



Arab American University – Jenin

**Faculty of Graduate Studies**

**SEMANTICS-BASED ONLINE RECRUITMENT SYSTEM**

By

**Aseel Bassam Mahmoud Kmail**

Supervisor

**Dr. Mohammed A. M. Maree**

**This thesis was submitted  
in partial fulfillment of the requirements for  
the Master's degree in  
Computer Science  
February / 2016**

**© Arab American University – Jenin 2016. All rights reserved.**

## **Semantics-based Online Recruitment System**

By

**Aseel Bassam Kmail**

This thesis was defended successfully on 27/2/2016 and approved by

Committee members

1. Dr. Mohammed Maree

2. Dr. Ahmad Ewais

3. Dr. Nidal Al-kafri

Signature



Three handwritten signatures in blue ink, each on a dotted line.

## **Declaration**

This is to declare that the thesis entitled "Semantics-Based Online Recruitment System" under the supervision of Dr. Mohammed Maree is my own work and does not contain any unacknowledged work or material previously published or written by another person, except where due reference is made in the text of the document.

Date: 20/2/2016

Name: Aseel Bassam Kmail

Signature:

## **Acknowledgements**

First and foremost I would like to thank my husband Mutasem and my son Mohammed.

Their help, support, patience, and understanding made this endeavor possible.

Also, it has been my endowment to work closely with Dr. Mohammed Maree, I have enjoyed the opportunity to learn from his knowledge and experience. He has been providing encouragement, assistance and support all the time.

Huge thanks, love and appreciation go to my brothers and sisters: Mahmoud, Lama, Lana, Mohammed and Amal for their support and belief in me and my work.

I wish to thank all of my friends who have supported me over the last few years: Waed Jabareen, Mervat Zuabi, Aya Kmail, Ahlam Hasan and Yusra Abu-Alrub.

Last but not least, and most importantly, I'd like to thank my parents, Eng. Bassam Kmail and Mrs. Seham Abu-Alrub for their unconditional love, help and support.

To all of you I dedicate this thesis.

## **Dedication**

I dedicate this thesis to my beloved father

**Eng. Bassam Kmail**

And

To my mother

**Seham Abu Al-rob**

You have successfully made me the person I am becoming

You will always be remembered

## **Abstract**

Due to the rapid development of job markets, conventional recruitment methods are becoming insufficient. This is because employers often receive an enormous number of applications that are difficult to process and analyze manually. To address this issue, several automatic recruitment systems have been proposed. Although these systems have proved to be more effective in processing candidate resumes (i.e. CVs) and matching them to their relevant job offers, they still suffer from low precision due to limitations associated with their underlying techniques. On the one hand, traditional keyword-based matching approaches ignore the hidden semantic dimensions in the content of job offers and resumes; and consequently a large portion of the produced results is irrelevant. On the other hand, the newer semantics-based approaches are penalized by limitations of the exploited semantic resources, namely semantic knowledge incompleteness and limited domain coverage. In this thesis, we propose an automatic online recruitment system that employs multiple semantic resources to discover the otherwise-unseen semantic dimensions of resumes and job posts. Additionally, the proposed system utilizes statistical-based concept-relatedness measures to further enrich the processed content with relevant concepts that were not initially recognized by the used semantic resources. The proposed system has been instantiated and tested using real-world data gathered from various recruitment resources on the WWW. Furthermore, the effectiveness of the employed techniques has been validated in a precision-recall based empirical framework and the results were published in two well-recognized international conferences in 2015.

## Table of Contents

<b>Declaration .....</b>	<b>ii</b>
<b>Acknowledgements .....</b>	<b>iii</b>
<b>Dedication.....</b>	<b>iv</b>
<b>Abstract .....</b>	<b>v</b>
<b>List of Tables.....</b>	<b>ix</b>
<b>List of Figures .....</b>	<b>x</b>
<b>List of Abbreviations.....</b>	<b>xi</b>
<b>1. Introduction .....</b>	<b>1</b>
1.1 Background and Motivations .....	3
1.2 Problem Statement and Research Questions .....	5
1.3 Research Methodology .....	7
1.3.1 Pre-processing of Unstructured Resumes/Job offers .....	7
1.3.2 System Evaluation .....	9
1.4 Contributions .....	10
1.5 Publications .....	11
1.6 Structure of the Thesis .....	11
<b>2. Background and Literature Review .....</b>	<b>13</b>
2.1 Background .....	13
2.2 Techniques/Approaches Employed by Online Recruitment Systems.....	16
A. Traditional NLP Keyword-based Techniques.....	16
B. Relevance-based Models .....	16
C. Semantics-based Approaches .....	17
D. Machine Learning Techniques .....	17
2.3 Classification of Existing Online Recruitment Systems .....	18
2.3.1 The Impact of Semantic Web Technologies on Job Recruitment Processes .....	20
2.3.2 EXPERT.....	20
2.3.3 On-Line Consistent Ranking on E-Recruitment: Seeking the Truth behind a Well- Formed CV.....	21
2.3.4 MatchingSem .....	22
2.3.5 Matching Resumes and Jobs based on Relevance Models .....	23
2.3.6 E-Gen .....	23

2.3.7	Application of Machine Learning Algorithms to an online Recruitment System .....	24
2.3.8	Convex .....	24
2.3.9	A Hybrid Approach to Managing Job Offers and Candidates .....	25
2.4	Comparative Analysis between Online Recruitment Systems .....	26
2.5	Summary .....	30
<b>3.</b>	<b>Semantically-enhanced and Statistical-based Online Recruitment System .....</b>	<b>31</b>
3.1	Introduction .....	31
3.2	General Overview of the Proposed System .....	32
3.2.1	First Phase .....	32
3.2.2	Second Phase .....	34
3.2.3	Third Phase .....	37
3.3	Summary .....	38
<b>4</b>	<b>Theoretical Basis and Detailed Steps of the Proposed System .....</b>	<b>39</b>
4.1	Theoretical Basis .....	39
4.2	Development Details of the First Phase of the Prototype System .....	44
4.2.1	Concept Identification and Extraction .....	44
4.2.2	Construction of Semantic Networks .....	47
4.2.3	Missing Background Knowledge Handler .....	48
4.2.4	Extraction of Semantic Relations .....	50
4.2.5	Matching Semantic Networks .....	51
4.3	Development Details of the Second Phase of the Prototype System .....	52
4.3.1	Refinement of Candidate Concepts .....	52
4.3.2	Construction of Semantic Networks – Updated Version .....	54
4.3.3	Missing Background Knowledge Handler – Updated Version .....	55
4.3.4	Further Enrichment of the Semantic Networks .....	57
4.4	Development Details of the Third Phase of the Prototype System .....	58
4.4.1	From Unstructured Resumes and Job offers into Semi-structured Documents .....	58
4.4.2	Semantic Networks Construction – Updated Version .....	61
4.4.3	Missing Background Knowledge Handler – Updated Version .....	63
4.4.4	Matching the Semantic Networks .....	64
4.5	Summary .....	66
<b>5</b>	<b>Experimental Evaluation .....</b>	<b>68</b>
5.1	Evaluating the Effectiveness of the First Version of the Prototype System .....	68



5.2	Evaluating the Effectiveness of the Second Version of the Prototype System.....	70
A.	Experiments Using Expert Judgments.....	71
B.	Evaluating the System’s Effectiveness When Utilizing the Statistical-based Techniques .....	73
C.	Comparison with State-of-the-Art Systems .....	74
5.3	Evaluating the Effectiveness of the Third Version of the Proposed System .....	75
A.	The Impact of Utilizing Feature Extraction Techqniues on the Effectiveness of the Proposed System.....	76
B.	Experiments Using Expert Judgments.....	78
5.4	Summary .....	79
<b>6</b>	<b>Conclusions and Future Work .....</b>	<b>81</b>
6.1	Conclusions .....	81
6.2	Challenges and Future Work.....	83
	<b>References .....</b>	<b>85</b>
	<b>Appendix A .....</b>	<b>89</b>
	الملخص باللغة العربية .....	93

## List of Tables

<b>Table 1. Classification of the studied online recruitment systems.....</b>	<b>28</b>
<b>Table 2. Results of applying the NLP steps .....</b>	<b>46</b>
<b>Table 3. NRD values for the concept “j2ee” .....</b>	<b>49</b>
<b>Table 4. Lists of candidate concepts .....</b>	<b>53</b>
<b>Table 5. Lists of refined concepts .....</b>	<b>53</b>
<b>Table 6. The result of submitting "jsp" to HS Dataset .....</b>	<b>55</b>
<b>Table 7. Results of applying candidate concepts identification module. ....</b>	<b>61</b>
<b>Table 8. Lists of refined concepts .....</b>	<b>61</b>
<b>Table 9. Subset of manually enriched O*NET dataset.....</b>	<b>63</b>
<b>Table 10. Precision and Recall Results .....</b>	<b>69</b>
<b>Table 11. Precision results of the automatically generated relevance scores .....</b>	<b>72</b>
<b>Table 12. P/R Results using/not using the statistical techniques .....</b>	<b>73</b>
<b>Table 13. P, R and F-measure results .....</b>	<b>74</b>
<b>Table 14. The third version of the system results using/not using feature extraction techniques .....</b>	<b>76</b>
<b>Table 15. Precision results using FE techniques .....</b>	<b>78</b>

## List of Figures

<b>Figure 1. Stages of the manual recruitment process as proposed by (Lang et al., 2011)</b> .....	14
<b>Figure 2. General architecture of the proposed system (Version 1)</b> .....	33
Figure 3. General architecture of the proposed system (Version 2) .....	35
<b>Figure 4. General architecture of the proposed system (Version 3)</b> .....	37
<b>Figure 5. Semantic networks for CV1</b> .....	47
<b>Figure 6. Semantic networks for P1</b> .....	47
<b>Figure 7. Enrichment of the job post (P1) semantic network</b> .....	51
<b>Figure 8. Semantic networks of job post (P2)</b> .....	54
<b>Figure 9. Semantic networks of resume (CV2)</b> .....	54
<b>Figure 10. Updated semantic networks of job post (P2)</b> .....	56
<b>Figure 11. Updated semantic networks of resume (CV2)</b> .....	56
<b>Figure12 . Top 5 related skills returned by HS dataset to the title “Java Programmer”</b> .....	57
<b>Figure 13. Enrichment of the semantic networks of job post (P2)</b> .....	58
<b>Figure 14. Job post (P3) semantic networks</b> .....	62
<b>Figure 15. Resume (CV3) semantic networks</b> .....	62
<b>Figure 16. Updated semantic networks of job post (P3)</b> .....	64
<b>Figure 17. Updated semantic networks of resume (CV3)</b> .....	64

## **List of Abbreviations**

**HR** - Human Resources

**NLP** - Natural Language Processing

**CV** – Curriculum Vitae

**NER** - Named Entity Recognition

**SRM** - Structured Relevance Models

**SVM** - Support Vector Machine

**POST** - Part-of-Speech Tagging

**NRD** - Normalized Retrieval Distance

**NNP** - Proper Noun, Singular

**NNPS** - Proper Noun, Plural

**NN** - Noun, Singular or Mass

**SRE** - Semantic Relation Extraction

**WSD** - Word Sense Disambiguation

**HS** - HiringSolved

**FE** - Feature Extraction

## **1. Introduction**

Hiring the right talent is one of the most important and challenging tasks for the Human Resources (HR) department in any organization (Strohmeier and Piazza, 2013, Raziq and Shaikh, 2015). This challenge is amplified in online recruitment systems as employers often receive an enormous number of resumes – some of which are uploaded as unstructured documents in different formats such as .pdf, .doc, and .rtf (Kessler et al., 2012), while others are uploaded according to specific forms prepared by employers (Bizer et al., 2005, Mochol et al., 2007, García-Sánchez et al., 2006) – that are difficult to manually process and analyze.

Recently, many companies have shifted to using automatic online recruitment systems (Colucci et al., 2003, Kumaran and Sankar, 2013, Sivabalan et al., 2014) in an attempt to reduce the cost, time, and efforts required for screening out applicants and matching candidate resumes to their relevant job offers (Sivabalan et al., 2014). As reported by SAT telecom (Pande, 2011), the shifting from using manual recruitment to online recruitment has led to 44% of cost savings and reduced the time to fill a vacancy from 70 to 37 days.

Consequently, several techniques/approaches have been employed by online recruitment systems. These approaches include Natural Language Processing (NLP) and keyword-based screening (Belkin and Croft, 1992, Senthil Kumaran and Sankar, 2012, Kumaran and Sankar, 2013), relevance feedback models (Kessler et al., 2009), analytic hierarchy processes (Faliagka et al., 2011), semantics-based techniques (Colucci et al., 2003, Trichet et al., 2004, Bizer et al., 2005, Mochol et al., 2006, Mochol et al., 2007, Kumaran and Sankar, 2013), and machine learning algorithms (Chung-Kwan et al., 2000, Kessler

et al., 2007, Faliagka et al., 2012a, Faliagka et al., 2012b, Hong et al., 2013a). Although these approaches have proved to assist employers in screening out irrelevant resumes, they still suffer from low precision ratio when matching candidate resumes to their relevant job offers (Suerdem and Akalin, 2015). This is mainly due to the limitations associated with the underlying techniques that are employed by current online recruitment systems. For instance, systems that employ NLP and automatic keyword-based matching techniques ignore the semantic aspects of the entities encoded in the content of job offers and resumes. This consequently has led to producing unsatisfying results for employers since many of the returned resumes can be classified as false positives (when the retrieved resumes are irrelevant) and/or false negatives (when relevant resumes are not retrieved). For example, when an employer seeks a "quality assurance engineer", the system will retrieve all applications containing this particular phrase. However, it is clear that using this technique, all applications containing "QA engineer" phrase will not be retrieved. Other approaches propose to exploit machine learning and feature extraction based algorithms, ontologies, and knowledge bases to overcome the drawbacks of traditional keyword-based matching techniques. Although employing such approaches has led to significant improvements, they still suffer from problems associated with the limited domain coverage of the exploited resources and the lack of semantic knowledge captured by such resources (training data, ontologies, and knowledge bases) (Bizer et al., 2005, Kumaran and Sankar, 2013, Maree and Belkhatir, 2015a).

Starting from this position, we propose building an automatic online recruitment system based on exploiting multiple semantic resources and statistical-based concept-relatedness

measures to match between resumes and their relevant job offers. By employing multiple semantic resources, we aim to discover and derive the hidden semantic aspects of resumes and job offers. Although this approach has proved to be more effective than other approaches that rely on a single semantic resource (Maree and Belkhatir, 2015a), it still suffers from limitations associated with limited domain coverage and missing background knowledge in the used semantic resources. Therefore, to address this issue we propose to utilize statistical-based concept relatedness measures to further enrich the processed content of resumes and job offers with entities that are not recognized by the exploited semantic resources.

The rest of this chapter is organized as follows. Section 1.1 presents the background and motivations behind our work. Section 1.2 describes the problems associated with existing online recruitment systems. The research methodology is presented in section 1.3. Section 1.4 defines our research scope and the obstacles that we attempt to overcome through the proposed system. Section 1.5 presents our recently accepted publications in the field of automatic online recruitment. The structure of our thesis is presented in section 1.6.

### **1.1 Background and Motivations**

Human Resources department is one of the most important departments in any organization (Strohmeier and Piazza, 2013). Among the top priorities of this department is to find and hire qualified individuals whom have all necessary skills and meet all job requirements that enable them to contribute in constructing a promising future for the organization (Parry and Wilson, 2009). Indeed, recruiting is one of the most important, yet difficult, tasks for any HR department. This is due to the fact that employers usually

receive an enormous number of applications that are difficult to manually filter and analyze. As a consequence, this issue has led to the development and growth of automatic recruitment systems that attempt to facilitate the recruitment process by effectively addressing the actual needs of both employers and job seekers.

Though the development of online recruitment systems has several advantages over the conventional recruitment process, these systems still suffer from major limitations and drawbacks. For instance, NLP based models employ keyword-based matching techniques to compare between the requirements of job offers and the qualification information of users' resumes. A major drawback of this model lies in the fact that it ignores the semantic aspects of the terms that are extracted from both job offers and resumes (Shin et al., 2015). On the other hand, relevance based models have proved to be ineffective in performing resume-to-job post matching on large-scale real-world datasets as stated by (Yi et al., 2007). In the newer semantics-based approaches, semantic resources (ontologies and knowledge bases) are exploited to derive the semantic aspects of job offers and resumes. Although these approaches have shown better results in accomplishing the matching task, they still suffer from limitations associated with the exploited semantic resources, namely semantic knowledge incompleteness a.k.a. missing background knowledge, and limited domain coverage issues (Maree and Belkhatir, 2015a).

Motivated by these observations, we propose exploiting multiple semantic resources to address the limitations of employing a single semantic resource and to derive the semantic dimensions of the content of resumes and job offers. In addition, we utilize



statistical-based techniques to alleviate the problem of semantic knowledge incompleteness in the exploited resources.

## **1.2 Problem Statement and Research Questions**

Automatic online recruitment systems aim to support employers in hiring job applicants whom have all necessary skills and competencies that match the requirements of job offers. This also means that such systems implicitly screen out irrelevant applicants. Recently, organizations are shifting to using automatic online recruitment systems in their HR departments in order to reduce the time, effort and cost required for hiring new employees (Sivabalan et al., 2014). Although these systems succeeded in improving the recruitment process, they still suffer from a number of drawbacks associated with the limitations of their underlying techniques as detailed earlier in section 1.1.

In this section, we present the research questions that we attempt to investigate and address during our research work. Some questions are concerned with examining the current online recruitment systems, their underlying techniques, and identifying their pros and cons. Moreover, other questions are related to studying the applicability of employing multiple semantic resources and statistical-based techniques in the domain of online recruitment. Particularly, in this thesis we address the following questions:

- How to tackle the low precision problem of automatic recruitment systems?

Automatic keyword-based online recruitment systems suffer from limitations associated with low precision ratio (i.e. large portion of the returned results is irrelevant) (Belkin and Croft, 1992). This is due to the fact that these systems ignore the semantic aspects of the content of resumes and job offers. On the other hand, systems that employ the newer

machine learning and semantics-based techniques still suffer from limitations, namely semantic knowledge incompleteness and limited domain coverage stemming from the resources (training data, ontologies, and knowledge bases) employed by such techniques. To address these limitations, we propose an automatic online recruitment system that employs multiple semantic resources to compensate for missing background knowledge as described below.

- How to address the problem of missing background knowledge in the employed semantic resources?

Existing semantics-based approaches are penalized by limitations associated with the exploited semantic resources, namely semantic knowledge incompleteness and limited domain coverage (Maree and Belkhatir, 2015b, Hu et al., 2009). To overcome these limitations, we propose building an online recruitment system wherein we employ multiple semantic resources to cooperatively capture the semantic aspects hidden in the content of both resumes and job offers. By the term “cooperatively” we mean that more than one semantic resource will be exploited to compensate for missing background knowledge in the used resources, and also to ensure wider domain coverage. The used resources are also supported by other statistical and NLP techniques to effectively retrieve relevant resumes that match the requirements of their corresponding job postings.

- Is there a publicly available dataset in the field of online recruitment to evaluate the effectiveness of the proposed methods and techniques?

In order to answer this question, we extensively searched for a publically available datasets in the domain of online recruitment. Unfortunately, we didn't find a dataset that consists of manually assigned relevance scores between resumes and their corresponding

job posts. And thus, we have constructed a dataset that consists of 500 resumes, and seven job posts. Also, we have manually assigned relevance scores between resumes and their corresponding job posts.

### **1.3 Research Methodology**

The following steps present the main phases and tasks that we carry out during our research work:

#### ***1.3.1 Pre-processing of Unstructured Resumes/Job offers***

- **From Unstructured Resumes/Job offers into Semi-Structured Documents**

At this step, the content of job offers and resumes is pre-processed based on performing the following NLP steps (more details are provided in section 4.2.1):

- Document segmentation.
- Text tokenization.
- Stop words removal.
- Part-of-Speech Tagging (POST)
- Named Entity Recognition (NER).

By performing the abovementioned steps, we convert unstructured resumes and job offers into semi-structured documents through extracting specific segments such as Educational Background / Educational Requirements and Job Experience information.

