JIRAE, Vol. 5, No. 1, April 2020, 19–22
e-ISSN 2407-7259
DOI: 10.9744/JIRAE.5.1.19-22

Improving Data Extraction System to Parse Data from Scraped Job Advertisements

Claudia Nathasia Jason[1,a]

[1]Department of Informatics, Petra Christian University, Surabaya, Indonesia
Information and Communication Technology, Fontys University of Applied Sciences Eindhoven, The Netherlands
[a]claudianathasiaj@gmail.com

Abstract. Extracting the information from an online job advertisement might be a little tricky. The information is wrapped with redundant information, called boilerplate, that is not related to the job at all. The information also needs to be segmented and classified into the right class or group. After the information has been classified, it is easier to find the features (e.g., required skills and required education) that make the later processing faster.

Keywords: Data parsing, job advertisement, data segmentation, data classification, feature extraction.

1. Introduction

The Internet provides so much information, including job advertisements. Job advertisements are updated almost every day and it is almost impossible to keep track of every single job advertisement that has been uploaded. The problem is job seekers will spend more time searching for the ideal job by gathering so much information. This process can be cut down by parsing the job advertisement and help people to find their ideal job by matching the features of the job and the skills of the job seekers. This is where feature extraction is needed. A job advertisement on the internet is wrapped with redundant information that needs to be removed before the data is matched to the job seeker's profile.

The job advertisement can be processed using Natural Language Processing (NLP). NLP is a branch of artificial intelligence to understand and make sense of the human language. NLP is hard for computers because human language is hard to understand. Human language has many rules with different levels and difficulties. The easy rule example is where a verb is changed by looking at whether the subject is singular or plural. The hard rule example is when people use sarcasm as the meaning changes. The after-result of job advertisement processing later will be used for people and job matching.

The research done is to find the best method to clean the job advertisement as most of the job advertisement website has unrelated advertisement inside the pages.

2. Research

The research is done by using the Development Oriented Triangulation or DOT framework that has been provided by Fontys University. The Framework focuses on what, why, and how of the research. This research focuses more on the product and is not to introduce new knowledge or theory. Furthermore, this research is focusing on finding out the best technique to resolve the problem.

There are 5 strategies in the DOT framework namely, library, field, lab, workshop, and showroom. For this research, only library, field, lab, and workshop are used. The library strategy covers reading and report writing. The field strategy covers the knowledge gain from experts. The lab and workshop strategies cover the implementation of the program.

2.1. Data Parsing steps

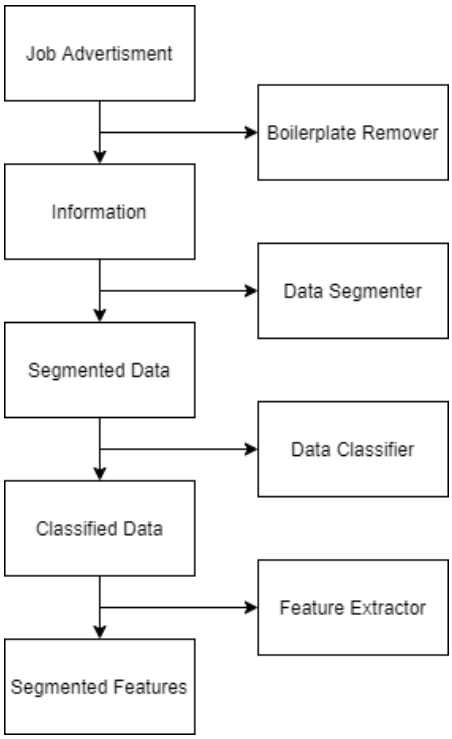

Figure 1. Process of Data Parsing

The data parsing has 4 steps that can be used as a pipeline or be done individually. The steps are boilerplate removal, data segmentation, data classification, and feature extraction. Each step is implemented in Python 3 and using the factory design pattern.

