Resume extraction with conditional random field method

Jason Anggakusuma¹, Viny Christanti Mawardi*², Manatap Dolok Lauro³
¹,²,³Informatics Engineering Department, Faculty of Information Technology, Universitas Tarumanagara

* viny@untar.ac.id

Abstract. Information extraction resume is a system used to carry out the automatic information extraction process on resumes of prospective employees to obtain key information including information on names, skills, educational experience, work experience, awards which are the main components in a curriculum vitae of prospective employees. Conditional Random Field (CRF) is one method that can be used to extract information. The step taken in the information extraction process is an annotation documents, tokenization, labeling on each token in the document, feature extraction, and the formation of models that will be used at the classification stage. In this study a system was built that could be used to extract CVs from users in pdf format. This study uses a CV from LinkedIn. Information extraction is done using 15 features and 11 classes. The evaluation of information extraction system for the employee's resume has a precision value of 85,052%, a recall value of 85,052%, and an f-measure of 78.527%.

1. Introduction
In the current era of technological development, almost all companies recruit employees through online-based sites. The online employee recruitment process is very useful for an effective employee selection process in cyberspace, and also candidate employees can send resumes to the company at the same time, without the need to come to every company that wants to apply. A company will select workers based on a company resume, which will retrieve important information on the resume such as educational background, applicant skills, work experience. Companies need to read one by one resume given by applicants.

Many portals that can be used online to find work and become a bridge of job seekers and companies, such as the online site LinkedIn.com that can collect 300 million personal resumes uploaded from each LinkedIn.com user [1]. The large number of data resumes, making it difficult to obtain useful information from each resume. In a resume there are also a few sentences and text that may be just additional information such as motto, education information in kindergarten or elementary school. Generally, there is information that is lacking or even unnecessary so that it can cause an abundance of information or commonly referred to as information overload [1]. It would be very easy if the results of resumes which are usually in the form of pdf can be directly changed in the form of a structured database so that the company can easily make selections, filters or analyzes of the advantages of an applicant.

The extraction process can be seen from the sentence "Tania Indrawan learns lots of things at IMAKTA", there is information on the name of the person "Tania Indrawan" and the organization "IMAKTA". So, the result of information extraction is important information contained in a text sentence. Generally, information extraction is a learning in Natural Language Processing which has a role to change information in the form of unstructured text into structured information [2]. The process
of extracting information from sentences can be done by several methods such as the Hidden Markov model, Conditional Random Fields (CRF) or other deep learning methods.

In this study a resume extraction information system was designed with the CRF method. The CRF method or algorithm was chosen in the design of employee resume data extraction application systems because the CRF method has many advantages from various aspects such as the CRF conditional nature which has more features and layers [3]. CRF is a method for constructing a probable model for labeling and sequential data sections [3]. The theoretical basic of the CRF method is to use conditional probabilities compared to the use of joint probabilities to distort patterns on labels based on patterns of observations. Jie Tang et al conducted research to find out the accuracy of the CRF (Conditional Random Fields) method for extracting a person's information from identification of the submitted biodata and having an accuracy of the data was 94% [4]. This research is only conducted on data in the form of biodata which usually has a simpler and more regular format than a resume.

In this study, we developed a resume data extraction application system that can be used to obtain information that is important to the HRD division automatically from employee resume text texts transforming employee resume information into tabular form. So that helps the selection of employees in the first stage quickly and does not require reading all the contents of a resume data text. This application can accept data input in the form of Indonesian resume in pdf by following the resume guide from LinkedIn. Then take information on names, skills, language, educational experience, work experience and awards. The results of this information will be displayed in tabular form.

2. INFORMATION EXTRACTION

Information extraction is a process of taking data needed for information retrieval, and the process of extracting information can be done with 2 approaches, namely based on rules or machine learning (machine learning). An employee data resume will contain a summary of each education, work experience, achievements, and capabilities of the employee itself. With the development of technology, employee data resumes are no longer sent using print media but rather, through cyberspace. In each resume that is commonly uploaded in cyberspace, generally a lot of information that is less important, not needed, and information that is not sought by workers seekers that can cause an abundance of information or referred to as information overload.