- **Candidate Concepts Identification**

The outcomes of the previously mentioned NLP steps are further processed and filtered in order to identify and produce lists of concepts that belong to both job posts and resumes. Then based on employing the term frequency-inverse document frequency (tf-idf) (Belkin and Croft, 1992) weighting scheme and a predefined list of features we refine the lists of concepts as described below.

- **Candidate Concepts Refinement**

In order to refine the lists of candidate concepts, we attempt to remove concepts that appear to have negative impact on the matching process and do not contribute to the actual hidden semantic dimensions (i.e. meanings) encoded in the content of resumes and job offers. This means that concepts that either belong to a list of pre-defined terms (e.g. contact info, address, birth date) or have low tf-idf weights are automatically removed from the lists of candidate concepts. Then based on the refined lists of concepts, we construct semantic networks that relate the remaining refined concepts with different types of taxonomic and semantic relations as described below.

- **Construction of Semantic Networks**

Unlike conventional approaches that carry out the matching task between job offers and resumes based on their keyword representation (including the synonymous terms of these keywords), our goal of this step is to extend such models by incorporating additional semantic and taxonomic relationships by employing multiple semantic resources. In this context, the created indices for job offers and resumes are represented as semantic

networks. Once the semantic networks are constructed the matching process is then performed as detailed below.

- **Matching the Semantic Networks**

During this step, the semantic networks of the resumes and job offers are matched based on an edit distance function. This function returns measures of similarity between the strings of the nodes in the semantic networks. In this context, the higher the similarity between the networks, the more a job post and resume pair are considered relevant to each other.

### ***1.3.2 System Evaluation***

To validate the effectiveness of the proposed online recruitment system, we use state-of-the-art indicators (precision and recall), in addition we have compared the produced results by the proposed system with state-of-the-art systems. It is important to mention that due to the lack of publicly available datasets in the field of online recruitment, we have manually constructed a dataset that comprises 500 resumes and seven different job offers. The collected resumes and job offers are unstructured documents in different formats such as (.pdf) and (.doc). We would like to point out that prior to carrying out the evaluation process, we have converted the collected resumes and job offers into semi-structured documents by extracting specific segments such as Educational Background/Educational Requirements and Job Experience information. After that, we analyzed the corpus of the resumes and job offers through employing NLP techniques and multiple semantic resources to represent the semantics aspects of the content of resumes and job offers. Details of these steps are presented in chapter 4.

## 1.4 Contributions

The main contributions of our research work are summarized as follows:

1. Proposing an automatic precision-oriented online recruitment system that integrates multiple semantic resources, statistical-based concept-relatedness measures and feature extraction techniques to match between job offers and their relevant candidate resumes. Accordingly and unlike traditional single semantic-resource based online recruitment systems, our proposed system attempts to discover the hidden semantic dimensions encoded in the content of resumes and job offers based on more than one semantic resource. Additionally the proposed system utilizes statistical based concept-relatedness measures to address the issue of semantic knowledge incompleteness in the exploited semantic resources. Hence, the used concept-relatedness measures further enrich the discovered concepts with additional newly obtained concepts that were not initially recognized by the used semantic resources. Moreover, we transform the matching process into a more precision-oriented task by employing feature extraction techniques that convert unstructured (documents that do not have a semantically-overt, easy for a computer structure) resumes and job post into semi-structured documents. In this context, the system matches segments of resumes to their relevant segments of job posts instead of matching the whole unstructured versions of resumes and job offers.
2. Due to the lack of publicly available datasets in the recruitment domain, we have constructed a dataset that comprises 500 resumes and seven job offers acquired from different online resources. This dataset is currently available for public use

for researchers in the field<sup>1</sup>. Additionally, this dataset has been exploited to validate the effectiveness of our proposed system in our research article that has been accepted for publication recently (Kmail et al., 2015b).

### **1.5 Publications**

In this section, we list our recently accepted publications in the field of online recruitment. In the first article, we propose to build a prototype of an online recruitment system that incorporates the following modules: Concept Identification and Extraction, Semantic Networks Construction, Missing Background Knowledge Handler, Extraction of Semantic Relations, and Matching Semantic Networks. While in the second article, we extend the prototype of the proposed online recruitment system by incorporating new modules such as Refinement of Candidate Concepts and Further Enrichment of the Semantic Networks.

- Aseel B. Kmail, Mohammed Maree, and Mohammed Belkhatir.: “MatchingSem: Online Recruitment System based on Multiple Semantic Resources.” In the Proceedings of the 12<sup>th</sup> International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'15), pp. 2654 – 2659, (2015). DOI: 10.1109/FSKD.2015.7382376
- Aseel B. Kmail, Mohammed Maree, Mohammed Belkhatir, Saadat, M. Alhashmi.: “An Automatic Online Recruitment System based on Exploiting Multiple Semantic Resources and Concept-relatedness Measures.” In the Proceedings of the 27<sup>th</sup> IEEE International Conference on Tools with Artificial Intelligence (ICTAI'15), pp. 620 – 627, (2015). DOI: 10.1109/ICTAI.2015.95

### **1.6 Structure of the Thesis**

The rest of this thesis is organized into the following chapters. Chapter 2 introduces a background about automatic online recruitment systems. Additionally, in this chapter we present a comprehensive comparative analysis between existing online recruitment

---

<sup>1</sup> <https://www.dropbox.com/sh/fr8uqkaqk0gpyti/AACKvR844rU2lz4LntjV1ZVha?dl=0>

systems/approaches and classify them according to different categorization criteria. In chapter 3, we present a general overview of the architecture of our proposed online recruitment system and highlight the changes between the different versions of the proposed system. We present the theoretical basis and provide a detailed description of the methods and techniques that we employ in the proposed system in chapter 4. Chapter 5 presents the evaluation of the results produced by the proposed system using precision-recall indicators. Additionally, in this chapter we compare the produced result by our system to one of the existing online recruitment systems. In chapter 6, we draw the conclusions and highlight the future extensions of our research work.



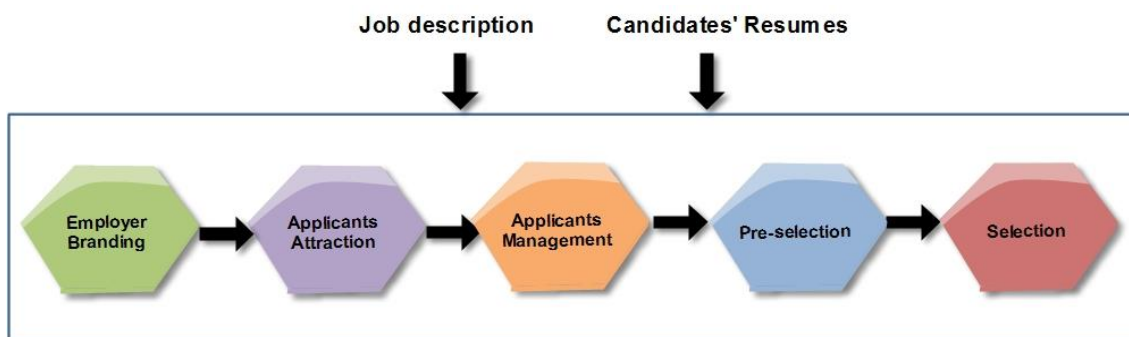
## **2. Background and Literature Review**

The aim of this chapter is to provide a state-of-the-art survey that covers several topics including recruitment (both manual and automatic recruitment strategies), approaches and techniques used by online recruitment systems, and the pros and cons of current online recruitment approaches. Accordingly, we start this chapter with a background about the recruitment process, and how this process has evolved from manual to automatic recruitment. Then, in section 2.2 we provide a general overview of existing techniques/approaches that are employed in the automatic recruitment domain. After that, in sections 2.3 and 2.4 we present a comprehensive comparative analysis between existing online recruitment systems/approaches, highlight their strengths and weaknesses and classify them according to different categorization criteria such as the goal of each system, implementation techniques/approaches, type of input, type of output and testing and evaluation method respectively. Finally, we summarize this chapter in section 2.5.

### **2.1 Background**

Recruitment is defined as the process of generating a pool of job seekers who are valuable for the company, have all necessary skills and expertise and meet all job requirements that enable them contribute in constructing a promising future for the organization (Sivabalan et al., 2014). Indeed, manual recruitment is one of the most difficult, time-consuming and tedious tasks for any HR department. In order to simplify this task, the manual recruitment process is divided into many phases. For example, (Carroll et al., 1999) recommend dividing the manual recruitment process into 4 stages: 1) decision making of whether a vacancy needs to be filled 2) job analysis 3) job description production 4) and a person specification. While other researchers such as

(Breaugh and Starke, 2000) propose to divide the manual recruitment process into 5 main phases: 1) short-term applicant attraction 2) long-term applicant attraction 3) applicant management, 4) pre-selection 5) final selection of candidates. On the other hand, Färber and his colleagues (Färber et al., 2003) split the manual recruitment process into two main phases: 1) attraction phase which includes planning and execution activities such as determination of target group, employee branding and attraction of direct applications and 2) selection phase which also includes planning and execution activities such as determination of selection criteria, pre-screening and final selection. A recent proposition has been made by (Lang et al., 2011) where the authors propose to divide the manual recruitment process into five stages as illustrated in Figure 1.



**Figure 1. Stages of the manual recruitment process as proposed by (Lang et al., 2011)**

As shown in Figure 1, the first stage of the recruitment process is “Employer Branding”. This stage and the second stage aim to create a good reputation for the organization in order to attract a large number of qualified applicants. During these phases, employers exploit different tools and instruments such as:

- General-purpose job boards (e.g. Monster.com and HotJobs.com).
- Job ads (e.g. adds in press and websites).
- Events (e.g. workshops and sessions).
- Employee referrals.

- Multimedia instruments.
- E-recruiting service providers.

The third stage is the “Management Phase”. During this stage, employers contact with applicants and manage their selection process. In this context, the selection process is separated into pre-selection and selection stages. During the pre-selection stage, applicants’ resumes and certificates are checked to screen out inappropriate candidates. While in the selection stage, applicants that are not screened out during the pre-selection stage are evaluated in order to make final recruitment decisions.

In the past, many organizations used manual recruitment to hire employees through collecting resumes from traditional media such as newspapers, magazines, job agencies and web sites. Then, recruiters select candidate applicants by skimming their resumes at a glance. After that, candidates are invited for interviews to test their communication skills, and language proficiency. Although manual recruitment performs well in screening out unqualified applicants, it still has limitations associated with the required effort, cost and time (Färber et al., 2003, Sivabalan et al., 2014) to match resumes to their relevant job offers. To address this issue, several online recruitment systems have been proposed to automate the recruitment process (Faliagka et al., 2012a, Hong et al., 2013a, Kumaran and Sankar, 2013). These systems are more preferred by employers and job seekers than traditional recruitment methods due to their advantages (Pande, 2011). For example, online recruitment systems are cost effective, easy to use, have proper targeting in any field or industry, generate fast response, allow to build up a database of candidates for talent searching, enable employers to present more information regarding the required job skills and competencies and allow them to have better access to talents (Pande, 2011,

Sivabalan et al., 2014). By reviewing state-of-the-art online recruitment systems, we can clearly find that they have employed different techniques and approaches for automating the manual recruitment process. In the following section, we provide more details about these techniques and approaches.

## **2.2 Techniques/Approaches Employed by Online Recruitment Systems**

Several techniques and approaches have been proposed to construct automatic online recruitment systems (Lee, 2007, Faliagka et al., 2011). In this section, we classify these techniques and discuss the major drawbacks and limitations that are associated with each technique.

### **A. Traditional NLP Keyword-based Techniques**

These techniques mainly depend on exact matching between keywords extracted from the content of job offers and candidate resumes. Systems that employ such techniques suffer from low precision wherein a large portion of the returned results is irrelevant. This is because keyword-based techniques ignore the underlying semantic aspects of the terms that are extracted from both job offers and resumes (Kumaran and Sankar, 2013)

### **B. Relevance-based Models**

Relevance-based models are usually built from known relevant resumes to a specific job post (Lavrenko and Croft, 2001). While in Structured Relevance Models (SRM) approach (Yi et al., 2007), relevance models are built from highly ranked documents. In this context, relevance models are used to compensate for vocabulary variations between resumes and job descriptions. Similar job offers are grouped by matching a candidate job description with a collection of job descriptions. After that, resumes that are relevant to those job descriptions are used to construct relevance models to capture terms that are not

explicitly mentioned in job descriptions. A major problem of these approaches is their low precision when tested against large-scale real-world datasets (Yi et al., 2007).

### **C. Semantics-based Approaches**

As stated in (Mochol et al., 2006), the exploitation of semantic resources in the recruitment domain assists in using shared vocabularies to describe job descriptions and resumes. The authors of (Trichet et al., 2004, Bizer et al., 2005, Mochol et al., 2007, García-Sánchez et al., 2006) propose using automatic recruitment systems that employ semantic resources that have been built based on integrated classifications and standards. In (Mochol et al., 2007), the authors propose using a human resource ontology (HR-ontology) to gain uniform representation of resumes and job offers and to accomplish the matching process at the semantics level. Another semantics-based system is EXPERT (Kumaran and Sankar, 2013) which constructs ontology documents that describe both job offers and resumes based on the concept linking approach (Senthil Kumaran and Sankar, 2012), and then ontology documents of job offers are mapped to ontology documents of resumes. Although these approaches have shown better results in accomplishing the matching process, they still face significant problems concerned with the development of complete and reliable ontologies that capture up-to-date knowledge about specific domains (Maree and Belkhatir, 2015a).

### **D. Machine Learning Techniques**

A number of machine learning algorithms are exploited in the online recruitment domain for data analysis and information extraction. These algorithms include neural networks (Chung-Kwan et al., 2000), support vector machines (Kessler et al., 2007), decision trees (Ramar and Sivaram, 2010) and clustering (Hong et al., 2013b). Among the systems that

employ machine learning techniques is E-Gen (Kessler et al., 2007). The authors of this system propose to automate the recruitment process through classifying and analyzing unstructured job offers using vectorial and probabilistic models. In addition, Support Vector Machine (SVM) classification algorithms are employed to annotate segments of job offers with appropriate topics and features. As reported in (Kessler et al., 2012), the main drawback of machine learning approaches is that they produce high error rates as they rely on manually-developed training corpora.

### **2.3 Classification of Existing Online Recruitment Systems**

In this section, we present a comprehensive comparative analysis of existing online recruitment systems/approaches and classify them according to the following categorization criteria:

- **Goal of the system:** some of the studied systems are only concerned with matching job offers to their relevant resumes (i.e. either in or out model), while others focus on ranking applicants' resumes according to their relevance scores.
- **Implementation techniques/approaches:** another criterion that we use to classify online recruitment systems is the techniques/approaches that are employed by each system. These techniques include keyword-based matching, semantics and knowledge based methods, machine learning algorithms, and a hybrid of these approaches.
- **Type of input (structured, unstructured, or semi-structured):** the type of input recognized by online recruitment systems is also of particular importance. This input (acquisition method of both resumes and job offers) can be obtained in the form of structured or semi-structured documents generated by filling specific

forms prepared by the employer, or it can be gathered as unstructured documents uploaded by both employers and job seekers in different formats such as .pdf and .doc.

- **Type of output:** another important criterion that we consider for categorizing online recruitment systems is the type of output that each system produces. Basically, the output produced by online recruitment systems can belong to one of two categories. In the first category, the produced results are characterized by their relevance/non relevance to a given job post. The systems of the second category extend this approach by producing ranked results. In this context, such systems do not only filter a given set of resumes (i.e. match/ not match), but they also recommend highly ranked resumes to their relevant job offers.
- **Testing and evaluation method:** different evaluation mechanisms have been carried out to test and evaluate the effectiveness of the proposed recruitment systems, and to find whether the returned results (resumes) by each system are truly relevant to their corresponding job offers. Some researchers have conducted experiments using real-world scenarios and manually-crafted datasets, while others have implemented prototypical systems wherein they tested the overall effectiveness of the employed techniques. It is important to mention that evaluating the techniques and approaches exploited in constructing online recruitment systems is of great interest as they can be successfully adopted in practical settings and have their positive impact on the revenue models of the companies that adopt them.

For the rest of this section, we will describe the characteristics of different online recruitment systems and classify them according to the introduced set of categorization criteria.

### **2.3.1 The Impact of Semantic Web Technologies on Job Recruitment Processes**

Christian Bizer et al. (Bizer et al., 2005) propose to match job offers to their relevant resumes based on employing semantic resources. In this context, a human resource ontology (i.e. semantic resource) is constructed by integrating widespread standards and classifications to annotate job offers and candidate resumes. In order to collect candidate resumes (i.e. applicants' information), web-based application forms are used to acquire CVs as semi-structured documents. Then, they utilize the human resource ontology to derive the semantic aspects of the produced semi-structured resumes and job offers. Finally, the semantic matching algorithm is employed to generate a list of qualified applicants.

Although semantics-based approaches enhance the effectiveness of online recruitment systems (Mochol et al., 2007), they are penalized by limitations of the exploited semantic resources, namely semantic knowledge incompleteness and limited domain coverage (Maree and Belkhatir, 2015a). On the other side, in the proposed approach, the authors rely on web-based application forms to acquire applicants' information as semi-structured resumes. This would be a tedious and time-consuming task for applicants (Sivabalan et al., 2014).

### **2.3.2 EXPERT**

The goal of the proposed system by (Kumaran and Sankar, 2013) is to match between resumes and job offers based on employing semantics and knowledge based methods in a



similar manner to the previously mentioned system. However, in order to start the matching process, this system first produces ontological representations of resumes and job offers to capture knowledge encoded in the content of resumes and job offers. After that, the ontology documents (ontological representations) of resumes are mapped to ontology documents of job offers to retrieve relevant candidates. In this context, ontology mapping (Kalfoglou and Schorlemmer, 2003) approach is utilized to determine the correspondences between the concepts of the produced ontology documents.

To measure the effectiveness of the proposed system, the authors evaluated its precision in assigning relevance scores between job offers and applicant resumes. In order to accomplish this task, two CV sets are used. The first CV set consists of structured resumes, while the second CV set consists of unstructured resumes and job offers. The results show high precision and recall ratios indicating the effectiveness of employing semantics and knowledge-based methods in the domain of online recruitment. Nevertheless, when we compare their system with our previous work (Kmail et al., 2015b), we find that our proposed system has been more effective and precise in matching resumes to job offers.

### **2.3.3 On-Line Consistent Ranking on E-Recruitment: Seeking the Truth behind a Well-Formed CV**

In the work that is presented in (Faliagka et al., 2014), job applications are evaluated and ranked by exploiting semantics-based matching techniques and machine learning algorithms. First, the proposed system extracts a set of features from the applicants' LinkedIn profiles and matches them semantically against job offers' descriptions. In order to accomplish this task, a single semantic resource has been constructed by domain experts to derive the semantic aspects of resumes and job offers. In addition, linguistic

analysis is utilized to analyze candidates' blogs to extract features that reflect their personality traits and social behaviors. After that, supervised machine learning algorithms are employed to generate a list of qualified applicants ranked according to their relevance. Although employing machine learning and semantics-based techniques have proved to assist employers in screening out irrelevant resumes, they still suffer from limitations, namely semantic knowledge incompleteness and limited domain coverage stemming from the resources (training data, ontologies, and knowledge bases). This system is evaluated in a real-world recruitment scenario by comparing manually calculated scores between resumes and job offers to those produced by the system. The results have shown high accuracy except for job offers that require special skills.

#### **2.3.4 MatchingSem**

MatchingSem (Kmail et al., 2015a) is an online recruitment system that matches unstructured documents (resumes and job offers) based on employing multiple semantic resources and statistical-based techniques. The proposed system first employs NLP tools to find and extract lists of candidate concepts from the content of both resumes and job offers. Next, existing semantic resources are cooperatively employed to analyze the lists of candidate concepts at the semantics level. When a concept is not recognized by the used semantic resources, statistical-based concept-relatedness measures are then used to address this issue.

