institution during a certain period of time. The heart of the invention is to treat the problem of extracting these zones and sub zones
20 as a classification problem. Remodeling the problem to a well-defined classification task allows us to treat the problem systematically by utilising a decision tree induction algorithm as a tool.

FIG 1 shows a computer based system deploying the invention. At step 101, resume is
25 inputted to the system in the form of a digital copy via an electronic network, such as the Internet.

Step 102, the zoning process, identifies the major zones in the resume. The zoning process in step 102 will be further elaborated using FIG 2, 3, 4, 5 and 6.

30

This zoning process produces a first-level segmented resume as shown in 103. Each of

the first-level zones are sub-zoned or segmented further in step 104. The sub zoning process in step 104 identifies sub-zones in specific zones like education and work experience zones. At step 105, a fully segmented resume is obtained as output. The sub zoning process in step 104 will be further elaborated using FIG 2, 3, 4, 5 and 6.

5

FIG 2 shows that the zoning/sub zoning process is a 2-step process. The training phase (201) of the system induces a decision tree classifier (202). The decision tree classifier is in turn used during the running phase to segment the resume (203). The System training phase (201) is further illustrated in FIG 3 while the zoning/sub zoning process (203) is illustrated in FIG 4, 5 and 6.

System training phase

In FIG 3, training text is received in step 301. The training text needs to be transformed into the form of normalised resume text with its lines pre-classified by a human expert.

15 First, normalised resume text is produced by a parser (302, 303). The parser parses the resume text and divides it into lines and elements (302).

Each line is represented in term of a feature set. The parser populates the feature set with appropriate values in step 303. The feature set at the minimum will consist of the following:

- 20 * The zone header present in the line
- * Features denoting whether the line is preceded or followed by empty line breaks.
- * Features denoting whether a company, date and position title and other significant elements are present in the line

25 A human expert will then assign one of the following classifications to each line in step 304 for zoning training set:

- * Work Experience Zone Boundary
- * Education Zone Boundary
- * Personal Particular Zone Boundary
- 30 * Reference Zone Boundary
- * Skill Zone Boundary

- *Other Zone Boundary
- * No Boundary

The following classifications will be assigned for creating the sub zoning training set:

- 5 * start of pattern.
- * continuation of pattern
- * end of pattern
- * anti-pattern

- 10 The classification information is essentially information on the patterns that serve as boundaries that separate the resume text into zones and sub zones. Thus, a training text is a real world resume text that is pre-classified in order to provide information on the pattern involved as boundary that separate the text into zones and sub zones whereby a classification is defined for each line in the training text.

15

After step 304, the training text set will be represented in terms of $\{l_1, l_2, l_3, \dots, l_n\}$ for a training text set having n lines. Each line l_i is in turn represented in a m-tuple $\{a_1, a_2, a_3 \dots a_{m-1}, c_i\}$ where $a_1, a_2, a_3 \dots a_{m-1}$ represents the $m-1$ features in the feature set together with their associated values as assigned by the parser while c_i represents the

20 classification assigned to l_i by the human expert.

The feature set is not fixed but instead would be determined by implementation consideration, subject to adjustment as appropriate. The parser will determine the appropriate value of each feature in the feature set for each line in the training text. The

25 result will be a table-like structure with each row representing a single line in the training set. The columns of the table will have a column allocated to each feature in the feature set, in which the value will be the appropriate value determined by the parser. There will be a final column for the classification for each line, in which the value will be filled by the information given in the pre-classified training text.

30

Step 305 takes the training text set representation produced in step 304 as input, and

induces a decision tree classifier by using a decision tree induction algorithm such as the ID3 (described in *Induction of Decision Tree*, Quilan, Machine Learning 1, 1995, pp81-106, Kluwer Academic Publishers).

5 FIG 4 shows the decision tree classifier induced during the training phase. For simplicity sake in aid of the explanation, the sample shown is only a fragment of the real decision tree classifier. The feature set used in this example consists of the following features:

- 10 * Current Zone Header {Work Experience, Education, Personal Particular, None and others}
 - A feature indicating the zone header occurs on this line. The value is 'None' if no zone header is present
- * Next 2 lines have position title {True, False}
 - A flag indicating if a position tile occurred in the next 2 lines
- 15 * Next 2 lines have education institute {True, False}
 - A flag indicating if an education institute occurred in the next 2 lines
- * Next 2 lines have date {True, False}
 - A flag indicating if a date occurred in the next 2 lines
- * Next 2 lines have qualification {True, False}
 - 20 - A flag indicating if a qualification occurred in the next 2 lines

Once the decision tree classifier is created, the actual zoning and sub zoning process is ready to be carried out.