2.2. Datasets

The data were collected to measure the quality of the system. Each step has its datasets as the measure for each step has different desired output. Boilerplate removal and data segmentation datasets are in Dutch. Data classification and feature extraction are in English. Each step is language independent if the training data is available for any language.

3. Result and Discussion

3.1. Boilerplate Removal

Taken from the dictionary of Meriam-Webster, boilerplate is a standardized text. The definition is almost the same in programming. The boilerplate code is a code that can be used repeatedly with small or without any change(s). The parts that are considered as boilerplate for this project are all the parts that are not related to the job. These parts are considered boilerplate because it holds a certain structure that might be similar to other websites.

There are a couple of methods and libraries to remove the boilerplate from the collected data. Every single one of them has different strengths and weaknesses. Take the example of Justext library removal methods. Justext takes all the visible texts on the websites and then removes all the texts that are shorter than ten words. This works on news websites because most of the time, the main information is longer than ten words, but it doesn't work in the job advertisement. Job advertisements sometimes have really short information such as email, phone number, and job location. This relevant information can't be left out.

The other way to remove the boilerplate is the structure recognition method. This method is used by most libraries. Beautiful soup library is using this method to get only the main content text, but it doesn't remove the advertisement that is showed to recommend job seekers.

Instead of making a new structure recognition library, we decided to remove the advertisement after beautiful soup did its job. This was done to prevent inaccurate results for the matching system due to misleading information from another job advertisement inside the data. Most of the job advertisements are separated by a section.

3.1.1.

The dataset for boilerplate removal quality measurements was collected. It contains 102 HTML pages from 34 different websites. The collected data were job advertisements in Dutch.

Table 1. Boilerplate Removal Score

Average Value(%)	**65.45**
Highest Value(%)	76.84
Smallest Value(%)	46.43

Table 1 shows the score of beautiful soup library. The result from the boilerplate removal is compared to the data that has been cleaned manually by the researcher and validated manually to check the accuracy for Beautiful Soup Library.

3.2. Data Segmentations

In this step, the long text is separated into two or more segments or groups that have the same context. Segmentation is a process of dividing a text into a group of words where it has a similarity based on the chosen parameters. Segmentation is done in the preprocessing phase because each job advertisement has different topics and segments that explain the details of the job. The segmentation will make the data classification easier because the topic will be more specific.

There are a lot of methods for text segmentation, but for this project, we chose to try the *texttiling* methods and *whitespace* segmentation methods after short research and discussion with the experts within the company. The methods were chosen due to their speed and accuracy.

3.2.1. Texttiling

The *texttiling* method can be found within the Natural Language Toolkit Package or more known as NLTK. The toolkit provides processing libraries for classification, tokenization, stemming, and parsing in Python. This toolkit is widely used in the text processing field. *Texttiling* works by calculating similar words inside one group of words called pseudo-sentence. The pseudo-sentence is separated into a fixed size. and does not follow the regular punctuation. *Texttiling* is using pseudo-sentence to prevent a big score of differences between long and short sentences that might happen if a regular punctuation sentence is used. The words are converted into tokens for easier calculation. The calculation is known more as the scoring method.

There are two types of scoring methods that are given by the NLTK, block comparison, and vocabulary introduction. Blocks on *texttiling* are a group of pseudo-sentences. For the block comparison method, the calculation is done by looking at how many words are similar or common. This is based on a belief that a segment might have a lot of common words as it is still in one topic. On the other hand, the vocabulary introduction method calculates how many new words are introduced in the other blocks.

3.2.2. Whitespace Segmentation

Whitespace segmentation is a method to separate text by space or a new line between text. The method is working well for the segmenting paragraph, as it is common to separate the text with a new line. It is also a common structure for job advertisement as it separates between job introduction, description, and requirements. This method is good for separating the text into short segments, but for this project, long segments are better for the next step.