To evaluate the effectiveness of methods and techniques employed in the proposed system, an experimental instantiation is conducted by comparing manually assigned scores between resumes and job offers and those produced by the proposed system in the same manner as carried out in (Faliagka et al., 2014). Although the system shows high

precision and recall ratios for most of the examined job offers, it suffers from low precision and recall ratios for job positions that require specific years of experience.

### **2.3.5 Matching Resumes and Jobs based on Relevance Models**

This system has been proposed in (Yi et al., 2007) to match semi-structured resumes and job offers in real-world large scale recruitment scenarios. It as well supports applicants ranking according to their similarity scores. Relevance models are usually built from known relevant resumes to a specific job post. While in Structured Relevance Models (SRM) approach, relevance models are built from highly ranked documents. In this context, relevance models are used to compensate for vocabulary variations between resumes and job descriptions. Similar job offers are grouped by matching a candidate job description with a collection of job descriptions. After that, resumes that are relevant to those job descriptions are used to construct relevance models to capture terms that are not explicitly mentioned in job descriptions. A major problem of this approach is its low precision when tested in large-scale real-world recruitment scenarios.

### **2.3.6 E-Gen**

E-Gen (Kessler et al., 2007, Kessler et al., 2009) is an automatic recruitment system that matches unstructured resumes to their relevant job offers. E-Gen has been built based on employing Support Vector Machine (SVM) classification algorithms in order to annotate segments of job offers with the appropriate topics and features. Additionally, E-Gen addresses the issue of ranking applicants according their relevance score by utilizing the vector space model. In this context, job offers and resumes are transformed into vector space representations, and then similarity measures for their associated vectors are computed. After that, relevance feedback is utilized to expand the job post vector

representation with terms extracted from relevant candidate resumes. Next, similarity measures are recomputed in order to produce better results. Experimental instantiation of the proposed system is conducted to prove its effectiveness in real-world recruitment scenario. However, the utilized SVM classification algorithms are subjective to high error rates since they depend on manually developed training corpora (Kessler et al., 2012).

### **2.3.7 Application of Machine Learning Algorithms to an online Recruitment System**

Faliagka and his colleagues (Faliagka et al., 2012a) propose to construct an online recruitment system based on employing machine learning algorithms. The proposed system starts by analyzing job offers and semi-structured resumes acquired by web-based forms and applicants' LinkedIn profiles. Then, machine learning algorithms are utilized to produce a list of qualified applicants ranked according to their relevance. In this context, the ranking process mainly focuses on learning a scoring function that calculates relevance scores between resumes and their relevant job offers. Therefore, a set of training data is collected by domain experts to further learn the required scoring function. An Experimental instantiation of the proposed system has been installed to validate its effectiveness in matching resumes against job offers. Although the authors argue that the produced results are satisfying in identifying applicant's personality traits, the consistency of the produced results (list of qualified applicants ranked according to their relevance) is highly dependent on the offered job offers. For example, it is difficult to learn a scoring function for senior positions which require specific experience and skills.

### **2.3.8 Convex**

Convex (Dan, 2004) is an automatic recruitment system that has been built to match unstructured/semi-structured resumes to job offers. The proposed system starts by

employing a single domain-specific knowledge base in an attempt to extract concepts from both job offers and candidate resumes. If the used knowledge base fails in identifying a specific concept due to the limitations of the used knowledge base, namely semantic knowledge incompleteness and limited domain coverage, then extraction techniques are utilized to compensate for missing background knowledge. Concepts extraction techniques include shallow natural language parsing, and heuristics. On the one hand, shallow natural language parsing uses two domain-independent, language-specific NLP techniques to extract noun phrases as concepts (barrier word algorithm (Bourigault, 1992) and parts-of-speech tagging). On the other hand, rule-based heuristics are employed by domain experts to further extract other relevant concepts that were not captured by NLP techniques. Once concepts are extracted, the matching process produces a list of qualified applicants. To validate the effectiveness of the proposed system the authors compare manually assigned relevance scores between resumes and job offers with those produced automatically by the proposed system. The results show that Convex is better than those employ keyword-based or statistical-based techniques. However, the proposed system employs a single knowledge base, accordingly if the concepts in resumes and job offers are not captured by the used knowledge base, then the system fails drastically in finding relevant applicants.

### **2.3.9 A Hybrid Approach to Managing Job Offers and Candidates**

The work presented in (Kessler et al., 2012) is an extended version of E-Gen system that utilizes a hybrid approach that combines statistical-based algorithms and vector space representations to precisely match resumes and job offers. The proposed system appends a summarization module to exclude irrelevant information contained in resumes and

cover letters according to specific compression criteria determined by employers. The updated version of the system composed mainly of three main modules:

- A module to extract information of job offers.
- A module to analyze the resumes and cover letters.
- A module to compute relevance scores between resumes and job offers.

According to the authors, in order to evaluate the precision of the proposed system, experimental validations are carried out on a huge dataset that consists of 1917 resumes and three job offers. Although the produced results are satisfactory, they depend on a summarization criteria determined by the employer. Accordingly, the produced results are inaccurate and prone to error rates particularly if the determined compression criteria exclude information that has significant value in the matching process.

#### **2.4 Comparative Analysis between Online Recruitment Systems**

As shown in Table 1, we have conducted a comprehensive comparative analysis between existing online recruitment systems/approaches and classified them according to different categorization criteria. We can see that the goal of the majority of the above mentioned systems is ranking applicants according to their relevance scores. While the minority of the systems only focus on matching between resumes and job offers. On the other hand, the type of input varies from one online recruitment system to another. Some systems accept unstructured resumes and job offers as input, while others are concerned with structured or semi-structured resumes and job offers. We would like to point out that – in the context of our work – we aim to analyze and match unstructured resumes to job offers. Concerning the employed techniques and approaches, it is clear that semantics

based techniques and machine learning algorithms are the dominant techniques and have been exploited by most of the systems.

**Table 1. Classification of the studied online recruitment systems**

Index	System/Approach	Goal of the system	Implementation techniques/approaches	Type of input	Type of output	Testing and evaluation method
2.3.1	The Impact of Semantic Web Technologies on Job Recruitment Processes	Matching candidate applications to job offers	Semantics and knowledge based methods	Semi-structured resumes and job offers	List of qualified applicants	A prototypical implementation of the system without using experiments in real-world scenario
2.3.2	Expert	Matching resumes to job offers	Semantics and knowledge based methods	Structured / unstructured resumes and job offers	List of qualified applicants	Evaluated in real-world recruitment scenario using two data sets of Structured and unstructured resumes and job offers
2.3.3	On-line Consistent Ranking on E-recruitment: Seeking the Truth Behind a Well-Formed CV	Ranking applicants according to their relevance score	Semantics based methods and machine learning algorithms	Structured resumes and job offers	List of applicants ranked according to their relevance scores	Evaluated in real-world recruitment scenario. The system shows good accuracy except for job offers that require special skills
2.3.4	Matchingsem	Matching resumes to job offers	Semantics based methods and statistical-based techniques	Unstructured resumes and job offers	List of qualified applicants	Evaluated in real-world recruitment scenario. The system shows good accuracy except for job offers that require special years of experience
2.3.5	Matching Resumes and Jobs Based on Relevance Models	Ranking applicants according to their relevance score	Structured relevance models (SRM).	Semi-structured resumes and job offers	List of qualified applicants	Evaluated in a large-scale real-world recruitment scenario by comparing manually assigned scores and those produced by the system



**Table 1. (continued)**

<b>Index</b>	<b>System/Approach</b>	<b>Goal of the system</b>	<b>Implementation techniques/approaches</b>	<b>Type of input</b>	<b>Type of output</b>	<b>Testing and evaluation method</b>
2.3.6	E-Gen	Ranking applicants according to their relevance score	Machine learning algorithms and vector space model	Unstructured job offers and resumes / cover letters	List of applicants ranked according to their relevance scores	Evaluated in real-world recruitment scenario.
2.3.7	Application of Machine Learning Algorithms to an Online Recruitment System	Ranking applicants according to their relevance score	Machine learning algorithms and linguistic analysis	Structured / unstructured resumes and job offers	List of applicants ranked according to their relevance scores	Evaluated in real-world recruitment scenario. The results show that the system is effective in identifying personality traits
2.3.8	A Hybrid Approach to Managing Job Offers and Candidates	Ranking applicants according to their relevance score	Machine learning algorithms, statistical-based techniques and vector space model	Unstructured job offers and resumes / cover letters	List of applicants ranked according to their relevance scores	Evaluated in real-world recruitment scenario. They used a huge dataset of 1917 resumes and three job offers
2.3.9	Convex	Matching resumes to job offers	A hybrid of Semantics based methods, shallow natural language processing and heuristics.	Unstructured / semi-structured resumes and job offers	List of qualified applicants	Evaluated in real-world recruitment scenario by comparing manually assigned scores with those produced by the system

This is due to the fact that semantic resources play a crucial role in discovering the semantic aspects hidden in the content of both resumes and job offers.

Considering the testing and evaluation methods of the studied systems and approaches, we can notice that some experiments don't express the precision of the proposed system since they don't rely on a real-world recruitment scenario. But, in our proposed system, we measure its effectiveness using a real-world recruitment scenario in the same fashion as proposed in (Kumaran and Sankar, 2013, Dan, 2004, Kessler et al., 2012, Faliagka et al., 2014). To judge the quality of results generated from these systems, manually assigned relevance scores (a.k.a. expert judgments or ground truth) are usually compared to their corresponding automatically generated scores by the system.

## **2.5 Summary**

The purpose of this chapter was to present a literature review about the recruitment process (manual recruitment and online recruitment). We have elaborated the techniques/approaches employed in constructing online recruitment systems including traditional NLP keyword-based techniques, relevance based models, semantics-based approaches and machine learning techniques. And further, we classified the studied online recruitment systems according to different categorization criteria such as type of input/output, goal of the system, implementation techniques/approaches and testing and evaluation methods. We concluded that semantics-based techniques are the dominant techniques that have been employed by most of existing systems and approaches. This is due to the fact that semantic resources are capable of representing several domains and derive the semantic aspects of resumes and job offers.

### **3. Semantically-enhanced and Statistical-based Online Recruitment System**

#### **3.1 Introduction**

Semantic resources have been extensively used in different fields such as knowledge representation and management (Brewster et al., 2004, Maree and Belkhatir, 2015a), semantics-based search (Bhagdev et al., 2008, Shin et al., 2015) and online recruitment (Trichet et al., 2004, Bizer et al., 2005, Mochol et al., 2007, Faliagka et al., 2014, Kessler et al., 2012). In these areas, semantic resources are employed to capture knowledge represented in various relevant domains by formally and explicitly specifying concepts used in each domain, and the constraints on their use and reuse (Gruber, 1993).

In the domain of online recruitment, knowledge encoded in either a single semantic resource (Mochol et al., 2007), or in multiple semantic resources (Kmail et al., 2015b) is used to discover the hidden semantic dimensions of the content of resumes and job offers. Examples of these semantic resources are WordNet (Miller, 1995) and YAGO3 (Mahdisoltani et al., 2015). Details of these semantic resources are listed below.

- WordNet: is a generic semantic resource that is created manually to cover different domains. It groups the concepts into sets of synonyms called synsets. These synsets are connected with different types of taxonomic and semantic relations such as hypernymy, meronymy, and hyponymy. In the context of our work, WordNet is primarily used for automatic text analysis and Word Sense Disambiguation (WSD). Additionally, we utilize the knowledge encoded in WordNet to construct semantic relations between the different concepts that we extract from the content of both resumes and job postings .

- YAGO3: is a high quality huge semantic resource that was developed at the Max Planck Institute for Computer Science in Saarbrücken. YAGO3 is automatically derived from online repositories of organized human knowledge, namely Wikipedia, and other online structured sources such as GeoNames. The current version of this semantic resource contains more than 10 million entities (such as persons, movies and cities). Additionally, it contains more than 120 million facts about these entities. The precision of YAGO3 has been tested resulting in a confirmed 95.03% of accuracy which makes it a special and trusted resource of semantic information (Medelyan et al., 2009).

In the next section, we present an overview of our proposed online recruitment system and clarify the overall architecture of the proposed system. Finally, we summarize this chapter in section 3.3.

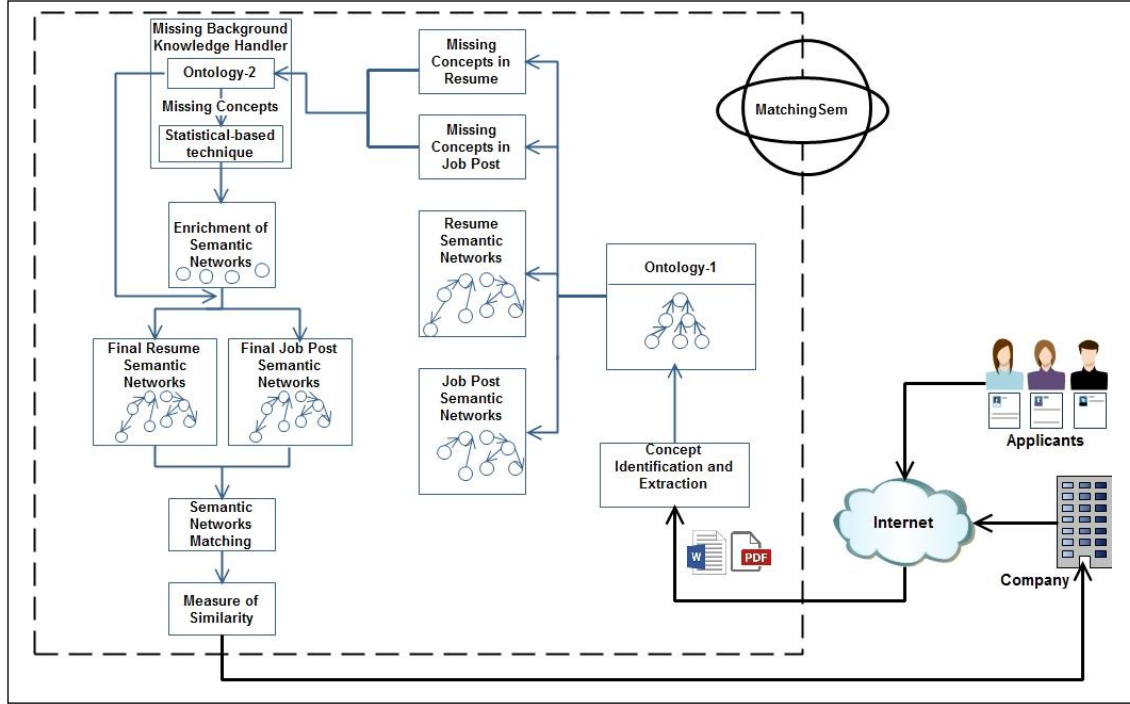
### **3.2 General Overview of the Proposed System**

In this section, we present a general overview of our proposed automatic online recruitment system wherein we combine multiple semantic resources and statistical-based concept-relatedness measures to effectively match between job offers and their relevant resumes. Screenshots of the proposed online recruitment system are shown in Appendix A. Also, it is important to point out that the development of the proposed system has been carried out through three major phases as described in the following sections.

#### **3.2.1 *First Phase***

During this phase, we have constructed the first version (named as MatchingSem1) of our proposed online automatic recruitment system based on employing knowledge captured in multiple existing semantic resources, namely WordNet and YAGO2 (Hoffart et al.,

2011). Both resources have been exploited to effectively match between candidate resumes and job offers. In addition, we have utilized statistical-based concept-relatedness measures to alleviate the problem of semantic knowledge incompleteness (i.e. missing entities) in the used resources. Figure 2 depicts the overall architecture of the first version of the proposed system.



**Figure 2. General architecture of the proposed system (Version 1)**

As show in Figure 2, the first version of the proposed system comprises several modules that are organized in the following order:

First, the *Concept Identification and Extraction (CIE)* module is used to identify and extract lists of candidate concepts from the content of both job postings and resumes. To do this, we employ various Natural Language Processing (NLP) tools such as n-gram tokenization, stop words removal, and Part-of-Speech Tagging (POST). Then, the system takes the identified lists of concepts (from both the segments of job postings and resumes) as input in order to construct semantic networks that connect the derived concepts with

different types of taxonomic and semantic relations such as hypernymy, meronymy, related-to, and synonymy.

To build the semantic networks, we first map each concept from the lists of candidate concepts to its correspondences in WordNet. At this step, we may have some concepts that are not defined in this semantic resource. Accordingly, we utilize the *missing background knowledge handler* in an attempt to find those concepts in the second semantic resource (YAGO2 in this version of the prototype system). We would like to point out that although we are using more than one semantic resource, we were still faced with the fact that some of the candidate concepts are not found in any of the used resources. To address this issue, we utilized the statistical-based semantic relatedness technique to measure the semantic closeness between the concepts that are missing in the used semantic resources and those that are defined in them. By employing this technique, we are able to obtain an additional set of proposed concepts to further expand the semantic networks of resumes and job offers. After this step, the matching algorithm takes the constructed and enriched semantic networks as input and produces measures of relatedness between the resumes and job offers based on the semantic relatedness between their semantic networks. Experimental validation of this prototype system has been conducted and the produced results of the proposed system are discussed in more details in chapter 5.

### **3.2.2 Second Phase**

By experimentally evaluating the effectiveness of the first version of the proposed prototype system, we found that the produced matching results were satisfactory and closely related to the manually assigned relevance scores between the job offers and their

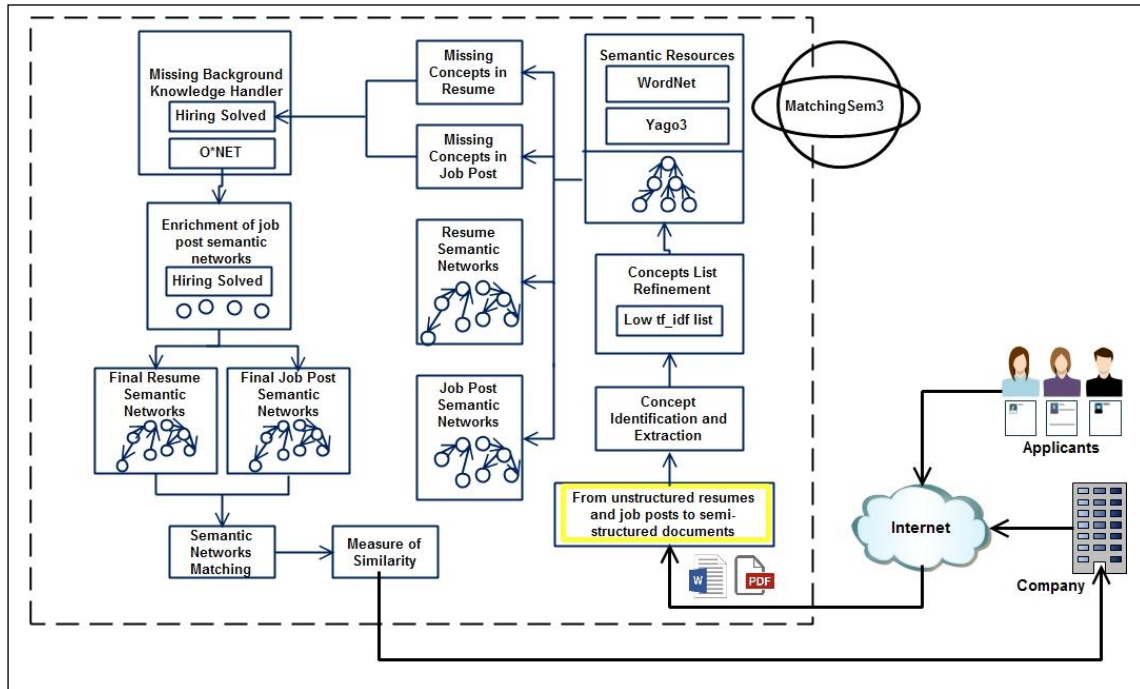


As shown in Figure 3, the new *Refinement of Candidate Concepts* module has been incorporated into the second version of the proposed system in order to refine the lists of concepts through removing those that appear to be of little value and don't have significant contribution to the actual meaning of the content of resumes and job offers. Examples of these concepts are those that usually fall under specific sections in the resume such as: candidate's name, contact info, etc. In addition, concepts that have low tf-idf weights (Belkin and Croft, 1992) are removed such as (core, accessing, address, etc.). More details can be found in section 4.3.1. Next, the third module of the proposed system takes the refined lists of concepts (from the segments of both the job offers and resumes) as input to construct semantic networks in which concepts are connected by various types of semantic relations (derived from WordNet and YAGO2). As we have pointed out earlier, during this step, we may find that some concepts are not defined in the exploited semantic resources. These concepts are then submitted to the *Missing Background Knowledge Handler* wherein we utilize an additional resource (Hiring Solved (HS) dataset) (HiringSolved, 2015) to enrich the constructed semantic networks with additional semantically-related concepts. HS dataset defines a huge number of terms in the form of skills – either mentioned in job offers or resumes – and the weights of the semantic relatedness between those skills. In this context, semantically-relevant concepts are extracted and used to expand the constructed semantic networks. Finally, the matching algorithm takes the updated semantic networks as input and produces the measures of semantic closeness between them as output. To validate the effectiveness of this version of the proposed system, experimental validations were conducted by comparing the manually assigned relevance judgments between resumes and job offers and those produced



automatically by the system (Kmail et al., 2015b). The overall precision of the results produced by the second version of the system were promising and the system have proved to be more effective in matching between job postings and their relevant candidate resumes. However, we would like to point out that for some job offers (those that require a specific number for the years of experience or a certain educational level) the precision of the system was not satisfactory. To address this issue, we decided to update the current version of the system by incorporating a new Semi-structured Feature Extraction based Conversion module through which we attempt to extract such features from the content of resumes and job postings.