2.1. Named Entity Recognition (NER)

Named Entity Recognition (NER) is one feature extraction to identify or group each word into certain classes [5]. NER is used to check such as name, organization, location, and time. But as it develops, NER can check geographical locations, genes and so on. For example the sentence "Andy is the best chess player in Indonesia", and the label given is NAME, ORGANIZATION, TIME, LOCATION, AGE and O (excluding the given entity), then the result of the introduction of the entity from the given sentence is Andy / NAME is / O chess player / best O / O in / O Indonesia / LOCATION ". With this example, the results of the detection and recognition of entities are obtained, the word "Andy" as a name and the word "Indonesia" as a location. And the unknown word is O.

2.2. Conditional Random Field (CRF)

Conditional Random Field (CRF) is a framework for building probability models in labeling and segmenting data based on certain requirements [4]. To implement the CRF method, the Generalized Iterative Scaling (GIS) method, the algorithm of the GIS method chosen from IIS because when modeling the convergence slowdown occurs to find the value of features that help the process of normalization on average, thus chosen GIS methods that have a preference for maximizing log pick parameters must be chosen using some form of iterative technique, such as discussing the Maximum Entropy method. The difference only lies in the normalization formula used. Following is the GIS algorithm for the CRF method [6]:

\[ 
\alpha_j^{(n+1)} = \alpha_j^{(n)} + \frac{1}{\lambda} \log \left( \frac{\text{LP}_{j}}{\text{LP}_{j}^{(n)}} \right) \]  

(1)
Where $\alpha_{j}^{(n+1)}$ is feature weights to $-j$ for the iteration to $n + 1$, $\alpha_{j}^{(n)}$ is feature weights to $-j$ for iteration to $n$, $C$ is the sum of the feature values for all possible label and word pairs, and $E_{p_{f_j}}$: empirical value of the feature to $-j$.

If the weight value of the feature reaches convergent or has reached the specified number of iterations, then stop the iteration. If it hasn't reached one of the two conditions, then add the value of $n$ to 1 and repeat the second to fifth steps. After finishing calculating the weight value for each feature, the training phase is complete. In the testing phase, the test data will be documented, filtered and feature extraction is carried out. Then, the probability of label sequence will be calculated if the word order is known by the formula (2):

$$ p(t|C) = \frac{1}{Z(C)} \exp \left( \sum_{a}^{n} \alpha_{i} f_{i}(C, t) \right) $$

Where $p(t|C)$: the chance of label sequences if word order, $n$ is number of features, $t$ is order of labels, $C$ is word order, $\alpha_{i}$ is learning weights related to $f_{i}$, $f_{i}(C, t)$ is feature function to $-i$, and $Z(C)$ is normalization so that the number of opportunities is all possible sequence of labels for word order $C$ same with 1.

Equation (2) is actually the same formula in the MaxEnt method because both methods use the log-linear model. The formula above shows the current label, shows the previous label and shows the vector of the features used. For example, if the feature used is next word, then the feature vector is assumed to have the identity of the next word from the current position [8]. As with the Maximum Entropy method, equation (2) is calculated for all possible labels or classes for word order $x$. The result is a sequence of labels with the highest probability for word order $x$ or can be written in the form:

$$ t = \text{argmax} \left( \sum_{a}^{n} \alpha_{i} f_{i}(C, t) \right) $$

Normalization and exponent values are not used because they are the same for all label sequences.

2.3. Accuracy

Accuracy is a measurement with true value. The real value is the definition of a quantity or constant, the laws of geometry, and numbers obtained from a theory that has been agreed upon to be valid [8]. In making and designing a system, the process of evaluating the accuracy of the system is needed to determine the level of performance of the system.

The parameters used in general to measure the performance and accuracy of the system are:

1. Recall is the level of success of the system in finding back information.
2. Precision is the level of accuracy between the information requested by the user and the answers given by the system.
3. F-measure (F1) is the harmonic mean of recall and precision.

In the process of extracting information data, the evaluation process to get accuracy can be done by comparing the results of the system prediction class with the actual class results obtained manually. Then, a confusion matrix can be used to store the amount of data related to the actual class and the predicted class produced. From the results of the calculation of the confusion matrix, the calculation value of recall, precision, and f-measure for each target class. To get the value of recall, precision, and f-measure is to calculate the overall system average.