25 **Zoning / Sub Zoning Process**

FIG 5 shows the zoning/sub zoning process. A test text is received in step 501. The test text is a real-world resume with no pre-classified information. It is the job of this process to classify each line in the test text in order to identify the relevant boundary type. The contents are as follows:

WORKING EXPERIENCE

Company name: Renesas E-learning Malaysia

Date Joined: 25 March 2002

Position: Senior Design Engineer

Specialisation: Education Courseware Design

Description:

EDUCATION

1999-2002

BSc (IT) Multimedia University

5

Step 502 and 503 are the parsing processes. The same parser that was used during the training phase is used here to produce a normalised representation of the resume test text, $\{l_1, l_2, l_3, \dots, l_n\}$ for a test text having n lines. Each line is in turn represented in terms of the m -tuple $\{a_1, a_2, a_3 \dots a_{m-1}, c_i\}$ where $a_1, a_2, a_3 \dots a_{m-1}$ represents the $m-1$ features in the feature set together with their associated values as assigned by the parser. c_i in the tuple will be unknown as of this stage.

10

In step 502, the test text will be divided into 11 lines:

WORKING EXPERIENCE	1
Company name: Renesas E-learning Malaysia	2
Date Joined: 25 March 2002	3
Position: Senior Design Engineer	4
Specialisation: Education Courseware Design	5
Description:	6
.....	7
.....	8
EDUCATION	9
1999-2002	10
BSc (IT) Multimedia University	11

In step 503, the parser populates the values of the feature set. As a result, we obtain the

5 following representation for the 11 lines in the test text:

Line	Current Zone Header	Next 2 lines have position title	Next 2 lines have education institute	Next 2 lines have date	Next 2 lines have qualification
1	WE	True	False	False	False
2	None	True	False	True	False
3	None	False	False	True	False
4	None	False	False	False	False
5	Education	False	False	False	False
6	None	False	False	False	False
7	None	False	False	False	False
8	None	False	False	True	False
9	Education	False	True	True	True
10	None	False	True	False	True
11	None	False	False	False	False

Table 1 Features of the Lines

At step 504, a pointer to the line collection of the test text is initialised. Step 505 initialises the pointer to 1. We are now about to classify our first line in the test text.

Step 506 classifies the current line, line 1 of the test text. Each line with its respective representation is supplied to the classifier as input. The classifier takes 1 line as input, classifies it, and moves on to the next line. The classification step 506 is further illustrated in FIG 6.

Classification process

The classification is guided by the decision tree classifier created during the training phase. The process starts by checking the root node of the decision tree. The root node tells the classifier which of the feature in the feature set is to be checked. The classifier checks the value for that feature of the current line. The value is used as a guide to traverse to the next child node in the decision tree. The same process of checking the feature denoted by the decision tree node, then checking the value of that feature for the current line and lastly use it as a guide to traverse to the next child node, is repeated in a loop until the traversal reaches a leaf node. A leaf node is a node with no child node. The leaf node will show the appropriate classification to be assigned to the current line.

In FIG 6, step 601 receives the current line, l_i of our test text. Step 602 sets the root node of the decision tree in FIG 4 as the current node. Step 603 tests if the current node is a leaf node. Since the root node of the decision tree is not a leaf node, we proceed to step 605 to retrieve the feature as indicated in the current node, the feature is '*Current Zone Header*'. Next, step 606 sets the variable $val(Current\ Zone\ Header)$ to the value of the feature '*Current Zone Header*' for the current line. Referring to table 1, it is '*WE*'. Therefore, $val(Current\ Zone\ Header) = 'WE'$.

Next, step 607 looks for the edge from the current node that match with $val(Current\ Zone\ Header) = 'WE'$. If the edge exists (608), the current node is set to the child node pointed by the edge selected (610). In our specific example, the edge matching the value '*WE*' from the root node does exist. So we move to the new current node, node 402 in

FIG 4.

Next, following the flow in FIG 6, we are back at the conditional test 603. The current node is still not a leaf node, so we move to step 605 and 606 to obtain *val(Next 2 lines have position title)=True*. The matching edge exists so we move to the new current node (step 607, 608, 610). The new current node will be node 404 in FIG 4.

When the process flows back to 603 this time, the current node is indeed a leaf node. So, step 604, assign the classification indicated in the leaf node to the current line.
 10 Therefore, the 1st line of our test text is assigned with the classification '*WE Zone Boundary*'.