### 3.2.3 Third Phase



**Figure 4. General architecture of the proposed system (Version 3)**

In order to overcome the limitations associated with the low precision ratios for particular job offers that require specific educational background and experience, we have

integrated a new module – *From Unstructured Documents to Semi-Structured Documents* – that converts the original unstructured format of resumes and job offers into semi-structured formats. In this context, instead of matching unstructured versions of resumes and job offers, this version of the prototype system matches segments of resumes to their relevant segments of job offers. Experimental instantiation of the proposed system has been installed to demonstrate the effectiveness of the proposed approach in matching job posts to resumes. For more details on the carried out experiments, please refer to chapter 5.

### **3.3 Summary**

Our aim in this chapter was to present a general overview of our proposed online recruitment system and to clarify the overall architecture of the proposed system. Also, we elaborated that the development of the prototype system has been carried out through three major phases. In addition, we clarified the reasons behind moving from one phase into another.

## 4 Theoretical Basis and Detailed Steps of the Proposed System

In this chapter, we introduce the theoretical basis that lie behind our work and present the implementation details of the prototype system. Section 4.1 introduces the theoretical background and presents the formal definitions and characterizations of the methods and techniques that are used in our system. Then, we elaborate, in details, each of the development phases of the proposed system in sections 4.2, 4.3 and 4.4 respectively. Finally, we summarize this chapter in section 4.5.

### 4.1 Theoretical Basis

Before proceeding to present the details of the methods and techniques used in the proposed system, we formally define – in the context of our work – the terms “Semantic Resource”, “Semantic Network”, “Enrichment of the Semantic Networks”, “Normalized Retrieval Distance (NRD)”, “Jaro-Winkler Distance” and “tf-idf Weighting”.

#### Definition 1: Semantic Resource:

A semantic resource  $SR$  is *quadruple*,  $SR := \langle C, P, I, V \rangle$  where:

- $C$  represents the set of concepts that are defined in  $SR$ . The hierarchical relationship between concepts of the set  $C$  is a pair  $(C, \leq)$ , where  $\leq$  is an order relation on  $C \times C$ .

We call  $\leq$  the sub-concept relation.

- $P$  represents the set of properties defined over  $C$ .
- $I$  is the set of individuals also called instances of the concepts in  $SR$ .
- $V$  is the set of values defined over  $P$ .

For each job post and resume pair, the system takes the extracted lists of concepts (using the NLP techniques detailed in section 4.2.1) as input and produces as output the following sets of semantic networks:

- The set of semantic networks  $S\zeta_j = \{\zeta_{j1}, \zeta_{j2}, \zeta_{j3}, \zeta_{jn}\}$  that are automatically derived and constructed from the job post.
- The set of semantic networks  $S\zeta_r = \{\zeta_{r1}, \zeta_{r2}, \zeta_{r3}, \zeta_{rn}\}$  that are automatically derived and constructed from the resume.

To automatically construct such networks, we rely on the exploited semantic resources  $S = \{SR1, SR2, SR3\}$ . We formally define a semantic network as:

**Definition 2: Semantic Network:**

A semantic network  $\zeta := \langle CC, RR \rangle$  where:

- CC is the set of concepts captured by  $\zeta$ . These are the resume and job post concepts that are also captured in the used semantic resources.
- RR is the set of relations that connect the concepts in CC. Similar to CC, these relations are obtained from the used set of semantic resources  $S = \{SR1, SR2, SR3\}$ .

As we have discussed earlier in section 3.2.2, the exploitation of more than one semantic resource does not necessarily guarantee that concepts, their instances, and the relations that link the concepts in each and every domain are fully covered. Therefore, we may encounter the problem of unrecognized entities (either concepts or their instances) in the used semantic resources. To tackle this problem, we have proposed two approaches:

- 1) The first approach is employed in the first version of the prototype system wherein we utilize the *statistical-based concept-relatedness* technique to automatically measure the strength of the semantic closeness between the missing concepts and the concepts that are defined in the semantic networks. Formally, we describe this technique as follows:

### - Semantic Networks Enrichment

The enrichment algorithm takes the set of concepts that are not recognized by the used semantic resources  $S_{-missing} = \{c_1, c_2, c_3, \dots, c_n\}$  and the network  $\zeta$  as input, and produces as output for each  $cc \in CC$  in  $\zeta$  a set of  $S(cc) \subseteq S_{-missing}$ .

Where:

- $S(cc)$  represents the proposed expansion candidates for  $cc$ . A candidate  $c \in S_{-missing}$  can be a single-word or compound-word from  $S_{-missing}$ . The proposed set  $S(cc)$  can be obtained using the Normalized Retrieval Distance (NRD) (Kmail et al., 2015a) algorithm. In this algorithm, we use a threshold value  $v$  ( $v=0.70$ ) based on Equation 1 to automatically decide upon which concepts should be considered as potential enrichment candidates and which concepts should be excluded from the set of enrichment candidates.
- 

$$S(cc, v) := \{c \in S_{-missing} \mid NRD(cc, c) \leq v\} \quad (1)$$

- 2) The second approach for tackling this problem is used in the second and third versions of our proposed system. In this approach, we exploit *HiringSolved Dataset* in order to enrich the semantic networks of both job offers and resumes. Formally, we define the updated process of semantic networks enrichment as follows:

### - Semantic Networks Enrichment – Updated Version

The semantic networks enrichment process takes a given semantic resource  $SR$  (*HiringSolved* in our work) and a given concept  $cr$  as input and produces for  $cr$  a set  $S(cr) \subseteq T(SR)$  as output.

Where:

- $S(cr)$  is the set of suggested enrichment candidates for  $cr$ . A candidate  $t \in T(SR)$  is a single-term or compound-term from  $SR$ .
- $T(SR)$  is the set of entities defined in  $SR$ .

**Definition 3: Normalized Retrieval Distance (NRD):**

We employ the NRD algorithm (Maree et al., 2011) which is an adapted form of the Normalized Google Distance (NGD) technique proposed in (Cilibrasi and Vitanyi, 2007).

The NRD determines the semantic closeness between two terms as follows:

Given two terms  $C_{mis}$  and  $C_{in}$ , the NRD between both terms can be measured as follows:

$$NRD(C_{mis}, C_{in}) = \frac{\max\{\log f(C_{mis}), \log f(C_{in})\} - \log f(C_{mis}, C_{in})}{\log M - \min\{\log f(C_{mis}), \log f(C_{in})\}} \quad (2)$$

Where:

- $C_{mis}$  is a concept that is not captured in  $SR$ .
- $C_{in}$  is a concept that is captured in  $SR$ .
- $f(C_{mis})$  is the number of hits retrieved for  $C_{mis}$ .
- $f(C_{in})$  is the number of hits retrieved for  $C_{in}$ .
- $f(C_{mis}, C_{in})$  is the number of hits retrieved for  $C_{mis}$  and  $C_{in}$ .
- $M$  represents the number of indexed Web pages by the search engine.

In order to compute the similarity between the concepts captured in the constructed semantic networks, we employ one of the well-known and most commonly used distance function techniques, namely the Jaro-Winkler Distance function (WINKLER, 1999).

This function is formally defined as:

**Definition 4: Jaro-Winkler Distance function:**

As shown in Equation 3, this edit distance algorithm is usually used to measure the number of edit operations (addition, deletion, and replacement) required to transform one string ( $s$ ) into another ( $t$ ). In this context, and based on the number of required edit operations, we can judge whether two strings are similar or not. For instance, we can consider both concepts (programming) and (programing) as two equivalent concepts since the number of edit operations required to transform one string into another is very low.

$$JWinkler(s, t) = \frac{1}{3} \cdot \left( \frac{m}{|s|} + \frac{m}{|t|} + \frac{m - t'}{m} \right) \quad (3)$$

Where:

- $s$ : represents the first string/word
- $t$ : represents the second string/word
- $m$ : is the number of matching characters between  $s$  and  $t$
- $t'$ : is the number of transpositions

As we mentioned earlier in section 3.2.2, the produced lists of candidate concepts included some extremely common concepts that appeared to be of little value in helping find resumes matching their relevant job offers. To eliminate such concepts, we employ the following tf-idf weighting algorithm.

**Definition 5: Term frequency-inverse document frequency (tf-idf) Weighting algorithm:**

The tf\_idf weighting algorithm (Belkin and Croft, 1992) assigns a term  $t$  a weight  $w$  in a document  $d$ :

$$tf - idf_{t,d} = tf_{t,d}.idf_t \quad (4)$$

It is important to mention that we employ the tf-idf weighting algorithm at the corpus level in order to remove the set of concepts  $S(cs)$  that appear to be of little value among the set of candidate concepts  $A(cc)$  – obtained using the NLP pre-processing tools introduced in section 4.2.1. The set  $S(cs)$  is obtained based on a threshold value  $v$  using Equation 5.

$$S(cs, v) = \{ cs \in A(cc) \mid tf - idf(cs) \leq v \} \quad (5)$$

As we mentioned earlier, the development of the proposed system is carried out through three major phases. The implementation details of each phase is detailed and described as follows.

## 4.2 Development Details of the First Phase of the Prototype System

During this phase, we have implemented the first version of our proposed automatic recruitment system through incorporating multiple semantic resources and statistical-based techniques to match between candidate resumes and their relevant job offers. The prototype system comprised several modules organized according to their priority of execution as explained in the following sections.

### 4.2.1 Concept Identification and Extraction

First, we employed several NLP techniques to pre-process the content of resumes and job postings as detailed below:

- Resume/job post segmentation: the content of resumes/job postings (either in .pdf or in .doc format) is segmented into units (paragraphs or sentences) in order to prepare them for further processing and analysis.



- n-gram text tokenization: we carry out this step in order to tokenize the text in each unit into unigram, bigram and trigram tokens.
- Stop words removal: we pre-defined a list of stop words such as (a, the, we, in, as,...etc.) that have no semantic significance and do not contribute to the actual meaning of the derived units from the content of the resumes and job posts. Such stops words are automatically removed to enhance the system's performance and effectiveness during the matching process.
- Part-of-Speech Tagging (POS-Tagging): during this step, each extracted token is assigned to its part of speech category such as noun, verb, adjective, etc. For example, programmer is a noun. This step enables us to better understand the basic constituents of each unit and accordingly decide upon which tagged-entities to include in the matching process.
- Named Entity Recognition (NER): during this step, each token is assigned a named-entity category based on a set of pre-defined categories such as Person, Organization, and Location. For example, Ahmad is a person. Similar to the previous step, NER assists in distinguishing between the different types of named-entities and consequently including/excluding some entities based on the category that they belong to.

After applying the abovementioned NLP steps, lists of candidate concepts are identified and extracted. In this context, by “candidate concepts” we mean meaningful terms that are used to construct semantic networks and to accomplish the matching process between job offers and resumes. The next example clarifies the process of identifying and extracting candidate concepts based on utilizing the abovementioned NLP steps.

**Example 1: Concept Identification and Extraction (Phase 1).**

- Part of a Job Post (P1)

- We need a programmer.
- The programmer must have experience in Java programming language (j2ee).

- Part of a CV (CV1)

- I have studied Marketing.
- I Live in Java the island of Indonesia.

In this example, we have considered one segment (one unit) of both the job post and the resume. Accordingly, the text tokenization is performed and stop words are removed according to the pre-defined list of stop words. Then, both the POST and the NER steps are carried out using the Stanford CoreNLP suite (Manning et al., 2014). In the context of our work and after analyzing the produced results by both techniques, we have decided to include nouns (NNP, NNPS, NN) in the lists of candidate concepts. The results of applying these steps are shown in Table 2 below.

**Table 2. Results of applying the NLP steps**

<b>Candidate concepts extracted from job post (P1)</b>	<b>Candidate concepts extracted from resume (CV1)</b>
programmer	marketing
experience	java
java	island
programming language	indonesia
j2ee	

After identifying the lists of candidate concepts, we proceed further with constructing semantic networks based on those concepts as described in the next section.

#### 4.2.2 Construction of Semantic Networks

Once the lists of candidate concepts are identified and extracted, we submit them to WordNet semantic resource. Concepts (in addition to the semantic and taxonomic relations that may hold between them) that are recognized by this semantic resource are then used for automatically constructing semantic networks. Figures 5 and 6 illustrate the process of constructing semantic networks for P1 and CV1 (presented in Example 1 in section 4.2.1).

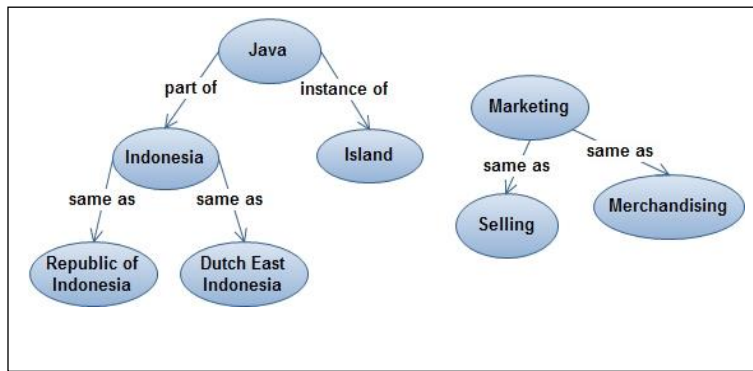


Figure 5. Semantic networks for CV1.

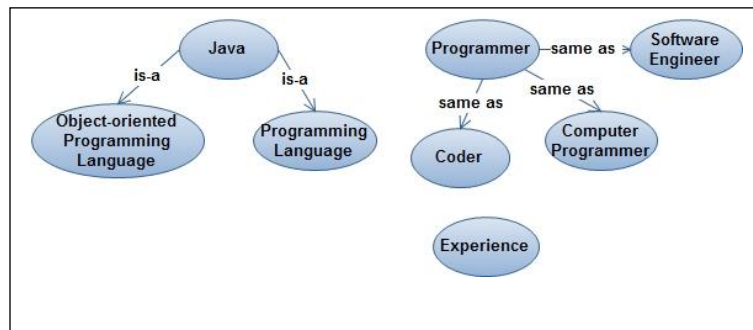


Figure 6. Semantic networks for P1.

When we search for the term “Java” in WordNet, we find it has three different senses (i.e. meanings) as follows:

1. Java -- (an island in Indonesia south of Borneo; one of the world's most densely populated regions)
2. Coffee, java -- (a beverage consisting of an infusion of ground coffee beans; "he ordered a cup of coffee")

3. Java -- (a simple platform-independent object-oriented programming language used for writing applets that are downloaded from the World Wide Web by a client and run on the client's machine).

Though this concept has three different meanings, it is obvious that we use the third meaning when constructing the semantic networks for job post (P1). This is because the rest of the meanings of the term “Java” (1 and 2) are not conceptually related to the concepts of job post P1. Accordingly, only the first meaning (1) should be considered in the construction of the semantic networks of CV1. To accomplish this task, we perform automatic Word Sense Disambiguation (WSD) (Agirre and Edmonds, 2006) of concepts to specify the correct sense for each term based on its surrounding textual content. Furthermore, the synonyms of each disambiguated term are also added to the produced networks.

The rest of the tokens that are not captured by WordNet such as "j2ee" are then submitted to the missing background knowledge handler as explained below.

#### ***4.2.3 Missing Background Knowledge Handler***

We utilize this module to tackle the problem of semantic knowledge incompletes in WordNet. Details of this module are described in the following sub-sections.

- Employing a Second Semantic Resource (YAGO2 Ontology)

Concepts that are not recognized by WordNet are further handled by YAGO2 ontology to address the problems of missing background knowledge and limited domain coverage that lie in WordNet. Accordingly, semantic relations that are defined in YAGO2 are exploited for enriching and expanding the initially produced WordNet-based semantic networks. However, using a second semantic resource may not solve the missing background knowledge problem since some concepts may still not be recognized by both

semantic resources. In this case, we employ our proposed statistical-based concept-relatedness technique (Kmail et al., 2015a) to judge whether the missing concepts can be suggested as enrichment candidates for the produced semantic networks.

- Statistical-based Concept-relatedness Techniqiue

For concepts that are not captured in both semantic resources, we employ the NRD concept-relatedness technique. We use this technique to decide whether such concepts can be proposed as enrichment candidates to expand the constructed semantic networks. In this scenario, the NRD algorithm computes the semantic closeness between concepts that are not captured by the used semantic resources and those that exist in the sematic networks. The following algorithm illustrates the process of computing the values of semantic closeness.

---

<b>Algorithm 1. NRD Algorithm</b>	
<b>Input:</b> Concepts that are not recognized in the used semantic resources ( $C_{mis}$ ) and concepts identified in them ( $C_{in}$ )	
<b>Output:</b> Values of semantic closeness between $C_{mis}$ and $C_{in}$	
1: $answer \leftarrow \langle \rangle$ ;	
2: for each $c_{mis} \in C_{mis}$	
3:   for each $c_{in} \in C_{in}$	
4: $answer \leftarrow \text{NRD}(c_{mis}, c_{in})$	
5:   end for	
6: end for	
7: <b>return</b> $answer$	

---

The NRD algorithm takes pairs of terms as input, and produces values that represent their semantic closeness as output. In Table 3 below, we show the values of semantic closeness between the concept “j2ee” and the rest of the job post's concepts: {programmer, experience, java, programming language}.

**Table 3. NRD values for the concept “j2ee”**

Concept	programmer	experience	Java	Programming language
Concept				
j2ee	0.71	1.15	0.69	0.64

#### 4.2.4 Extraction of Semantic Relations

In this algorithm, obtaining semantic closeness values by the NRD function is a prerequisite for finding the actual semantic relation(s) that may exist between semantically related concepts. Accordingly, to find such relations, we have defined a list of patterns in the same manner as proposed in (Maree and Belkhatir, 2015a, Maree et al., 2011). In this context, for each pair of semantically close concepts, the *Semantic Relation Extractor* (SRE) function retrieves the number of their hits after sending them including each of the patterns to a group of search engines. After this step, pairs of concepts are linked using the relations that achieve the highest values by the SRE function. For example, using the results computed by the NRD function, we were able to find that “j2ee” is semantically close to the terms "Java" and "programming language". But, we didn’t know the actual relation that may hold between them. Hence, we employed the SRE function by sending the following patterns as queries  $Q_i$  to a group of search engines:

- $Q1 = \text{"j2ee is a programming language", which outputs 3,724 hits result}$
- $Q2 = \text{"j2ee is a part of programming language", which outputs 0 hits result}$
- $Q3 = \text{"j2ee is same as programming language", which outputs 0 hits result}$

According to the number of hits retrieved for  $Q_i$ , relations in the patterns were proposed to enrich the job post semantic networks with the concept "programming language" as depicted in Figure 7.