3. Formatting the text

The process of testing the prospective employee’s resume data extraction application program system using Conditional Random Field is done by testing the model to classify and extract the resume of prospective employee in the form of PDF / text. The test carried out is to do a comparison of 5 different types of data comparisons.
The process of testing the prospective employee's resume data extraction application program system using Conditional Random Field is done by testing the model to classify and extract the resume of prospective employee in the form of PDF / text. The test carried out is to do a comparison of 5 different types of data comparisons. Data used in the system using the Conditional Random Field method is in the form of pdf as text. The amount of description text of the prospective employee's resume used will be divided into 11 classifications (Figure 1). The classifications used in the form of PDF files and explanations are as follows: Contact (Contact), Skill, Language, Award, Name, Job name, Job type, Time of work, Name of education, Type of education, Time of education.

Fig 1. Sample Resume PDF file downloaded from LinkedIn.com

Output data that will be generated after extracting the extraction module contained in the form of results that have been processed by the model that has been made and has been classified with entities that have been determined by the program makers. The results are displayed according to the name of the entity that has been determined. Examples of extraction output results are in table 1.

| Table 1. Example of Results of Data Resume Extraction |
|---------------------------------|---------------|
| Label                  | Information                  |
| Contact                | www.linkedin.com/in/tania-indrawan-178865157 |
| Skill                  | Microsoft Word Public Speaking Easily Adaptable |
| Language               | English | Indonesian |
| Award                  | The Most Favorite Sophomore of the Year Beswan Djarum CPA Australia | |
|                        | Student Ambassador Junior of the Year | |
| Name                   | Tania Indrawan                  |
| Job Name               | Universitas Tarumanagara        | Ikatan Mahasiswa Akuntansi Tarumanagara |
| Job Type               | Internship Staff at Secretary of Accounting Program | Assistant Secretary of Accounting Program |
| Job Date               | January 2019-Present | January 2018-March 2018 |
| Edu Name               | Universitas Tarumanagara        | SMA Kemurman II Social |
| Edu Type               | Bachelor's degree | Science |
| Edu Date               | 2016-2020 | 2013-2016 |

What will be tested at the stage of comparison of different types of data is to determine the percentage and amount of test data and training data that will be used, the collection of data used is 350 data resume prospective employees. In this experiment, 88% training data (298 data) and 12% test data (39 data) are used, 80% training data (207 data) and 20% are used test data (39 data), 50% training data (40 data) and 50% test data (40 data) description in table 2.
Table 2. Amount of Training Data and Test Data for Each Data Comparison Type

| Percentage Training Data | Percentage Testing Data | Amount of Training Data | Amount of Test Data |
|--------------------------|-------------------------|-------------------------|--------------------|
| 50%                      | 50%                     | 40                      | 40                 |
| 80%                      | 20%                     | 207                     | 39                 |
| 88%                      | 12%                     | 298                     | 39                 |

Data collection of 350 data resumes will be labeled on each word that has been manually documented. The labeling process takes place, after the resume data has gone through the stages of tokenization and adjustment (removing noise in the data). The tokenization and adjustment process aim to label each token in the data according to the appropriate label. We can see the amount of tokenization can be seen at table 3.

Table 3. Amount of Tokenization Data

| Percentage Training Data | Amount of Training Data | Number of Training Data Tokens |
|--------------------------|-------------------------|-------------------------------|
| 50%                      | 40                      | 9327                          |
| 80%                      | 207                     | 48491                         |
| 88%                      | 298                     | 69476                         |

4. RESULT AND DISCUSSION
This application can accept data input in the form of Indonesian resume in pdf by following the resume guide from LinkedIn as we can see the Home menu at figure 2.

![Fig 2. Home Menu](image)

Then take information on names, skills, language, educational experience, work experience and awards. The results of this information will be displayed in tabular form as seen at figure 3.

![Fig 3. Result in Tabular Form](image)

We tested 350 data by dividing them into sections. The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections the formatting shown in table 2 should be used. The training model used in this test is based on the model with the highest evaluation results in the first test, namely the model that has been trained with 298 training data. Accuracy and evaluation results show at table 4 for datasets that have 88% training data, and show 94.96% for tokens that have correctly predicted at 88% training data, with the
percentage of precision at 94.96%, the percentage of recall at 93.95%, and the percentage of F-Measures at 94.45%.