At this point, the process flows back to FIG 5. Step 507 tests if the current pointer is at the last line of the collection. If not, the pointer is incremented by 1 (508) and step 506
 15 classifies the next line in the collection, as indicated by the pointer. The same process is repeated in the iteration until the pointer is equal to the last line in the collection (507). At that point, the process returns a classified test text (509). The result is:

Line	Current Zone Header	Next 2 lines have position title	Next 2 lines have education institute	Next 2 lines have date	Next 2 lines have qualification	Classification
1	WE	True	False	False	False	WE Zone Boundary
2	None	True	False	True	False	None
3	None	False	False	True	False	None
4	None	False	False	False	False	None
5	Education	False	False	False	False	None
6	None	False	False	False	False	None
7	None	False	False	False	False	None
8	None	False	False	True	False	None
9	Education	False	True	True	True	Education Zone Boundary
10	None	False	True	False	True	None
11	None	False	False	False	False	None

Table 2 Classified lines in a test text

A fully classified test text as shown in Table 2 would give information on the zone boundaries. Therefore, the problem of text zoning is thus accomplished as shown below.

5

WORKING EXPERIENCE Company name: Renesas E-learning Malaysia Date Joined: 25 March 2002 Position: Senior Design Engineer Specialisation: Education Courseware Design Description:
EDUCATION 1999-2002 BSc (IT) Multimedia University

The process will be continued to the sub zoning phase within the work experience zone and education zone. The process is similar to the zoning phase except that it is based on another decision tree classifier, one specifically used for sub zoning.

10

An example is provided below which will illustrate both the zoning and sub-zoning.

Personal Particulars
WORKING EXPERIENCE Company name: Renesas E-learning Malaysia Date Joined: 25 March 2003 Position: Senior Design Engineer Specialisation: Education Courseware Design Description:
Company name: Pentasoft (M) Sdn. Bhd Date Joined: 2 May 2002 Position: Trainee Design Engineer Description:
EDUCATION 1999-2002 BSc (IT) Multimedia University 1994-1999 High School, Maxwell

- 5 Once the zoning/sub zoning is completed, other automated information extraction methods can be used to process the segmented resume text to extract various information for populating the database of a recruitment firm or a employing company such as:

10

Position Title:	Senior Design Engineer
Company:	Renesas E-learning Malaysia
Date:	25 March 2002

- This invention allows the utilisation of a computer to automatically induce the heuristic rules needed to do the text zoning. The heuristic rules are represented in the form of a decision tree. The input needed is only a pre-classified training text set. This offers
- 5 significant advantage over hand coded heuristic rules because the computer-induced rules are more reliable. Furthermore, the cost of preparing a pre-classified training text set is significantly less than the cost of manually hand-coding and maintaining the rules for text zoning.
- 10 While in the foregoing specification this invention has been described in relation to certain preferred embodiments thereof and many details have been set forth for purpose of illustration, it will be apparent to those skilled in the art that the invention is susceptible to additional embodiments and that certain of the details described herein can be varied considerably without departing from the basic principles of the invention.

CLAIMS

1. A computer-assisted text zoning method for a semi-structured text comprising the steps of:

5 preparing a training set by manually annotating resume text with boundary information that separates the text into zones and sub zones
inducing a plurality of decision tree classifiers for zoning and sub zoning by
running a decision tree induction algorithm on a pre-classified training text set;
and

10 utilising the decision tree classifiers to segment text into individual text zones and sub zones containing domain-specific information;
whereby,
the domain-specific information is extracted thereafter from the text zones by any text extraction algorithm.

15

2. The text zoning method of claim 1, wherein preparation of a pre-classified training text set comprises:

running the training text set through a parser in order to divide the text into lines and elements;

20 populating appropriate values into a feature set that represent each of the line;
and assigning a classification to each of the line manually.

3. The text zoning method of claim 1, wherein segmenting the text comprises:

25 running the test text through a parser in order to divide the text into lines and elements;

populating appropriate values into the feature set that represent each of the line;
and classifying each line into one of the possible classes by referring to the decision tree classifier.

30

4. The text zoning method of claim 1, wherein classifying each line into one of the possible classes comprises the steps of:
starting at the root node of the decision tree as the current node;
obtain the value for the feature as indicated in the current node;
5 traverse to the next child node from the current node, pointed by the edge matching the value of the feature as indicated;
assigning a default class if no edge is found;
the same process being repeated until the current node is a leaf node,
whereby,
10 the classification indicated in the leaf node is assigned to the current line.
5. The text zoning method of claims 1 to 4, wherein each of the text sub-zones are nested inside a top-level text zone.
- 15 6. The text zoning method of claim 5, wherein the semi-structured text is a electronic resume text prepared by a job-seeker.
7. The text zoning method of claim 5, wherein the top level zones are working experience, education, personal particular, skill, reference, project and other
20 information categories found in a resume, and the sub-zones are chronological events or sub-sections contained within some of the said top-level zones such as education, work experience and reference.