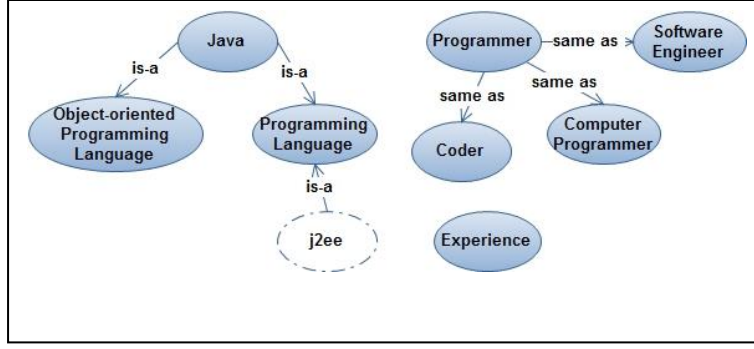


Figure 7. Enrichment of the job post (P1) semantic network

#### 4.2.5 Matching Semantic Networks

Once semantic networks are constructed, the matching process starts. During this process, semantic networks are matched according to the Jaro-Winkler distance function (formally defined in section 4.1). This function is utilized in algorithm 2 to find the measures of similarity between the semantic networks of the job post and the resume pair.

---

**Algorithm 2. Name-based technique for finding the similarity between the resume semantic network ( $SN_R$ ) and the job post semantic network ( $SN_J$ )**

---

**Input:**  $SN_R$  and  $SN_J$

**Output:** Measure of similarity based on correspondences set  $S$

---

```

1:  $answer \leftarrow \langle \rangle$ ;
2: for  $i \leftarrow 0$ ;  $i < SN_J.Length$ ;  $i++$ 
3:   for  $j \leftarrow 0$ ;  $j < SN_R.Length$ ;  $j++$ 
4:      $answer \leftarrow \mathbf{JWinkler}(SN_J[i] \ SN_R[j])$ 
5:     if( $answer < v$ ) then
6:       add( $SN_J[i], SN_R[j]$ ) to  $S$ 
7:     end if
8:   end for
9: end for
10: return  $similarity$ 

```

---

The algorithm measures the similarity between the compared strings (Line 4). When the similarity values is greater than 0.92 (tested and discovered empirically), then both strings are assumed to be equivalent and moved to the set of corresponding concepts  $S$  (Lines 5 and 6). For example, if we have the concept "Object Oriented Programming Language" in the semantic networks of the job post, and the concept "Object\_Oriented Programming Language" in the resume semantic networks, then using the Name-based

technique we find that they are equivalent. Hence, each resume is ranked according to the similarity between its semantic networks and the job post's semantic networks.

### **4.3 Development Details of the Second Phase of the Prototype System**

During this phase, new modules are integrated into the updated version (version 2) of the prototype system to achieve the following goals.

- a) Refine the lists of candidate concepts by removing concepts that have little value and appear to be of little significance in helping in the matching process.
- b) Enrich the constructed semantic networks with newly obtained concepts that were not initially recognized by the used semantic resources.

We present the details of the new/updated modules below.

#### ***4.3.1 Refinement of Candidate Concepts***

At this step, we define a list of pre-defined terms that appear to be of little value in helping find resumes matching their relevant job offers. Examples of those concepts are: contact info, address, date of birth, etc. In addition, we utilize the tf-idf weighting algorithm in order to identify and remove concepts that may have negative impact on the matching process. Accordingly, concepts that either belong to the list of extremely common concepts or have low tf-idf weights are removed from the lists of candidate concepts as illustrated in the following example.



### Example 2: Refinement of Candidate Concepts (Phase 2)

- Part of a Job Post (P2)

- We are seeking a programmer who is looking to take his experience to the next level.
- Our programmer is required to have 2+ years of experience in Java programming language (e.g. jsp).

- Part of a CV (CV2)

- I have worked as a Software engineer. And I have the following skills: java, j2ee, jsp, xml.

It is important to point out that prior to refining the lists of candidate concepts, we perform the NLP pre-processing steps to extract the lists of candidate concepts as described in section 4.2.1. The results of applying these steps are shown in Table 4 below.

**Table 4. Lists of candidate concepts**

Candidate concepts extracted from job post (P2)	Candidate concepts extracted from resume (CV2)
programmer	software engineer
experience	java
level	j2ee
java	jsp
programming language	xml
jsp	

After extracting candidate concepts, they are refined as shown in Table 5.

**Table 5. Lists of refined concepts**

Job post (P2) refined concepts list	Resume (CV2) refined concepts list
programmer	software engineer
java	java
programming language	j2ee
jsp	jsp
	xml

Once the lists of candidate concepts are refined, the semantic networks that represent the taxonomic and semantic relations between them are constructed as shown below.

#### 4.3.2 Construction of Semantic Networks – Updated Version

Unlike the first version of the prototype system, the second version employs two semantic resources during this module, namely WordNet and YAGO2 in order to construct semantic networks from the lists of refined candidate concepts.

##### - WordNet Semantic Resource

As mentioned earlier in section 4.2.2, each concept is submitted to WordNet in order to extract the semantic and taxonomic relations that hold between it and other concepts in the list of refined concepts. As a result of this step, semantic networks that represent resumes and job offers are constructed. Figures 8 and 9 depict the output of employing the semantic networks construction module.

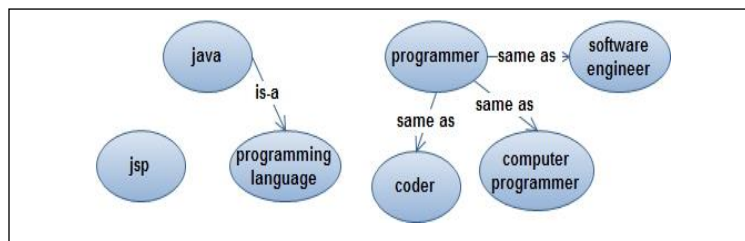


Figure 8. Semantic networks of job post (P2)

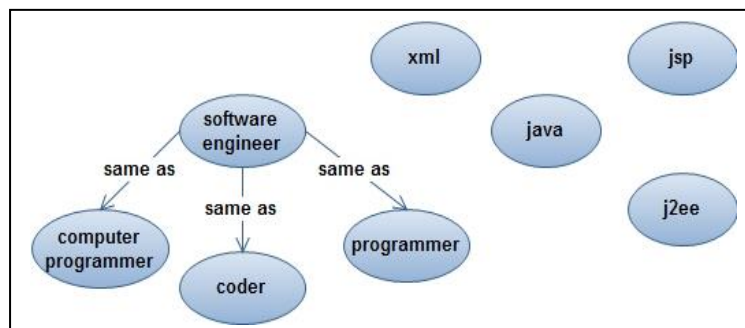


Figure 9. Semantic networks of resume (CV2)

The rest of concepts that are missing from WordNet ontology are then submitted to YAGO2 ontology.

#### - **YAGO2 ontology**

Concepts that are not defined in WordNet are submitted to YAGO2 ontology. Accordingly, semantic relations that are defined in YAGO2 are also exploited to expand the constructed semantic networks. However, we would like to point out that even using a second semantic resource like YAGO2 may not solve the missing background knowledge problem since some concepts such as "jsp" are not defined in it. Therefore, concepts that are not recognized in WordNet or in YAGO2 are further submitted to the missing background knowledge handler.

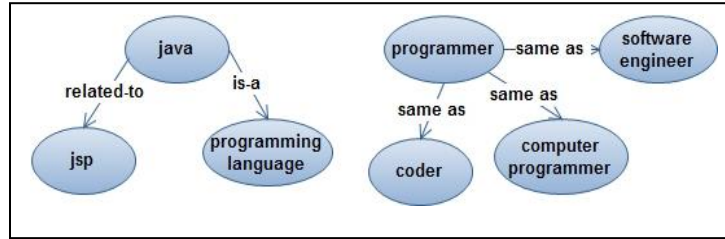
#### **4.3.3 Missing Background Knowledge Handler – Updated Version**

Unlike the first version of the prototype system, we exploited HiringSolved (HS) dataset to compensate for missing background knowledge presented in the used semantic resources. HS dataset defines a huge number of terms in the form of skills – either mentioned in job offers or resumes – and the weights of semantic relatedness between them.

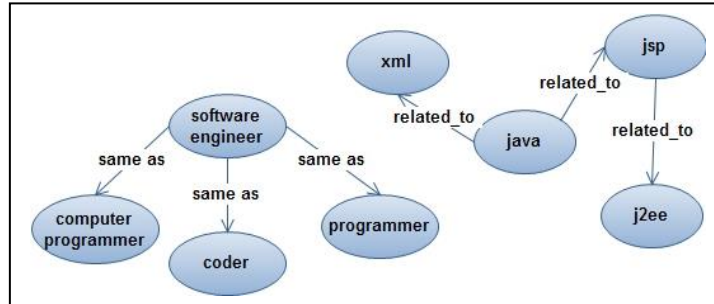
**Table 6. The result of submitting "jsp" to HS Dataset**

<b>Term</b>	<b>Relatedness measure</b>
servlets	1.00
j2ee	0.94
jdbc	0.92
tomcat	0.90
ejb	0.76
struts	0.75
hibernate	0.62
xml	0.60
java	0.56

For example, although the term "jsp" was not recognized by any of the exploited semantic resources, when we submit it to HS dataset we get a set of statistical-based semantically-relevant terms to this term as shown in Table 6. The weights shown in the table represent measures of semantic relatedness between the submitted term and its semantically-related terms. Following to this step and based on the results of applying the missing background knowledge handler, the semantic networks are updated and enriched as depicted in Figures 10 and 11.



**Figure 10. Updated semantic networks of job post (P2)**

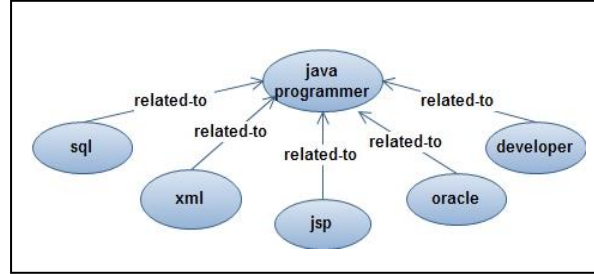


**Figure 11. Updated semantic networks of resume (CV2)**

As shown in Figures 10 and 11, concepts in the semantic networks of (P2) and (CV2) are connected with the newly obtained concepts from HS dataset. For instance, we can see that the degree of semantic relatedness between the terms “j2ee” and “jsp” is 0.94. We replace this semantic relatedness value by the “related-to” relation and use it to connect both concepts.

#### 4.3.4 Further Enrichment of the Semantic Networks

Semantic networks extracted from job offers represent the reference to which the semantic networks of the resumes are matched. In this context, and since some of the required skills may not be explicitly defined by the employer, we further enrich the semantic networks of the job offers by automatically adding new skills obtained from HS dataset. To carry out this step, we submit the job titles to HS dataset to obtain a set of related skills to each title. For instance, when submitting the job title ("java programmer") of job post (P2) to HS dataset, it returns the list of skills shown in figure 12. As highlighted in the previous section, we replaced the measures of semantic relatedness with the “related-to” relation and only considered the top 5 related skills returned by HS dataset.

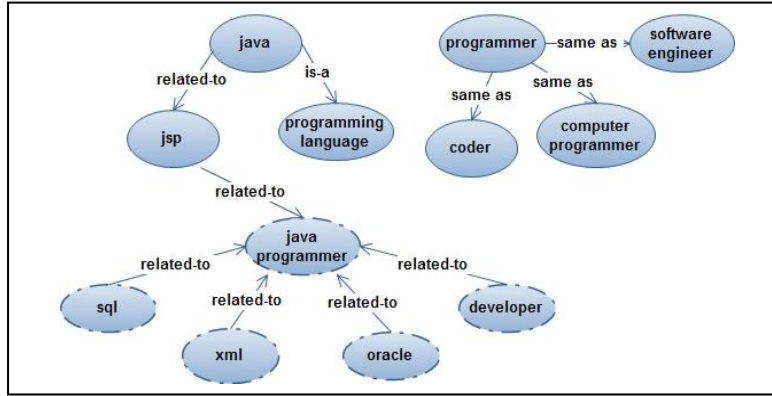


**Figure12 . Top 5 related skills returned by HS dataset to the title “Java Programmer”**

To enrich the semantic networks of job post (P2) with the elements of  $S(cr)$ , we follow the following procedure:

- If an element  $ce \in S(cr)$  already exists in the semantic networks of (P2), then we retain  $ce$  in its position in the networks. For example, since the element "jsp" is already defined in the semantic networks of (P2), we keep this element in its position in the network.

- If an element  $ce \in S(cr)$  does not exist in the semantic networks of (P2), then we update the networks by adding the job title as a new node, and then we attach it to all other elements of  $S(cr)$  that do not exist in the semantic networks of (P2).



**Figure 13. Enrichment of the semantic networks of job post (P2)**

Once the semantic networks of the resumes and job offers are constructed, they are matched in a similar manner as we did during the matching process (previously presented in section 4.2.5.) between the semantic networks of resumes and job offers in the first version of the prototype system.

#### **4.4 Development Details of the Third Phase of the Prototype System**

During this phase, feature extraction techniques are integrated into the third version of the prototype system in order to convert the original resumes and job offers into semi-structured documents. Accordingly, the system matches segments of resumes to their relevant segments of job offers instead of matching unstructured versions of resumes and job offers. In the following sections we present the details of the newly incorporated/updated modules.

##### **4.4.1 From Unstructured Resumes and Job offers into Semi-structured Documents**

During this module, unstructured resumes and job offers are converted into semi-structured documents based on employing feature extraction techniques. These

techniques include NLP techniques and rule-based regular expressions. As detailed earlier in section 4.2.1, the NLP steps include resume/job post segmentation, n-gram tokenization, stop words removal, POST and NER. In this module, we particularly improved the NER technique through expanding the defined rules to cover more entities. We have labeled the newly introduced entities as: “DEGREE”, “EXPFIELD” and “EDUCATION FIELD” through using the RegexNER from the Stanford CoreNLP suite. Examples of these rules are shown below.

The following example clarifies the process of converting unstructured resumes and job offers into semi-structured documents.

Bachelor of (Arts Laws Science Engineering)	DEGREE
PhD	DEGREE
Master of (Arts Laws Science Engineering)	DEGREE
M.Sc.	DEGREE
B.Sc.	DEGREE
Information Technology	EDUCATION FIELD
CS	EDUCATION FIELD
Computer Science	EDUCATION FIELD
Software engineer	EXPERIENCE
Java programming language	EXPERIENCE

**Example 3:** Converting unstructured resume and job post into semi-structured documents.

- Part of a Job Post (P3):

<p><b><i>What you need for this position</i></b></p> <ul style="list-style-type: none"> <li>• <i>3+ years of experience in Java programming language (e.g. jsp)</i></li> <li>• <i>Bachelor of Science in Computer Science.</i></li> <li>• <i>You should be a programmer who is looking to take his experience to the</i></li> </ul>
---

- Part of a CV (CV3):

**Key skills and experience**

Strong core Java, j2ee, jsp, xml development experience.  
 Ability to develop creative solutions for complex problems.  
 I have worked as a Software engineer for 2 years.

**Education**

B.Sc. in CS.  
 M.Sc. in CS.

In this example, we convert the unstructured segments of P3 and CV3 into semi-structured units as described below.

## Semi-structured version of CV3

```
<Applicant Info>
<Experience>
<Years>2</Years>
<Field>Software engineer</Field>
</Experience>
<Education>
<Degree> B.Sc.</Degree>
<Field>CS</Field>
</Education>
<Education>
<Degree> M.Sc.</Degree>
<Field>CS</Field>
</Education>
</Applicant Info>
```

## Semi-structured version of P3

```
<Job post Info>
<Experience>
<Years>3</Years>
<Field>Java programming language </Field>
</Experience>
<Education>
<Degree> Bachelor of Science</Degree>
<Field> Computer Science</Field>
</Education>
</Job post Info>
```

First we employ the regular expressions in order to identify job experience segment (paragraph or sentence). Some of the used regular expressions are shown below.

- 1: [0-9]+(\\-[0-9]+)?\\+? years .+ experience
- 2: .+? months .+ experience
- 3: work.+ .+ years

Then, the NLP techniques are performed to extract the specified number for the years of experience (tokens that its POST is “NUMBER” or “DURATION”) and the specified experience field (tokens that its NER is “EXPERIENCE”). After that, we identify



educational background info such as the educational degree (tokens that its NER is “DEGREE”) and the education field (tokens that its NER is “*EDUCATION FIELD*”).

Once resumes and job offers are converted into semi-structured documents, lists of candidate concepts are extracted and identified as described in section 4.2.1. The produced lists of candidate concepts are shown in Table 7.

**Table 7. Results of applying candidate concepts identification module.**

<b>Candidate concepts extracted from the segment of job post (P3)</b>	<b>Candidate concepts extracted from the segment of resume (CV3)</b>
programmer	Core
experience	development
java	Experience
programming language	Java
jsp	j2ee
level	Jsp
	Xml
	software engineer
	ability

After extracting candidate concepts, they are refined as detailed in section 4.3.1. The lists of refined concepts that pertain to P3 and CV3 are shown in Table 8.

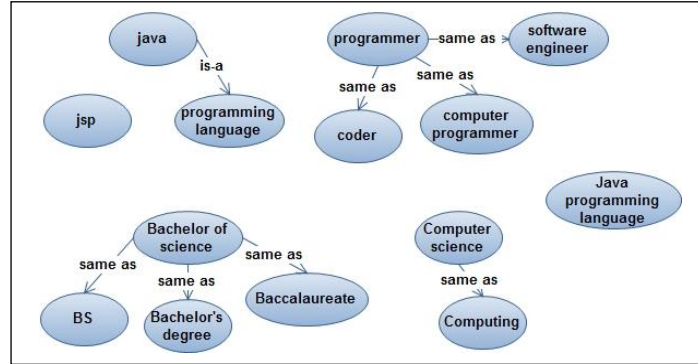
**Table 8. Lists of refined concepts**

<b>Job post (P3) refined concepts</b>	<b>Resume (CV3) refined concepts</b>
programmer	java
java	j2ee
programming language	jsp
jsp	xml
	software engineer

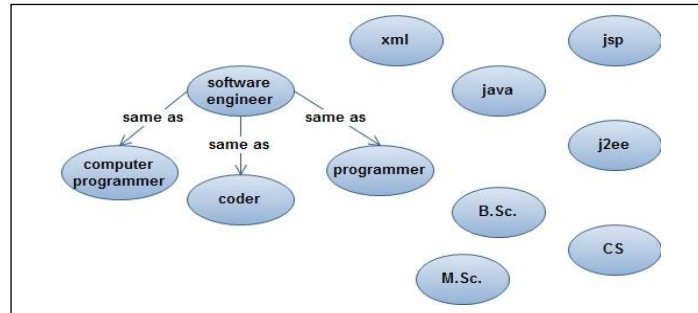
#### **4.4.2 Semantic Networks Construction – Updated Version**

In this section, we identify the details of constructing semantic networks that represent the lists of refined candidate concepts and the semi-structured documents. As described

earlier in section 4.3.2, each concept is submitted to WordNet ontology in order to extract the semantic and taxonomic relations that hold between it and other concepts. Figures 14 and 15 depict the output of performing semantic networks construction module.



**Figure 14. Job post (P3) semantic networks**



**Figure 15. Resume (CV3) semantic networks**

The rest of concepts that are missing from WordNet ontology are then submitted to YAGO3 ontology. It is important to point out that during this phase we have shifted from using YAGO2 to YAGO3 ontology due to the following reasons:

- It captures a broader number of local entities and facts.
- The precision of this ontology has been tested with a confirmed 95.03% of accuracy resulting in a better accuracy than that of YAGO2.

Accordingly, semantic relations that are defined in YAGO3 are also exploited to expand the constructed semantic networks. We would like to point out that even we have shifted to using YAGO3, we are still faced with the problem of missing concepts such as "jsp" in this ontology. Therefore, concepts that are not recognized in WordNet or in YAGO3 ontologies are submitted to the missing background knowledge handler as follows.