Table 4. Accuracy and Evaluation Results for All Types of Testing Data

| Percentage Training Data | Precision | Recall | F-Measure |
|--------------------------|-----------|--------|-----------|
| 50%                      | 72.97%    | 81.52% | 77.01%    |
| 80%                      | 94.16%    | 92.90% | 93.53%    |
| 88%                      | 94.96%    | 93.95% | 94.45%    |

From a total of 350 data used, a process of sharing training data and test data with a certain percentage of 10 times resulted in 10 different test datasets being produced. Each dataset will be tested and averaged for its accuracy to obtain an average accuracy of the system. From a total of 350 data used, the process of distributing training data was 88% and the test data was randomly generated 12% 10 times to produce 10 different datasets. Table 5 is the average of the recall, precision, and f-measure values of 10 datasets.

Table 5. Accuracy, Recall, Precision, and F-Measure 10 Dataset results for random test data

| Dataset | Accuracy | Precision | Recall | F-Measure |
|---------|----------|-----------|--------|-----------|
| 1       | 66.667%  | 76.19%    | 84.21% | 80.00%    |
| 2       | 72.7274% | 88.89%    | 80.00% | 84.21%    |
| 3       | 71.05%   | 90.00%    | 77.14% | 83.08%    |
| 4       | 69.44%   | 78.13%    | 81.97% | 86.21%    |
| 5       | 64.00%   | 84.21%    | 72.73% | 78.05%    |
| 6       | 75.00%   | 100.00%   | 75.00% | 85.71%    |
| 7       | 72.00%   | 85.71%    | 81.82% | 83.72%    |
| 8       | 64.71%   | 81.48%    | 75.86% | 78.57%    |
| 9       | 83.33%   | 90.91%    | 90.91% | 90.91%    |
| 10      | 53.85%   | 75.00%    | 65.63% | 70.00%    |
| Average | 69.278%  | 85.052%   | 78.527%| 74.05%    |

The results of the accuracy and evaluation with the parameters recall, precision, and f-measure show that the results obtained are quite stable and there is only 1 dataset that has a value difference that is not too far away, namely in dataset 10 which has a difference of 10-11% accuracy with the second lowest dataset. From the evaluation results it can be seen that the model tested against 10 random test data produces an average value of accuracy of 69.278%, precision 0.85052, recall 0.85052, and f-measure 0.78527.

For testing by dividing 5 types of datasets, the best results obtained using test data are at a percentage of 88% of training data and 12% of test data with the highest recall, precision, and f-measure values compared to other comparisons. And for the average evaluation results of the 10 datasets can be used as accuracy of the tested model. By using the model obtained an average result with an accuracy of 69.278, a precision value of 85.052, a recall value of 85.052, and an f-measure of 78.527.

5. Conclusion

Based on the design, several conclusions can be drawn as follows:
1. This system can be a reference for job seekers in a company to choose employees quickly.
2. This system can extract and classify any information that has been determined.
3. In making the application system, the Conditional Random Field (CRF) method has an accuracy of 69.278, a precision value of 85.052, a recall value of 85.052, and a f-measure of 78.527 in the process of extracting information on the prospective employee’s resume data.
6. References

[1] Chen, Jie, Chunxia Zhang, and Zhendong Niu. "A two-step resume information extraction algorithm." Mathematical Problems in Engineering (2018).

[2] Jurafsky, Daniel, and James H. Martin. "Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing." Upper Saddle River, NJ: Prentice Hall (2008).

[3] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[4] Tang, Jie, Mingcai Hong, Duo Liang Zhang, and Juanzi Li. "Information extraction: Methodologies and applications." In Emerging Technologies of Text Mining: Techniques and Applications, pp. 1-33. IGI Global, 2008.

[5] Permana, Hadi. "Named Entity Recognition Menggunakan Metode Bidirectional Lstm-Crf Pada Teks Bahasa Indonesia." PhD diss., Universitas Komputer Indonesia, 2019.

[6] Curran, James R., and Stephen Clark. "Language independent NER using a maximum entropy tagger." In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, pp. 164-167. 2003.

[7] Curran, James R., and Stephen Clark. "Investigating GIS and smoothing for maximum entropy taggers." In 10th Conference of the European Chapter of the Association for Computational Linguistics. 2003.

[8] Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields for relational learning." Introduction to statistical relational learning 2 (2006): 93-128.