3.2.3. Quality Measure

The dataset for data segmentation quality measurements was collected. The dataset contains 102 HTML pages from 34 different websites. The collected data are job advertisements in Dutch.

Table 2. Methods comparison

Criteria	Whitespace Segmentation	TextTiling
Average Time	0.067 seconds	0.28 seconds
The final result	A lot of segments	Segments divided by Paragraph
Functionality	• Language-Independent • Results in many short segments	• Language-Independent by using different language stopwords. • Results in long segments
Maintainability	Easy to maintain	The toolkit is still getting updated

Table 2 shows us the comparison of the methods. The methods are compared to their results and not on the Pk Error Metrics, where the probability of two sentences is identified whether it belonged in the same document or not. *Texttiling* was selected as the long segments are better for the system. The chosen method has been considered by looking forward to the classification step as the longer text will give a better classification class.

Table 3. Data Segmentation Score

Average Score (%)	82.54
Highest Score (%)	86.07
Smallest Score (%)	43.58

Table 3 shows us the data segmentation using the *texttiling* method. The upgraded algorithm is the use of the *texttiling* method combined with the data segmentation. The score is generated using the comparison from the result and the data that has been cleaned manually.

3.3. Data Classification

In this phase, the data are labeled into several classes. The classes that are used in this project are *introduction, description, benefits, employer description, and requirements.* Classification is a process of labeling data according to context. There is a lot of data classification methods that can be used for this function. After a short research, the best algorithms are Support Vector Machine (SVM) and Neural Network (NN).

3.3.1. Support Vector Machine

Support Vector Machine, or known as SVM, is an algorithm that classifies data by picking the best decision boundary from a given group or category by looking at the vectors. To do the SVM classification, first, the words need to be stemmed. The stemming process is a method to reduce a word to its stem (e.g., *ophalen* and *ophaal* are stemmed as *ophal*). After stemming is done, the text is converted to vectors. The vectors will be set into coordinates and will be used to calculate the boundaries that exist within the data. When the boundaries have been developed, then they can be used to classify the text.

In this research, linear SVM is used with a maximum of 1000 passes over the training data. The number was selected because it gave a better result than a higher or smaller number of passes.

3.3.2. Neural Network

Neural Network is the base of every deep learning method. This method is inspired by how the human brain works. The neural network used for this project is the feedforward neural network. The neural network method uses the labeled data provided and 'learn' from there. Later, the solution will come up with the best model and solution without any needs to specify the parameters manually.

The information flows through two ways in a neural network, the feedforward network and backpropagation. The feedforward network represents the condition where the network is learning. The feedforward network is where the network receives the information and triggers the next layer inside the network. Each layer receives the input from the previous layer and does the calculation inside the perceptron.

After learning, the network will get feedback. This process is called backpropagation. The feedback is done by comparing the output with the output that it should produce and using the difference to modify the weights. The information is propagated to the left side.

For the research, a single-layer neural network is used and a maximum of 10000 passes over the training data. This number was selected due to limited hardware resources and the relatively small number of data.

3.3.3. Quality Measures

The dataset for data classification quality measurements was given by the company. There are five classes with around 8000 data in each class that are used for data training. The data is in Dutch.

Table 4. F1-Score between the two methods

Class	SVM	Neural Network
Introduction	0.798	0.528
Description	0.883	0.683
Benefits	0.879	0.629
Employer Description	0.864	0.843
Requirement	0.859	0.546
Average F1-Score	**0.857**	0.646

Table 4 shows us the comparison for the data classification methods. F1-Score is calculating the harmonic mean of the precision and the recall for the test. Precision is looking at how many selected items are relevant and recall is to see how many relevant items are selected. For data classification, SVM works better as it scores higher than neural network in classifying the data.

3.4. Feature Extraction

Feature extraction is a process of determining a subset of features to be used in matching. The features contain the relevant information from the input data like skills, education, etc. The result of feature extraction will make the data processing faster since the data size will be reduced after the feature extraction and the quality of the matching will improve by extracting better features.