#### 4.4.3 Missing Background Knowledge Handler – Updated Version

Unlike the second version of the prototype system, the third version exploits two datasets, namely HiringSolved and O\*NET (USDOL, 2015) datasets. In this context, HiringSolved dataset is employed to compensate for missing background knowledge presented in the used semantic resources as described in section 4.3.3. While O\*NET is utilized to recognize concepts tagged in the produced semi-structured documents and that are not fully covered in the used semantic resources. It is important to point out that we have manually enriched this dataset with missing concepts to ensure broader domain coverage. A subset of this dataset is shown in Table 9.

**Table 9. Subset of manually enriched O\*NET dataset**

Term	Relation	Term
B.Sc.	same as	Bachelor of science, BSc, B.Sc, BS, Bachelors, Bachelor,,B.S.
M.Sc.	same as	Master of science, MSc, M.Sc, Master"s degree
CS	same as	Computer Science
SE	same as	Software Engineering
CSE	same as	Computer System Engineering
IT	Same as	Information Technology
Computer Science	related to	SE, CSE,IT
Computer Network Architect	related to	Network Analyst, Network Consultant, Network Engineer, Network Manager, Networking Systems and Distributed Systems Engineer, Systems Engineer, Telecommunications Analyst, Telecommunications Engineer

Based on the results of applying the missing background knowledge handler, the semantic networks are updated as depicted in Figures 16 and 17.

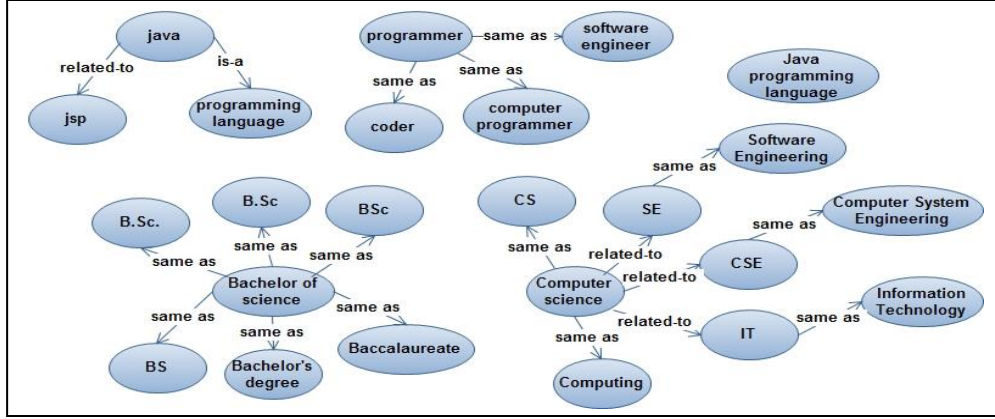


Figure 16. Updated semantic networks of job post (P3)

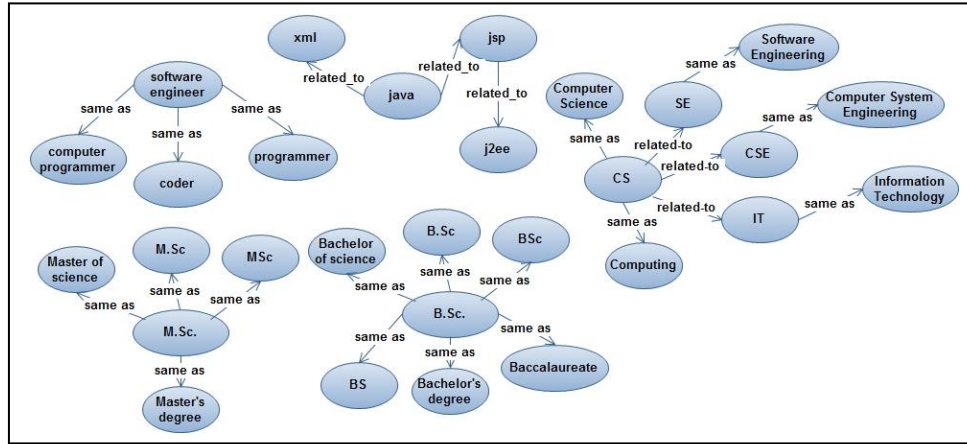


Figure 17. Updated semantic networks of resume (CV3)

#### 4.4.4 Matching the Semantic Networks

During the matching process, we use a multi-level matching algorithm to match between the semantic networks of resumes and job offers. Firstly, we match the semantic networks that represent the acquired/ required “educational background information”. Secondly, we match the semantic networks that represent job experience information in both resumes and job posts. And finally, we match the semantic networks of candidate concepts. In this context, we use Algorithm 2 to match between the semantic networks as described in section 4.2.5. This algorithm produces as output a correspondences set  $S$ . This set includes common concepts between the semantic networks of resumes and job posts and

it is further used to find relevance scores between each resume and its relevant job post based on Equation 6. This equation is an adapted form of the candidate's relevance scoring (RS) formula that has been proposed in the Oracle Project Resource Management (Management, 2010). The formula for calculating the scoring percentage is as follows:

$$RS = \frac{[(Candidate\ Concepts\ Match) * (Candidate\ Concepts\ Match\ Weighting) + (Educational\ Level\ Match) * (Educational\ Level\ Match\ Weighting) + (Job\ Experience\ Match) * (Job\ Experience\ Match\ Weighting)]}{[(Candidate\ Concepts\ Match\ Weighting) + (Educational\ Level\ Match\ Weighting) + (Job\ Experience\ Match\ Weighting)]} \quad (6)$$

In the context of our work, we use the following formula to assign automatic relevance scores:

$$RS = \frac{|Scc|}{|CCj|} * 70\% + \frac{|Se|}{|CEj|} * 15\% + \frac{|Sx|}{|CXj|} * 15\% \quad (7)$$

Where:

- RS: is the relevance score assigned between a job post and a resume.
- Scc: the correspondences set of candidate concepts.
- CCj: the candidate concepts of the job post.
- Se: the correspondences set of concepts that describe educational background information.
- CEj: the concepts that represent educational background information in the job post.
- Sx: the correspondences set of concepts that describe job experience information.
- CXj: the concepts that represent experience information in the job post.

It is important to point out that, the weighting values are variable and can be determined according to the employers' preferences. In the context of our work, we have assigned the following weighting values:

- Candidate concepts weight = 70%.
- Educational level weight = 15%.
- Job experience weight = 15%.

We would like to highlight that although the weighting values are variable, we have decided upon using the above mentioned values since they are the actual values that have been manually assigned during the phase of constructing our testing ground truth. This accordingly ensures conducting a fair evaluation of the effectiveness of the proposed system (i.e. when comparing the automatically generated relevance scores by the system to the manually assigned scores).

#### **4.5 Summary**

The aim of this chapter was to present the theoretical basis that lie behind our work, and the formal definitions and characterizations of the methods and techniques that are used in our system. Also, we have demonstrated that the development of the proposed system has been carried out over three major phases. The first phase incorporated five modules that mainly focused on extracting candidate concepts from the content of resumes and job posts, constructing semantic networks from the identified candidate concepts based on employing multiple semantic resources, enriching the produced semantic networks with newly obtained concepts that were not recognized by the used semantic resources, and finally, matching the enriched semantic networks. Although the produced matching results were satisfactory, new modules were integrated into the second and third versions

of the prototype system. We have discussed the reasons behind integrating those modules with the help of elaborating examples during this chapter. For instance, we have explained that the “*Refinement of Candidate Concepts*” module has been integrated into the second version of the prototype system in order to remove concepts that have little value in helping find resumes matching their relevant job offers and my negatively impact the matching process. While in the third version of the proposed system, we have integrated the “*From Unstructured Resumes and Job Posts to Semi-structured Documents*” module in order to convert the original unstructured format of resumes and job offers into semi-structured formats. In addition, we have discussed the foundations of the used relevance scoring formula, and more specifically, the assigned weights for each component of the formula, and the reasons behind assigning those weights.

## **5 Experimental Evaluation**

In this chapter, we describe the experiments that have been carried out to evaluate the techniques of the proposed system. The evaluation process has been accomplished at three successive stages. The first stage validates the first version of the prototype system. Then, in the second stage, we evaluate the newly incorporated modules in the second version of the prototype system and compare the results produced by this version of the system with one of the well-recognized state-of-the-art recruitment systems. And finally, in the third stage, we evaluate the effectiveness of the most recently updated version of our proposed system (version 3) and compare between the precision of the produced results when utilizing feature extraction techniques against not utilizing them in the matching process. We have implemented the prototype of the proposed system using Java programming language and conducted the experiments using a PC with core i5 CPU (2.1GHz) and (4 GB) RAM. The operating system is Windows 10.

The rest of this chapter is organized as follows. Section 5.1 presents the first stage of the experimental evaluation. The second stage of the system's evaluation is discussed in section 5.2. Section 5.3 presents the results of evaluating the third version of the proposed system. Concluding discussions on the conducted experiments are presented in Section 5.4.

### **5.1 Evaluating the Effectiveness of the First Version of the Prototype System**

In this section we present the experiments that we have carried out to evaluate the effectiveness of the first version of the prototype system. The effectiveness of the prototype system is evaluated based on how precise it is in automatically assigning relevance scores between candidate resumes and their corresponding job offers. To



accomplish this task, we have conducted a series of experiments on a dataset that consists of 100 resumes (downloaded from Amrood website<sup>2</sup>) and three job offers (obtained from Asal website<sup>3</sup>).

In order to provide a ground for evaluating the quality of the produced results, we manually identified all possible relevance judgments between the acquired resumes and their relevant job postings. We built expert judgments based on our knowledge and experience in the same fashion as presented in (Kessler et al., 2012). Then, we compared the manually assigned relevance scores to those automatically produced by the prototype system. We used the Precision/Recall (P/R) indicators in order to measure the quality of the produced results where:

$$P = \frac{|[\text{relevant resumes}] \cap [\text{retrieved resumes}]|}{|[\text{retrieved resumes}]|} \quad (8)$$

$$R = \frac{|[\text{relevant resumes}] \cap [\text{retrieved resumes}]|}{|[\text{relevant resumes}]|} \quad (9)$$

**Table 10. Precision and Recall Results**

<b>Job position</b>	<b>P</b>	<b>R</b>
QA engineer	71%	83%
Junior iOS developer	80%	100%
Senior java software engineer	60%	75%

As shown in Table 10, the first and the second job offers (i.e. QA engineer and junior iOS developer), focus on certain required technical skills such as (familiarity with Objective-c, Xcode, writing manual tests, etc.). And thus, it is possible to construct and match

---

<sup>2</sup> <http://www.amrood.com/resumelisting/listallresume.htm>

<sup>3</sup> <http://www.asaltech.com/careers/>

semantic networks without having the obstacles of extracting specific information related to experience and education. Consequently, the produced results show good precision and recall ratios.

However, the third job post has lower precision and recall values due to the fact that its job description includes the requirement of having 3+ years of professional java development experience. We would like to point out that we did not tackle this this type of requirements in the current version of the prototype system. However, this was among the pressing issues that required incorporating further feature extraction techniques in order to be able to effectively respond to such requirements specified by employers. To address this issue, we have incorporated a feature extraction module in the third version of the prototype system.

## **5.2 Evaluating the Effectiveness of the Second Version of the Prototype System**

To measure the effectiveness of the second version of the proposed system, we evaluated its precision in assigning relevance scores between job offers and applicant resumes. To accomplish this task, we have expanded the size of the dataset by including 500 resumes downloaded from Amrood website<sup>4</sup> and other local job portals, and using seven different job offers obtained from Monster<sup>5</sup>. The manually constructed dataset has a size of 35.5 MB of documents represented in different document formats such as (.pdf) and (.doc) and contains 1296360 words.

In order to carry out the experiments, we analyzed the corpus of resumes and job offers through employing the NLP techniques described in section 4.2.1. Then, we utilized statistical-based measures to refine the lists of candidate concepts. Next, we used the

---

<sup>4</sup> <http://www.amrood.com/resumelisting/listallresume.htm>

<sup>5</sup> <http://jobs.monster.com>

semantic resources to construct the semantic networks of job offers and resumes. Additionally, the constructed networks were further enriched based on HS dataset. And finally, the resulting networks were automatically matched and different relevance scores were produced by the system.

#### ***A. Experiments Using Expert Judgments***

In order to provide a ground for evaluating the quality of the results produced by the second version of the system, we manually calculated all relevance scores between each job post and its relevant resumes. Then, we compared the manually calculated scores to those produced by the system. In this context, we used the Precision (P) indicator in order to measure the quality of our results. This measure is defined as follows:

Precision (P): is the Percentage Difference between the manually assigned relevance scores (between each job post and its relevant resumes) and those automatically generated by the system.

$$P = \frac{|V_{manual} - V_{automatic}|}{\frac{V_{manual} + V_{automatic}}{2}} * 100\% \quad (10)$$

Where:

- $V_{manual}$ : is the manually assigned relevance score between each resume and job post.
- $V_{automatic}$ : is the automatically calculated relevance score between each resume and job post.

As shown in Table 11, for each job post, we compared between the manually assigned relevance score for each resume and its corresponding relevance score that is automatically produced by the system. We considered six resumes per job post. Each job

post requires a different set of qualifications. The first job post requires skills in java, jsp, jsf, html, and javascript, and five years of experience. The second job post divides the required qualifications into two categories: i) Obligatory: having 6-8 years of experience in developing web applications using .Net technologies (asp.net, c#, mvc vb.net, etc) and ii) Optional: having experience in jquery, vb Script, and ajax. The third job post focuses on Microsoft sql related skills.

**Table 11. Precision results of the automatically generated relevance scores**

Job post	Resumes	Manual score	Automatic score	P (%)
Programmer	IT-CRM	0.16	0.26	0.53
	IT-Programming-51	0.33	0.30	0.91
	Software Developer	0.30	0.30	1.00
	IT-Tele-Software	0.58	0.60	0.96
	IT-Programming-94	0.20	0.26	0.74
	Network Admin	0.10	0.13	0.74
.Net developer	IT-CRM	0.55	0.55	1.00
	IT-Programming-51	0.40	0.36	0.92
	Software Developer	0.60	0.44	0.70
	IT-Tele-Software	0.50	0.55	0.91
	IT-Programming-94	0.60	0.50	0.88
	Network Admin	0.20	0.27	0.71
Database developer	IT-CRM	0.37	0.33	0.89
	IT-Programming-51	0.21	0.14	0.60
	Software Developer	0.45	0.44	0.89
	IT-Tele-Software	0.50	0.48	0.96
	IT-Programming-94	0.40	0.40	1.00
	Network Admin	0.35	0.33	0.92

As we can see in Table 11, the manual scores that were assigned for each resume are very close to the automatically calculated scores by the system. This is due to the fact that we employ multiple semantic resources that represent the semantic aspects of resumes and job offers. Additionally, we exploited statistical concept-relatedness measures to compensate for missing background knowledge and to enrich the list of concepts that are extracted from the job offers with relevant concepts that were not recognized by the used semantic resources.

However, we can find that for some particular results the percentage difference was large. For example, when matching the second job post ".Net developer" and "Software Developer" resume, the difference is (0.30 i.e. 100% - 70%). This is because the job post ".Net developer" has optional requirements in its job description such as (having experience in jquery, vb Script and ajax). This optional requirement is not distinguished from other obligatory requirements by our system and thus the manual score for the resume is larger than the automatic score. In order to solve this problem, we plan to assign different weights for optional and obligatory requirements, and then use these weights in computing the relevance scores between job offers and resumes.

### ***B. Evaluating the System's Effectiveness When Utilizing the Statistical-based Techniques***

In this section, we compare between the produced results by the system when we utilize the statistical concept-relatedness measures against when only using the multiple semantic resources. We used the Precision/Recall (P/R) indicators in order to measure the quality of the produced results as defined in Equations 8 and 9.

**Table 12. P/R Results using/not using the statistical techniques**

Job post		P/R Results without using the statistical techniques		P/R Results using the statistical techniques	
		P	R	P	R
1	Programmer	0.41	1.00	1.00	0.83
2	Java software engineer	0.71	0.69	0.88	0.97
3	Database developer	0.35	0.83	0.90	0.75
4	Senior QA enginer	0.45	0.91	0.84	0.91
5	Software quality engineer	0.48	1.00	0.82	1.00
6	Senior database administrator	0.23	0.80	1.00	0.80

As shown in Table 12, we were able to achieve promising precision results for most of the job offers. Additionally, it was obvious that a significant improvement on the produced results was achieved when utilizing the statistical-based concept-relatedness techniques. This is because when using these techniques we were able to refine the lists of candidate concepts on the one hand, and further enrich them with more related concepts on the other.

### C. Comparison with State-of-the-Art Systems

In this section, we compare the results produced by our systems with EXPERT system (Kumaran and Sankar, 2013) which is one of the state-of-the-art semantics-based automatic recruitment systems.. Both systems were tested against the dataset obtained from <http://www.amrood.com/resumelisting/listallresume.htm>. To accomplish the comparison task, we used the Precision (P) and Recall (R) indicators as defined in Equations 8 and 9. Additionally, we used the F-measure indicator as defined in Equation 11.

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (11)$$

Table 13. P, R and F-measure results

System Indicator	Our system	EXPERT
P	0.91	0.89
R	0.88	0.93
F-measure	0.89	0.87

As shown in Table 13, our proposed system was able to achieve better results than EXPERT system. The reason behind this is that – unlike EXPERT system – we are exploiting multiple semantic resources to derive the semantic aspects of resumes and job offers. In addition, we utilize HiringSolved dataset to compensate for missing background

knowledge and to enrich job offers with skills that are not explicitly mentioned by the employer. It is important to mention that we are incorporating other features in the matching algorithm to improve the effectiveness of the third version of the proposed system. Accordingly, we integrate a features extraction module to extract features such as educational background and years of experience from applicants' resumes. We believe that incorporating these features will lead to improving the results produced by the system as detailed in the next section.

### **5.3 Evaluating the Effectiveness of the Third Version of the Proposed System**

To validate the effectiveness of the third version of our proposed system, we have conducted experiments on the same dataset that has been used to evaluate the effectiveness of the second version of the proposed system. In order to carry out the experiments, we started by converting unstructured resumes and job offers into semi-structured documents by identifying segments that describe the educational background and job experience information. Then, we analyzed the corpus of the resumes and job offers through employing NLP techniques as described in section 4.2.1. After that, we utilized multiple semantic resources to construct the semantic networks of resumes and job offers. Additionally, HS dataset was utilized to further enrich the constructed networks with additional concepts that were not recognized by the employed semantic resources. And finally, we have carried out the matching process through comparing parts (segments) of resumes to their relevant parts (segments) of job offers instead of matching the whole resume documents to the pool of job posts.

In this section, we discuss our experiments in terms of two different aspects. First, we discuss the experiments that we carried out in order to compare between the produced

relevance scores by our system when utilizing feature extraction techniques against when not using them. In other words, our aim of this step is to measure the impact of using features extraction techniques on the effectiveness of the proposed system. Second, we experimentally demonstrate the effectiveness of the proposed system (after integrating feature extraction techniques) in assigning relevance scores between job offers and their relevant resumes.

#### ***A. The Impact of Utilizing Feature Extraction Techniques on the Effectiveness of the Proposed System***

In this section, we compare between the results produced by the system when we utilize feature extraction (FE) techniques to extract the experience and educational background information against when not utilizing them. By this we mean that we compared the manually calculated scores to those produced by the system when considering feature extraction techniques and when only using candidate concepts identification modules.

