Designing Rules for Accounting Transaction Identification based on Indonesian NLP

I Iswandi*, I S Suwardi and N U Maulidevi
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia

*irvan.iswandi10@gmail.com

Abstract. Recording accounting transactions carried out by the evidence of the transactions. It can be invoices, receipts, letters of intent, electricity bill, telephone bill, etc. In this paper, we proposed design of rules to identify the entities located on the sales invoice. There are some entities identified in a sales invoice, namely: invoice date, company name, invoice number, product id, product name, quantity and total price. Identification these entities using named entity recognition method. The entities generated from the rules used as a basis for automation process of data input into the accounting system.

1. Introduction
Accounting is an information and measurement system that aims to identify, record, and communicate relevant, reliable, and comparable information about business activities. It helps assess opportunities, products, investments, and social and community responsibilities [1]. The accounting cycle begins with the recording of financial transactions. The process of recording in accounting should only be done based on the evidence of the transaction, either in the form of invoices, receipts, letters of intent, electricity bill, telephone bill, etc. [2]. The goal of accounting is to provide useful information for decision making.

Invoice as one of the evidence of accounting transaction have been widely studied. An invoice is a special type of form that contains certain regularities in structure and content that aid in processing. Invoices also contain certain regions containing information that must be extracted from the document. And most of what is printed on an invoice is vital information that needs to be recovered [3]. Current versions of form readers which presume an apriori known geometric layout of the forms cannot solve the requested extraction task because invoices significantly vary in their layout structure. Hence, no unique form specification can be defined which matches all invoice types [4].
This paper presents an early stage that needs to be done to realize the artificial intelligence application in accounting domain. Stages initial question is how to design rules for tagging entities in accounting transactions. A variety of techniques have been proposed for automatic tagging. These different techniques can be classified into three major categories: rule-based, statistical-based and transformation-based approaches, which have been applied in many languages [5].

Named Entity Recognition (NER) is a very important tool in most applications of Natural Language Processing (NLP), including machine translation, question answering, information retrieval, information extraction, automatic summarization, and others [6]. And NER plays a very important role in the process of extracting information about people, locations and times of a description of the words [7].

In defining the work related to parse words, it is important to recognize the pieces of information such as name, include the person's name, organization and location names. And also the expression of figures includes the time, date, money and expressions percentage. Identification reference to these entities in a text recognized as one important part of the Information Retrieval and referred to as Named Entity Recognition and Classification (NERC) [8].

Previous research conducted by David A. Kosiba, identified the invoice using the method of combining textual and graphical processing by analyzing the intersection line and the line features in the document. In this paper, we proposed the rule design to identify invoice based on Indonesian Natural Language Processing. At the beginning of this paper, presented the background of the preliminary drafting of this paper, then in the second section presented related research. In the third section presented the process undertaken related to the early stages of the development of artificial intelligence-based accounting application is to draw up some rules to identify an entity in a transaction evidence. In fourth section presented the conclusions of this paper.

2. Related Research

Research relating to the identification of the invoice was also conducted by several researchers. David A. Kosiba et al conducted the research to identify the invoice using the method of combining textual and graphical processing by analysing the intersection line and the line features in the document as well as searching for possible keywords such as item number; quantity, total, etc. Valid keyword search regions are determined by a specialized connected-component analysis before any OCR is performed. The results of the keyword search and the line analysis are combined to give the search regions for extracting the relevant data contained in the invoice. Graphics component analysis concentrates on the detection of lines and their intersections. And text component analysis is a specialized connected-component analysis. But there are many other items that need to be addressed before calling the system complete. Because not all invoices contain such regular box structures, so the line intersection features may not be compatible with their method [9].

T.A. Bayer, et al, proposed a generic system for processing invoices. The system consists of two components, an OCR tool which need not be adapted to the current domain and an information extraction component FRESCO which contains the knowledge about the domain. A desired item from invoices can be extracted in two different ways. Extraction by syntax driven search in unconstrained document regions. And Extraction by first detecting the location the item is expected and then using this information for syntax driven extraction in this location [10].

An automatic invoice-documents classification system based on the analysis of the graphical information present in the document and able to perform both closed (the number of classes is fixed) and open world (the number of classes increases during operational life) classification presented by C. Alippi, et al [11]. When the normalised value is above the threshold the class of the logo is the one with the highest value in Vnum and the logo is moved to its proper class file folder. When the classification provides a 0 value the invoice is moved into an ambiguous folder for a subsequent inspection and manual classification. An ambiguous classification generally arises either when the logo to be classified is rather noisy and unclear or the invoice, for mistake, does not belong to the envisaged class.
3. Design Rules

Accounting is an information and measurement system that identifies, records, and communicates relevant, reliable, and comparable information about an organization’s business activities. Identifying business activities requires selecting transactions and events relevant to an organization. Examples are the sale of iPhones by Apple and the receipt of ticket money by TicketMaster. Recording business activities requires keeping a chronological log of transactions and events measured in dollars and classified and summarized in a useful format. Communicating business activities requires preparing accounting reports such as financial statements. It also requires analysing and interpreting such reports.

Recording accounting transactions carried out by the evidence of the transactions. It that can be invoices, receipts, letters of intent, electricity bill, telephone bill, etc. In this paper, we proposed design of rules to identify the label located on the sales invoice, as shown in Figure 1. In Figure 1 shown one example of a sales invoice that apply in Indonesia.

![Figure 1. Sales invoice that apply in Indonesia.](image)

After analysis some invoices, then there are some entities that can be identified in a sales invoice, namely:

- Invoice Date
- Customer Name
- Invoice Number
- Product ID
- Product Name
- Quantity of Product
- Total Price

Identification entities are required as a basis for accounting records in the form of accounting journals. Recording accounting transactions can be done automatically by the entities that have been identified.