Data extraction was performed by determining the classes for each word. This method is inspired by Named entity recognition that locates and classifies named entities in text into pre-defined categories. To do the feature extraction, first, the data were trained with Part-Of-Speech tagging or POS tagging. The training was done by NLTK and Scikit-Learn Tool Library. The POS tagging helps to label the training data's words into categories such as nouns, verbs, adjectives, etc. For this project, the categories are educations, skills, functions, and industries.

After the data has been trained, the data will generate labeled data that are used to categorize the input data. The next step is to label each word from the input data by using the labeled data. After each word was labeled, the feature was extracted by picking the label on each text.

3.4.1. Quality Measures

The dataset for feature extraction quality measurements was collected. It contains 102 HTML pages from 34 different websites. The collected data are job advertisements in English.

Table 5. Feature Extraction Score

F1-Score	**0.763**

Table 5 shows the test result for the feature extraction using named entity recognition. This solution for feature extraction is not recommended because the solution can not recognize some features in the class. This was confirmed by checking on the result manually by the researcher. This could be improved by testing another method, such as NLP Pipeline, and try to collaborate it with the named entity recognition.

4. Conclusion

There are 4 tasks in data preprocessing: boilerplate removal, data segmentation, data classification, and feature extraction. Each task has its own best method and is combined into one prototype. Good results are obtained after the research and several tests. The boilerplate removal can be done by using the linguistic approach. It can be improved by looking into the structure of the boilerplate and later removing the irrelevant parts. The data segmentation can be done by using the *texttiling* library. The library provides a good number of segments. The data classification can be done by the Support Vector Machine (SVM). The SVM provides a good classification and a better result than neural network method that is investigated at the same time. The feature extraction can be done by the Named Entity Recognition. The method gives a good result based on the quality measure. The solution for the problem is by building the data parsing and extraction system based on the researched method

References

[1] Endrédy, I., & Novák, A. (2013). *More Effective Boilerplate Removal - the GoldMiner Algorithm*. Polibits, 48(ISSN 1870-9044), 79-83. doi: 10.17562/pb-48-10

[2] Garbade, M., 2018. *A Simple Introduction to Natural Language Processing*. [online] Medium. Available at: <https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32> [Accessed 20 July 2019].

[3] Hearst, M. (1997). *TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages*. Computational Linguistics, 23(1), 33-64. Retrieved from https://www.aclweb.org/anthology/J97-1003.pdf

[4] Nguyen, D. (2016). Using BeautifulSoup to parse HTML and extract press briefings URLs. [online] Computational Journalism, Spring 2016. Available at: http://www.compjour.org/warmups/govt-text-releases/intro-to-bs4-lxml-parsing-wh-press-briefings/ [Accessed 20 Jul. 2019].

[5] Patel, S. (2017). *Chapter 2: SVM (Support Vector Machine)—Theory*. [online] Medium. Available at: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 [Accessed 2 Jun. 2019].

[6] Pevzner, L., & Hearst, M. (2002). A Critique and Improvement of an Evaluation Metric for Text Segmentation. Computational Linguistics, 28(1), 19-36. Retrieved from https://www.mitpressjournals.org/doi/pdf/10.1162/089120102317341756

[7] Saveri, M. (2018). *What is boilerplate and why do we use it? Necessity of coding style guide*. [online] freeCode Camp.org News. Available at: https://www.freecodecamp.org/news/whats-boilerplate-and-why-do-we-use-it-let-s-check-out-the-coding-style-guide-ac2b6c814ee7/ [Accessed 20 Jul. 2019].

[8] Zhou, V. (2019). *Machine Learning for Beginners: An Introduction to Neural Networks*. [online] Victorzhou.com. Available at: https://victorzhou.com/blog/intro-to-neural-networks/ [Accessed 20 Jul. 2019].