**Table 14. The third version of the system results using/not using feature extraction techniques**

<b>Job post</b>	<b>Resumes</b>	<b>Manual scores</b>	<b>Automatic scores using FE techniques</b>	<b>Automatic scores without FE techniques</b>
Java Developer	IT-Programming	0.30	0.36	0.21
	IT-testing	0.30	0.29	0.14
	Network Admin	0.30	0.27	0.27
	IT-QA	0.38	0.45	0.30
	Software Engineer	0.26	0.19	0.04
	IT-CRM	0.15	0.18	0.18
Senior Test Engineer	IT-Programming	0.25	0.26	0.11
	IT-testing	0.46	0.5	0.35
	Network Admin	0.10	0.19	0.19
	IT-QA	0.45	0.46	0.31
	Software Engineer	0.15	0.22	0.07
	IT-CRM	0.10	0.11	0.11
Software Engineer	IT-Programming	0.38	0.40	0.25
	IT-testing	0.38	0.43	0.28
	Network Admin	0.30	0.23	0.23
	IT-QA	0.61	0.66	0.36
	Software Engineer	0.37	0.25	0.10
	IT-CRM	0.15	0.15	0.15



As shown in Table 14, we have three job posts, and for each job post we have six resumes. The first job post requires a java developer with the following characteristics: 5+ years of server side design and development experience, B.Sc. degree in Computer Science and knowledge in object oriented programming language such as (Java, C++), REST based web service development and http principles. The second job post requires 6+ years of professional experience related to system testing, Bachelor's degree in computer science or related field and knowledge and experience with tracking and testing tools such as Selenium, SoapUI, Remedy and Siebel. The third job post focus on looking for talented candidates with 6 years of software engineering experience responsibilities such as agile/iterative development methodologies (XP, SCRUM, etc.), object-oriented design and Java programming skills. As we can see in Table 14, the manual scores that were assigned for each resume are very close to those produced by the system when utilizing feature extraction techniques. For example, if we consider the first job post “Java Developer” and the fourth resume “IT-testing, that describes an applicant with Bachelor of Computer Applications degree (B.C.A) and 2.5 years of software testing experience”, we can see that the difference between the manually assigned score and the automatically generated score when utilizing feature extraction techniques is less than when not utilizing them. This is due to employing feature extraction techniques to convert the original resumes and job offers into semi-structured documents. Accordingly, the system matches segments of resumes to their relevant segments of job offers instead of matching unstructured versions of resumes and job offers.

Accordingly, the automatic score between “IT-testing” and “Java developer” is increased by 0.15 due to a match between the acquired and required educational background info.

However, for some particular results, integrating FE techniques doesn't affect the produced results. For example, when we consider the second job post "senior Test Engineer" and the third resume "IT-CRM that describes an applicant with master of computer applications degree and 2.2 years of experience in Client / Server based applications development and support", we can see that the automatic score equals the manually assigned score. This is due to the fact that there is no match between the required and acquired educational background and job experience info. And hence, capturing the experience and education information from the resume and job post doesn't affect (i.e. increase) the automatic score.

### ***B. Experiments Using Expert Judgments***

In this section, we evaluate the system effectiveness based on comparing the manually assigned relevance scores between resumes and their related job offers and automatically generated scores. In this context, we used the Precision (P) indicator in order to measure the quality of our results as defined in Equation 10.

**Table 15. Precision results using FE techniques**

<b>Job post</b>	<b>Resumes</b>	<b>Manual score</b>	<b>Automatic score</b>	<b>P (%)</b>
Java Developer	IT-Mobile	0.22	0.18	0.80
	IT-Systems	0.23	0.27	0.84
	Electronic eng	0.10	0.16	0.54
	IT_prog	0.30	0.30	1.00
Senior Test Engineer	IT-Mobile	0.10	0.07	0.65
	IT-Systems	0.10	0.16	0.54
	Electronic eng	0.10	0.03	0.54
	IT_prog	0.25	0.27	0.46
Database Developer	IT-Mobile	0.10	0.07	0.54
	IT-Systems	0.23	0.28	0.81
	Electronic eng	0.10	0.18	0.43
	IT_prog	0.24	0.23	0.96

As shown in Table 15, the manual scores that were assigned for each resume are very close to the automatically calculated scores by the system. This is because we have

integrated two new important factors (educational background and job experience info) in calculating relevance scores. These factors constitute 30% of the final result (relevance score). In addition, we have employed multiple semantic resources and statistical concept-relatedness measures to represent the semantic aspects of resumes and job offers and to further enrich them with concepts that are not recognized by the used semantic resources.

However, we can find that for some particular results the percentage difference was large. For example, when matching the second job post “Senior Test Engineer” and “Electronic eng” resume, the difference is (0.46 i.e. 100% - 54%). This is because the job post has optional requirements in its job description such as (having knowledge and experience with tracking tools such as Remedy, Siebel, or other industry standard). This optional requirement is not distinguished from other obligatory requirements by our system and thus the manual score for the resume is larger than the automatic score. In order to solve this problem, we plan to assign different weights for optional and obligatory requirements, and then use these weights in computing the relevance scores between job offers and resumes.

## **5.4 Summary**

In this chapter, we presented detailed discussion on the experiments that we have carried out to evaluate the effectiveness of the proposed online recruitment system on the one hand, and compared the produced results with one of the state-of-the-art systems on the other hand.

During this chapter, we divided the evaluation process into three stages. The first stage is concerned with evaluating the effectiveness of the first version of the prototype system.

Although the results show satisfactory precision and recall ratios, we also found that for particular job offers, the precision and recall results were not satisfactory due to missing background knowledge in the used semantic resources particularly for concepts that are related to “Required Skills” section. Additionally, the prototype system failed in identifying the experience and educational background info. The second stage evaluated the effectiveness of the second version of the prototype system, and it showed that employing HS dataset has led to significant enhancements on the produced results due to recognizing concepts that were not identified in the used semantic resources. However, for some job offers (those that require a specific number for the years of experience or a certain educational level) the precision of the system was not satisfactory. To address this issue, we incorporated a new Semi-structured Feature Extraction based Conversion module in the third version of the prototype system through which we attempt to extract such features from the content of resumes and job postings. Experimental evaluations for the produced matching results were satisfactory and closely related to the manually assigned relevance scores between the job offers and their relevant resumes.

## **6 Conclusions and Future Work**

In this chapter, we summarize our proposed approach for building an automatic online recruitment system and outline the future works and challenges related to exploiting semantic resources in building the proposed system. This chapter is organized as follows. Section 6.1 provides a summary for our research work and highlights the techniques and approaches that we employ in the proposed system. Section 6.2 discusses the future work and other challenges related to employing semantic resources in building online recruitment systems.

### **6.1 Conclusions**

Job-Resume matching task is one of the most important and challenging tasks for the Human Resources (HR) department in any organization (Strohmeier and Piazza, 2013). Accordingly, online recruitment systems are proposed to facilitate this challenge and to reduce the time, cost and effort required for matching resumes to their relevant job posts. Several techniques and approaches have been proposed to build online recruitment systems. However, these techniques and approaches suffer from limitations and drawbacks. For example, keyword-based techniques suffer from low precision ratios due to ignoring the semantic aspects of the entities encoded in the content of job offers and resumes. While the newer semantics-based approaches and machine learning algorithms suffer from drawbacks and limitations associated with the used resources (training data, ontologies, and knowledge bases), namely limited domain coverage and semantic knowledge incompleteness.

In this thesis, we had two main goals. First, we aim to tackle the low precision problem of automatic keyword-based recruitment systems and to address the problem of missing

background knowledge presented in online recruitment systems that employ a single semantic resource. Second, we aim to address the issue of the lack of publicly available datasets in the recruitment domain .To meet our two goals, we summarize our contributions as follows.

Our first contribution is the development of an automatic online precision-oriented recruitment system by combining multiple semantic resources, statistical-based techniques and feature extraction methods. Unlike traditional keyword-based online recruitment systems, our proposal derive the hidden semantic dimensions of concepts encoded in resumes and job offers through employing knowledge represented in multiple semantic resources. Although employing multiple semantic resources alleviated the problem of missing background knowledge when employing a single semantic resource, we found that those resources still suffer from semantic knowledge incompleteness particularly when it comes to the recognition of concepts that are mentioned in the “Required Skills” section of job offers. To address this issue, concept-relatedness measures presented in HiringSolved dataset are utilized to enrich the content of resumes and job offers with concepts that are not recognized by the used semantic resources. Moreover, to facilitate the matching task, feature extraction techniques are employed to convert unstructured resumes and job posts into semi-structured documents. In this context, the system matches segments of resumes to their relevant segments of job offers instead of matching unstructured versions of resumes and job offers.

The second contribution aims to construct a publically available dataset that comprises 500 resumes and seven job offers acquired from different online resources to address the issue of the lack of publicly available datasets in the recruitment domain.

## 6.2 Challenges and Future Work

Though the conducted experiments showed promising results, there are other remaining challenges and research problems that need to be addressed in the future work. Below we discuss these problems and outline proposals on how to address them in our future work:

- In chapter 4, we have explained how employing multiple semantic resources can address the problem of missing background knowledge and limited domain coverage presented in a single semantic resource. However, as pointed in that chapter, we were still faced with semantic knowledge incompleteness problem in the used semantic resources. To tackle this problem, we plan to exploit additional domain-specific semantic resources and classifications such as German Classification of Industrial Sector (WZ2003), North American Industry Classification System (NAISC), German version of the Human Resources XML (HR-BA-XML) and Skill Ontology developed by the KOWIEN Project. Utilizing these resources aims to ensure obtaining broader and deeper domain coverage of semantic knowledge and alleviate the problem of semantic knowledge incompleteness.
- Some job offers have “Optional Requirements” component in its job description. This component is not currently distinguished from other “Obligatory Requirements” component. However, when we convert resumes and job posts from unstructured documents into semi-structured documents both components fall under one category “Candidate Concepts”. Accordingly, we plan to separate this component into two different components and assign different weights for

each component, and then use these weights in computing the relevance scores between job offers and resumes.

- An equally important issue to evaluating the effectiveness of the proposed online recruitment system is to evaluate its performance. To do this, we plan to employ several optimization and normalization strategies to improve the run-time and efficiency of the utilized techniques in matching resumes to job offers.



## References

- AGIRRE, E. & EDMONDS, P. 2006. *Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology)*, Springer-Verlag New York, Inc.
- BELKIN, N. J. & CROFT, W. B. 1992. Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35, 29-38.
- BHAGDEV, R., CHAPMAN, S., CIRAVEGNA, F., LANFRANCHI, V. & PETRELLI, D. 2008. Hybrid search: effectively combining keywords and semantic searches. *Proceedings of the 5th European semantic web conference on The semantic web: research and applications*. Tenerife, Canary Islands, Spain: Springer-Verlag.
- BIZER, C., HEESE, R., MOCHOL, M., OLDAKOWSKI, R., TOLKSDORF, R. & ECKSTEIN, R. 2005. The impact of semantic web technologies on job recruitment processes. *Wirtschaftsinformatik 2005*. Physica-Verlag HD.
- BOURIGAULT, D. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. *Proceedings of the 14th conference on Computational linguistics - Volume 3*. Nantes, France: Association for Computational Linguistics.
- BREAUGH, J. & STARKE, M. 2000. Research on Employee Recruitment: So Many Studies, So Many Remaining Questions. *J. Manag.*, 26, 405-434.
- BREWSTER, C., O'HARA, K., FULLER, S., WILKS, Y., FRANCONI, E., MUSEN, M. A., ELLMAN, J. & BUCKINGHAM SHUM, S. 2004. Knowledge representation with ontologies: the present and future. *IEEE Intelligent Systems*, 72-81.
- CARROLL, M., MARCHINGTON, M., EARNSHAW, J. & TAYLOR, S. 1999. Recruitment in Small Firms: Processes, Methods and Problems. *Employee Relations*, 21, 236-250.
- CHUNG-KWAN, S., UI TAK, Y., HUY KANG, K. & SANG CHAN, P. 2000. A hybrid approach of neural network and memory-based learning to data mining. *Neural Networks, IEEE Transactions on*, 11, 637-646.
- CILIBRASI, R. L. & VITANYI, P. M. 2007. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19, 370-383.
- COLUCCI, S., DI NOIA, T., DI SCIASCIO, E., DONINI, F. M., MONGIELLO, M. & MOTTOLA, M. 2003. A formal approach to ontology-based semantic match of skills descriptions. *J. UCS*, 9, 1437-1454.
- DAN, C. A Hybrid Approach to Concept Extraction and Recognition-Based Matching in the Domain of Human Resources. In: JOHN, D., ed., 2004. 535-539.
- FALIAGKA, E., ILIADIS, L., KARYDIS, I., RIGOU, M., SIOUTAS, S., TSAKALIDIS, A. & TZIMAS, G. 2014. On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed CV. *Artificial Intelligence Review*, 42, 515-528.
- FALIAGKA, E., RAMANTAS, K., TSAKALIDIS, A. & TZIMAS, G. Application of machine learning algorithms to an online recruitment system. ICIW 2012, The Seventh International Conference on Internet and Web Applications and Services, 2012a. 215-220.
- FALIAGKA, E., RAMANTAS, K., TSAKALIDIS, A. K., VIENNAS, M., KAFEZA, E. & TZIMAS, G. An Integrated e-Recruitment System for CV Ranking based on AHP. WEBIST, 2011. 147-150.

- FALIAGKA, E., TSAKALIDIS, A. & TZIMAS, G. 2012b. An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet research*, 22, 551-568.
- FÄRBER, F., WEITZEL, T. & KEIM, T. 2003. An automated recommendation approach to selection in personnel recruitment. *AMCIS 2003 Proceedings*, 302.
- GARCÍA-SÁNCHEZ, F., MARTÍNEZ-BÉJAR, R., CONTRERAS, L., FERNÁNDEZ-BREIS, J. T. & CASTELLANOS-NIEVES, D. 2006. An ontology-based intelligent system for recruitment. *Expert Systems with Applications*, 31, 248-263.
- GRUBER, T. R. 1993. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5, 199-220.
- HIRINGSOLVED. 2015. *HiringSolved website* [Online]. Available: <https://hiringsolved.com/explorer>.
- HOFFART, J., SUCHANEK, F. M., BERBERICH, K., LEWIS-KELHAM, E., MELO, G. D. & WEIKUM, G. 2011. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. *Proceedings of the 20th international conference companion on World wide web*. Hyderabad, India: ACM.
- HONG, W., ZHENG, S., WANG, H. & SHI, J. 2013a. *A Job Recommender System Based on User Clustering*.
- HONG, W., ZHENG, S., WANG, H. & SHI, J. 2013b. A Job Recommender System Based on User Clustering. *Journal of Computers*, 8, 1960-1967.
- HU, X., ZHANG, X., LU, C., PARK, E. K. & ZHOU, X. 2009. Exploiting Wikipedia as external knowledge for document clustering. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Paris, France: ACM.
- KALFOGLOU, Y. & SCHORLEMMER, M. 2003. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18, 1-31.
- KESSLER, M., B, N., CHET, ROCHE, M., TORRES-MORENO, J.-M., EL-B, M. & ZE 2012. A hybrid approach to managing job offers and candidates. *Inf. Process. Manage.*, 48, 1124-1135.
- KESSLER, M., TORRES-MORENO, J. M., EL-B, M. & ZE 2007. E-Gen: automatic job offer processing system for human resources. *Proceedings of the artificial intelligence 6th Mexican international conference on Advances in artificial intelligence*. Aguascalientes, Mexico: Springer-Verlag.
- KESSLER, R., BÉCHET, N., TORRES-MORENO, J.-M., ROCHE, M. & EL-BÈZE, M. 2009. Job Offer Management: How Improve the Ranking of Candidates. In: RAUCH, J., RAŚ, Z., BERKA, P. & ELOMAA, T. (eds.) *Foundations of Intelligent Systems*. Springer Berlin Heidelberg.
- KMAIL, A. B., MAREE, M. & BELKHATIR, M. 2015a. MatchingSem: Online recruitment system based on multiple semantic resources.
- KMAIL, A. B., MAREE, M., BELKHATIR, M. & ALHASHMI, S. M. 2015b. An Automatic Online Recruitment System Based on Exploiting Multiple Semantic Resources and Concept-Relatedness Measures.
- KUMARAN, V. S. & SANKAR, A. 2013. Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping EXPERT. *Int. J. Metadata Semant. Ontologies*, 8, 56-64.

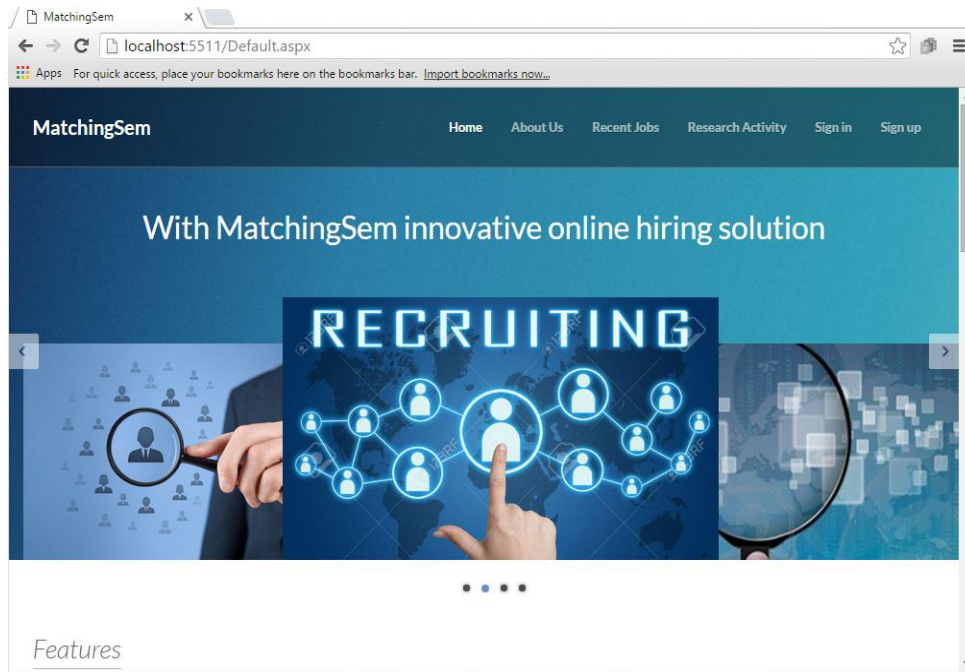
- LANG, S., LAUMER, S., MAIER, C. & ECKHARDT, A. Drivers, challenges and consequences of E-recruiting: a literature review. *Proceedings of the 49th SIGMIS annual conference on Computer personnel research*, 2011. ACM, 26-35.
- LAVRENKO, V. & CROFT, W. B. 2001. Relevance based language models. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. New Orleans, Louisiana, USA: ACM.
- LEE, I. 2007. An architecture for a next-generation holistic e-recruiting system. *Commun. ACM*, 50, 81-85.
- MAHDISOLTANI, F., BIEGA, J. & SUCHANEK, F. M. 2015. YAGO3: A Knowledge Base from Multilingual Wikipedias.
- MANAGEMENT, O. P. R. 2010. Available: [https://docs.oracle.com/cd/E18727\\_01/doc.121/e13493/T176493T176498.htm#I\\_cand\\_scorwght](https://docs.oracle.com/cd/E18727_01/doc.121/e13493/T176493T176498.htm#I_cand_scorwght) [Accessed 25/8/2015].
- MANNING, C. D., SURDEANU, M., BAUER, J., FINKEL, J., BETHARD, S. J. & MCCLOSKEY, D. 2014. The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.
- MAREE, M., ALHASHMI, S. & BELKHATIR, M. 2011. QuerySem: Deriving Query Semantics Based on Multiple Ontologies. *In: WANG, H., LI, S., OYAMA, S., HU, X. & QIAN, T. (eds.) Web-Age Information Management*. Springer Berlin Heidelberg.
- MAREE, M. & BELKHATIR, M. 2015a. Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies. *Knowl.-Based Syst.*, 73, 199-211.
- MAREE, M. & BELKHATIR, M. 2015b. Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies. *Knowledge-Based Systems*, 73, 199-211.
- MEDELYAN, O., MILNE, D., LEGG, C. & WITTEN, I. 2009. Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67, 716-754.
- MILLER, G. A. 1995. WordNet: a lexical database for English. *Commun. ACM*, 38, 39-41.
- MOCHOL, M., PASLARU, E. & SIMPERL, B. 2006. Practical guidelines for building semantic erecruitment applications. *International Conference on Knowledge Management, Special Track: Advanced Semantic Technologies (AST'06)*, 1-8.
- MOCHOL, M., WACHE, H. & NIXON, L. 2007. Improving the Accuracy of Job Search with Semantic Techniques. *In: ABRAMOWICZ, W. (ed.) Business Information Systems*. Springer Berlin Heidelberg.
- PANDE, S. 2011. E-recruitment creates order out of chaos at SAT Telecom. *Human Resource Management International Digest*, 19, 21-23.
- PARRY, E. & WILSON, H. 2009. Factors influencing the adoption of online recruitment. *Personnel Review*, 38, 655-673.
- RAMAR, K. & SIVARAM, N. 2010. Applicability of clustering and classification algorithms for recruitment data mining. *Int J Comput Appl* 4(5), 23-28.
- RAZIQ, A. & SHAIKH, A. 2015. Exploring recruitment & selection practices in Pakistani SMEs. *2015*, 4, 7.