Some rules defined to automatically identify the labels contained in a sales invoice. The rules are applied as an algorithm to determine the classification on a string. Below are some rules designed to identify the entities on a sales invoice:

a. Invoice Date

Rules are applied to determine the invoice date as follows:

- If string in contextual feature of month and string -1 is numeric and string +1 is numeric, then string -1 & string & string +1 is complete date
- If string content character ‘-‘ counted 2 (two), string before the first character ‘-‘ is numeric and string after the first character ‘-‘ until the second character ‘-‘ is in contextual feature of month and string after the second character ‘-‘ is numeric with length is four, then that string is date.
b. Customer Name
    Rules are applied to determine the customer name as follows:
    cont=NAMAPERUSH; morph=TitleCase|UpperCase|MixCase>ne=NAMAPERUSAHAAN

c. Invoice Number
    Rules are applied to determine the invoice number as follows:
    cont=IDFAKTUR; morph=TitleCase|UpperCase|MixCase>ne=NOMORFAKTUR

d. Product ID
    Rules are applied to determine the product id as follows:
    cont=KODEBRG; morph=TitleCase|UpperCase|MixCase>ne=KODEBARANG

e. Product Name
    Rules are applied to determine product name as follows:
    cont=NAMABRG; morph=TitleCase|UpperCase|MixCase>ne=NAMABARANG; ne+n=NAMABARANG

f. Quantity of product
    Rules are applied to determine quantity of product as follows:
    - Find the line of header of detail product in invoice
    - The line after header is the table of product
    - Find the lowest value in each line of product, by ignoring the numbers contained serial
      numbers and figures contained in the item name

g. Total Price
    Rules are applied to determine total price as follows:
    - Find the line of header of detail product in invoice
    - The line after header is the table of product
    - Find the highest value in each line of product, by ignoring the numbers contained serial
      numbers and figures contained in the item name

The entities generated from the rules can be used as a basis for making the automation process of data input into the accounting system.

4. Result and discussion
Some previous research that addresses associated with the identification of the invoice, there are several
methods used. Some of which identifies an invoice based on a combination of text and graphical, using
OCR analysis results are combined with the knowledge that is built and there are also studies that focus
on the analysis of the logo for the classification of invoice.
Of the few studies that have been done before, this study makes the draft a rule that can be used for
identification invoice Indonesian language. Rules designed referring to the text contained in the invoice.
There are some entities identified in a sales invoice, namely : invoice date, customer name, invoice
number, product id, product name, quantity of product, total price.
These rules require some contextual features as a basis for comparison text. Table 1 shows some of the
contextual features are used.

Table 1. List of contextual feature
Rule which has been designed can be used as a basis for building algorithms for identification invoice. The algorithm built into an intelligent accounting application.

5. Conclusions

Recording accounting transactions carried out by the evidence of the transactions. It can be invoices, receipts, letters of intent, electricity bill, telephone bill, etc. The rules has been designed to identify some entities in a sales invoice that apply in Indonesia. There are some entities identified in a sales invoice, namely: invoice date, customer name, invoice number, product id, product name, quantity of product, total price.

The rules which has been designed still has some drawbacks. Rules are only valid in the form of sales invoice in Indonesian language, but these rules can be developed so that it can apply to the other types of invoices.

The entities generated from the rules can be used as a basis for making the automation process of data input into the accounting system. This automation will be used in accounting intelligent system that will be the next research of this paper.

References

[1] John J W, Ken W S and Barbara C Fundamental accounting principles 20th edition
[2] Jerry J W, Paul D K and Donnalad E K Accounting principles 12th edition
[3] Delie M, Jian L and Jinwen T 2002 The design and implementation of a Chinese financial invoice recognition system Video/Image Processing and Multimedia Communications 4th EURASIP-IEEE Region 8 International Symposium on VIPromCom IEEE 79-82
[4] Cesarini F, Francesconi E, Gori M, Marinai S, Sheng J Q and Soda G 1997 Conceptual modelling for invoice document processing Database and Expert Systems Applications 1997 Proceedings Eighth International Workshop on IEEE 596-603
[5] Forsati R, Shamsfard M and Mojtahedpour P 2010 An efficient meta heuristic algorithm for post-tagging Computing in the Global Information Technology (ICCGI) 2010 Fifth International Multi-Conference on IEEE 93-98
[6] Ekbal A and Bandyopadhyay S 2009 Bengali named entity recognition using classifier combination Advances in Pattern Recognition, 2009. ICAPR’09. Seventh International Conference on IEEE 259-262
[7] Ali D A and Noah S A 2010 Semantically indexed and searched of digital images using lexical ontologies and named entity recognition Information Technology (ITSim) 2010 International Symposium in IEEE 3 1308-1314
[8] Nadeau D and Sekine S 2007 A survey of named entity recognition and classification Lingvisticae Investigationes 30 (1) 3-26
[9] Kosiba D A and Kasturi R 1996 Automatic invoice interpretation: invoice structure analysis Pattern Recognition 1996 Proceedings of the 13th International Conference on IEEE 3 721-725
[10] Bayer T A and Mogg S H U 1997 A generic system for processing invoices Document Analysis and Recognition 1997 Proceedings of the Fourth International Conference on IEEE 2 740-744
[11] Alippi C, Pessina F and Roveri M 2005 An adaptive system for automatic invoice-documents classification Image Processing 2005 ICIP 2005 IEEE International Conference on IEEE 2 II-526