- SENTHIL KUMARAN, V. & SANKAR, A. 2012. Expert locator using concept linking. *International Journal of Computational Systems Engineering*, 1, 42-49.
- SHIN, S., KO, J., EOM, S., SONG, M., SHIN, D.-H. & LEE, K.-H. 2015. Keyword-based mobile semantic search using mobile ontology. *J. Inf. Sci.*, 41, 178-196.
- SIVABALAN, L., RASHAD, Y. & NOR HASLINDA, I. 2014. *How to Transform the Traditional Way of Recruitment into Online System*, International Business Research; Mar 2014, Vol. 7 Issue 3.
- STROHMEIER, S. & PIAZZA, F. 2013. Domain driven data mining in human resource management: A review of current research. *Expert Syst. Appl.*, 40, 2410-2420.
- SUERDEM, A. & AKALIN, M. 2015. Using Conjoint Analysis to Determine the Requirements of Different Users for Designing Online Solution Tools: Job Matching Platform. In: BILGIN, M. H., DANIS, H., DEMIR, E. & LAU, C. K. M. (eds.) *Innovation, Finance, and the Economy*. Springer International Publishing.
- TRICHET, F., BOURSE, M., LECLERE, M. & MORIN, E. Human resource management and semantic Web technologies. Information and Communication Technologies: From Theory to Applications, 2004. Proceedings. 2004 International Conference on, 19-23 April 2004. 641-642.
- USDOL 2015. O\*NET Online.
- WINKLER, W. E. 1999. *The State of Record Linkage and Current Research Problems*, Internal Revenue Service Publication.
- YI, X., ALLAN, J. & CROFT, W. B. 2007. Matching resumes and jobs based on relevance models. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. Amsterdam, The Netherlands: ACM.

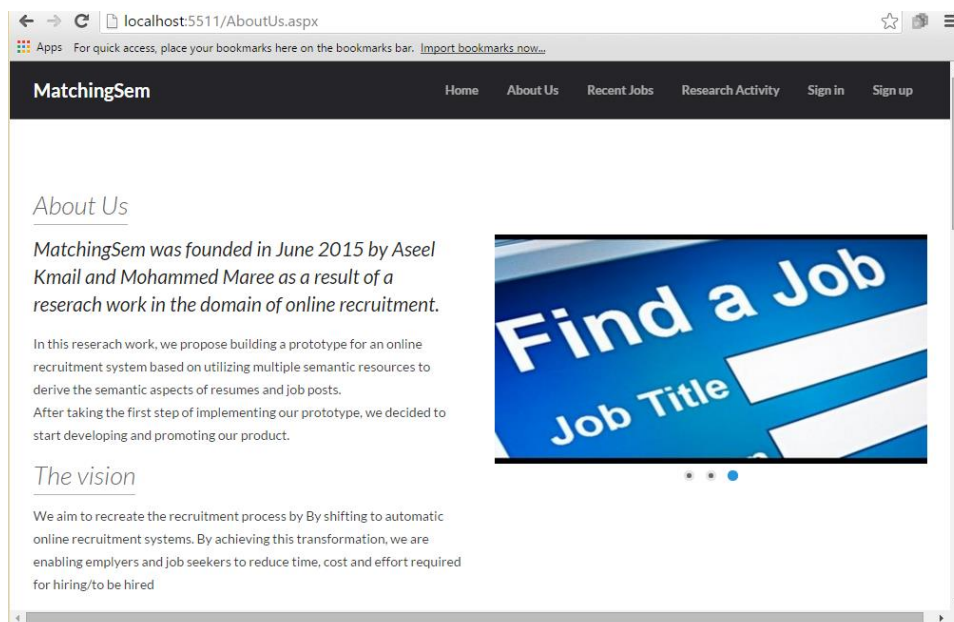
## Appendix A

### Screenshots of the online recruitment system

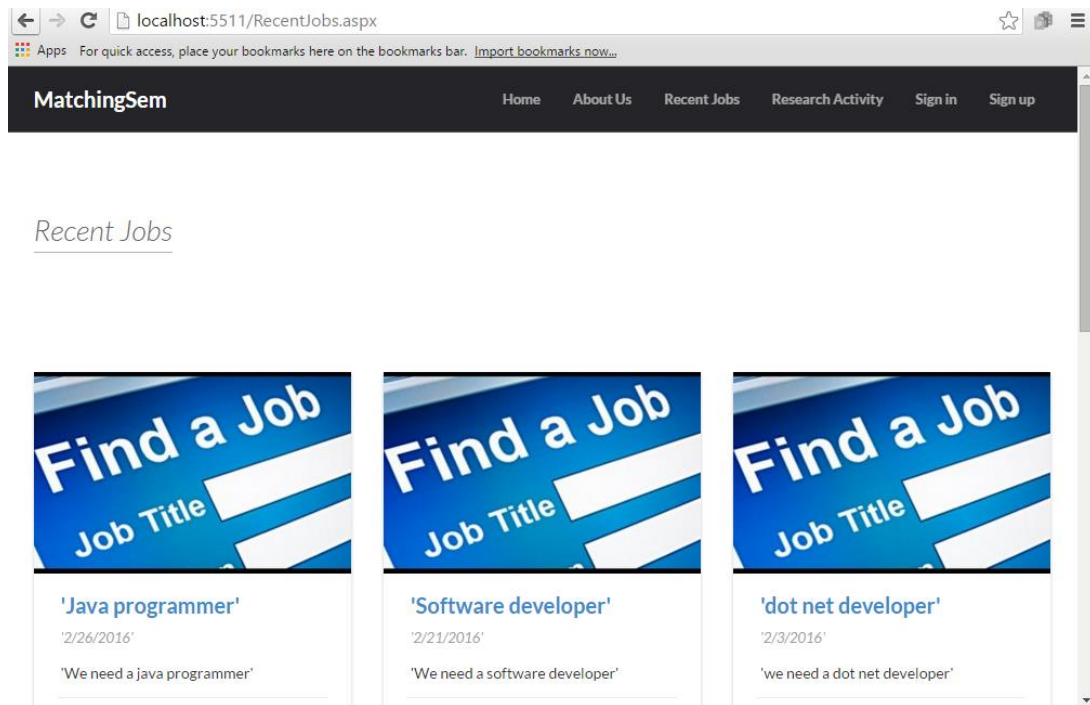
- The home page



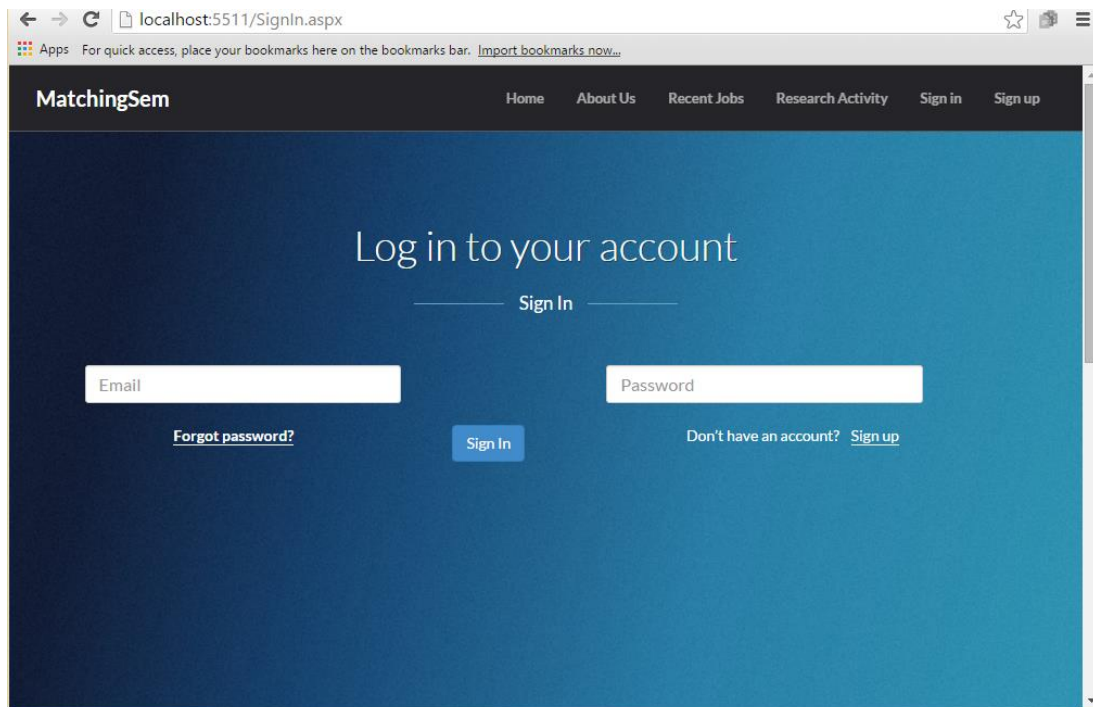
- About us page



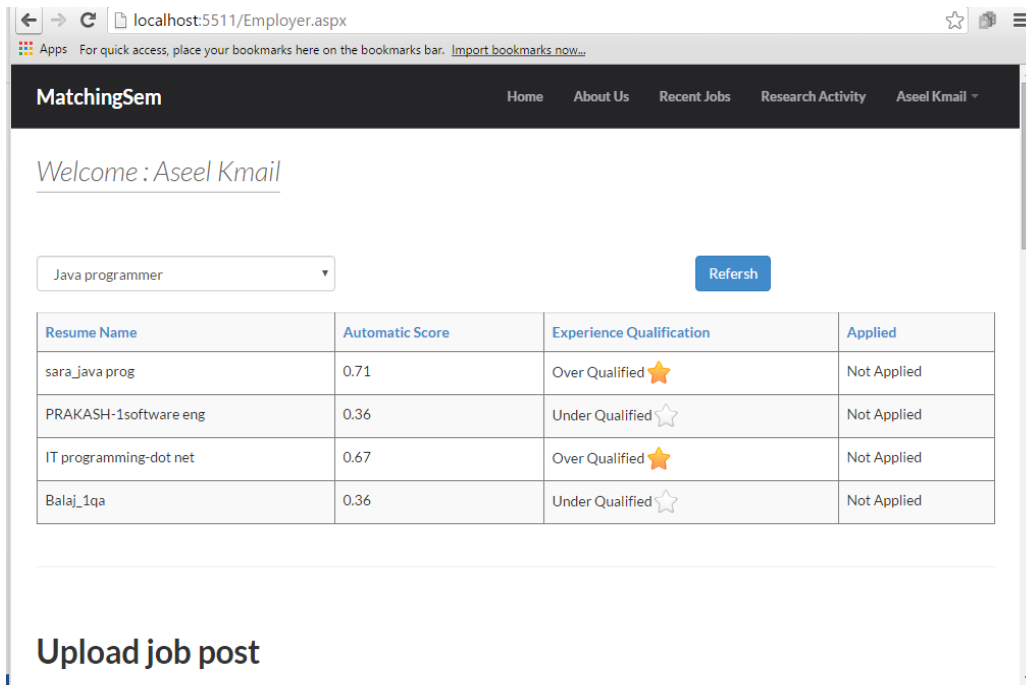
- **Recent jobs page**



- **Sign in page**



- **Employer page (view the automatically produced relevance scores)**

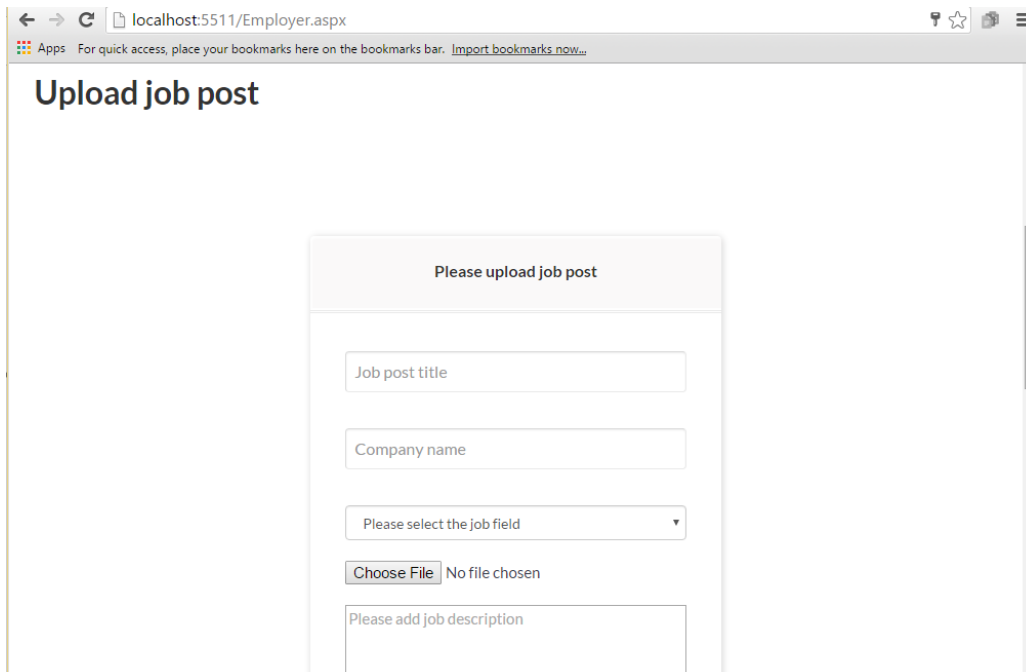


The screenshot shows the MatchingSem website interface. The browser address bar displays 'localhost:5511/Employer.aspx'. The page has a dark header with the site name 'MatchingSem' and navigation links: Home, About Us, Recent Jobs, Research Activity, and Aseel Kmail. Below the header, a welcome message 'Welcome : Aseel Kmail' is displayed. A search filter is set to 'Java programmer', and a 'Referesh' button is present. A table lists job matches with columns: Resume Name, Automatic Score, Experience Qualification, and Applied.

Resume Name	Automatic Score	Experience Qualification	Applied
sara_java prog	0.71	Over Qualified ★	Not Applied
PRAKASH-1software eng	0.36	Under Qualified ☆	Not Applied
IT programming-dot net	0.67	Over Qualified ★	Not Applied
Balaj_1qa	0.36	Under Qualified ☆	Not Applied

At the bottom of the page, there is a section titled 'Upload job post'.

- **Employer page (upload a new job post as unstructured document)**



The screenshot shows the 'Upload job post' section of the MatchingSem website. The browser address bar displays 'localhost:5511/Employer.aspx'. The page title is 'Upload job post'. A central form titled 'Please upload job post' contains the following fields:

- Job post title (text input)
- Company name (text input)
- Please select the job field (dropdown menu)
- Choose File button and No file chosen text
- Please add job description (text area)

- **Job seeker page (upload a CV as unstructured document)**

The screenshot shows a web browser window with the address bar displaying 'localhost:5511/JobSeeker.aspx'. The browser's bookmarks bar is visible with a link to 'Import bookmarks now...'. The website's header is dark with the 'MatchingSem' logo on the left and navigation links 'Home', 'About Us', 'Recent Jobs', 'Research Activity', and 'Lama Kmail' on the right. Below the header, a welcome message 'Welcome : Lama Kmail' is displayed. The main content area is titled 'Upload your CV'. It features a light gray box with the instruction 'Please upload your CV'. Inside this box, there is a dropdown menu labeled 'Please select the job field' and a file selection area with a 'Choose File' button and the text 'No file chosen'. Below the file selection area is a blue 'UPLOAD' button.

localhost:5511/JobSeeker.aspx

Apps For quick access, place your bookmarks here on the bookmarks bar. [Import bookmarks now...](#)

**MatchingSem** Home About Us Recent Jobs Research Activity Lama Kmail

*Welcome : Lama Kmail*

## Upload your CV

Please upload your CV

Please select the job field

Choose File No file chosen

UPLOAD



## الملخص باللغة العربية

يشهد العصر الحالي تطوراً سريعاً ومتلاحقاً في أسواق العمل وأساليب التوظيف المستخدمة فيها. ويأتي ذلك نتيجة لتزايد إعلانات الوظائف الشاغرة وأعداد المتقدمين لها. لذا، فقد أصبح من الصعب على أرباب العمل استقطاب واختيار الكوادر البشرية المناسبة لمتطلبات العمل. ووفق مقتضيات عصر السرعة، ظهرت الحاجة إلى استبدال وسائل التوظيف التقليدية بوسائل توظيف عصرية تدرج تحت مظلة "التوظيف الإلكتروني". والجدير بالذكر هنا أن فرق الباحثين والعلماء المعنيين في هذا المجال قد قاموا بتطوير أنظمة توظيف الكترونية تعتمد على تقنيات مختلفة. فالبعض من التقنيات المقترحة يعتمد على التوفيق بين إعلانات التوظيف والسير الذاتية للأشخاص الراغبين في العمل بناءً على الربط بين الكلمات المذكورة في كل منهما. وعليه، كلما ازداد عدد الكلمات المشتركة بينها، ازدادت درجة الملائمة بين متطلبات الوظيفة الشاغرة ومواصفات المتقدمين لها. أما البعض الآخر من التقنيات، فيقوم على استنباط مصطلحات جديدة (الكلمات المرادفة) لها علاقة بالكلمات المذكورة في كل من السير الذاتية وإعلانات التوظيف بناءً على استخدام مصادر دلالات المعاني (أو ما يعرف بالأنطولوجيا). ومن خلال استقراء أبعاد التقنيات المذكورة آنفاً وتشريح مضامينها، يتبين أنها ساهمت في تحسين مخرجات أنظمة التوظيف الإلكتروني بشكل عام. ولكن وعلى الرغم من ذلك، ما زالت هذه التقنيات تعاني من بعض المشاكل التي تتجسد في الاستراتيجيات المتبعة في بناء وتركيب البنية الهيكلية لهذه الأنظمة الحديثة. فمثلاً، نجد أن تقنية التوفيق التي تعتمد على استخدام الكلمات المذكورة في إعلانات التوظيف والسير الذاتية للمتقدمين محدودة في كفاءتها وتظهر نسب دقة متدنية؛ حيث أن نسبة كبيرة من المتقدمين للوظيفة تعتبر غير ملائمة لها لمجرد عدم ذكر المتقدمين نفس الكلمات الواردة في إعلان التوظيف. أما على صعيد تقنيات التوظيف التي تعتمد على مصادر دلالات المعاني، فهي أيضاً تعاني من مشاكل مرتبطة بمحدودية المصادر المستخدمة وعدم دقتها أحياناً في وصف المجالات التي من المفترض أن تشملها. من أجل حل المشاكل المذكورة آنفاً، فإننا نقترح في هذا البحث تطوير نظام توظيف إلكتروني يدمج ما بين استخدام مصادر دلالات المعاني (الأنطولوجيا) وتقنيات احصائية تُستخدم لقياس القوة الرابطة بين المفاهيم من أجل التوفيق بين متطلبات إعلانات التوظيف ومواصفات الأشخاص المتقدمين لها. وقد قمنا بالفعل ببناء أولي للنظام المقترح حيث اشتمل على التقنيات المقترحة، وقمنا أيضاً باختبار فعالية التقنيات المقترحة من خلال تنفيذ تجارب متعددة على مجموعة من إعلانات التوظيف والسير الذاتية. ونود أن نشير إلى أننا قد قمنا بنشر النتائج التي تم التوصل إليها في اثنتين من المؤتمرات

العلمية العالمية المعروفة والمصنفة في العام 2015، كما يجري العمل أيضا على تقديم بحث آخر الى أحد المجالات العلمية العالمية المحكّمة خلال هذا العام.