An Ontology-Based Information Extraction Approach for Résumés

Duygu Çelik[1,*] and Atilla Elçi[2]

[1] Computer Engineering Department, Istanbul Aydin University, Turkey
[2] Dept. Electrical-Electronics Engineering, Aksaray University, Turkey
duygucelik@aydin.edu.tr
atilla.elci@gmail.com

Abstract. A Curriculum Vitae (CV) or a résumé, in general, consists of personal information, education information, work experience, qualifications and preferences parts. Scanning or making structural transformation of the millions of free-formatted résumés from the databases of companies / institutions with human factor will result in the loss of too much time and human effort. In the literature, a limited number of studies have been done to change the résumés of the free-format to a structural format. The overall objective of the study is to infer required information such as user's experience, features, business and education from résumés of the potential user of a human resources system. In this article, we proposed an ontology driven information parsing system that is planned to operate on few millions of résumés to convert them structured format for the purpose of expert finding through the Semantic Web approach.

Keywords: Web Semantics, Web Ontology Language, Résumé, Semantic Search Agents, Information Extraction.

1 Introduction

When a résumé is mentioned, a written document comes to mind containing a person's education, work experience, skills, and personal information and so on. The résumés are created and sent as individual free-format, to make a job application via a brokerage firm or directly to companies which call for recruitment of staff. The purpose of the study is to elicit the requested information from each section, by dividing the résumés which may come from various sources on the Internet into the above sections with the semantic-based information extraction system and to make this information presentable in the form of relational data.

Owing to the fact that forming a general system which will be able to extract information from any kind of document is a very difficult application data mining systems is being developed specifically for the collapsed spans. Data acquisition which can be saved to the database from the *Résumés* or *Curriculum Vitaes (CVs)* written in free format is also one of the areas studied on. Therefore, Natural Language

[*] Corresponding author.

Q. Zu, B. Hu, and A. Elçi (Eds.): ICPCA-SWS 2012, LNCS 7719, pp. 165–179, 2013.

Processing (NLP) methods require a separate study for each language as the rules of language, semantic and grammatical structures are different from each other. Although, there are many studies on this issue in particular English language but not adequate study implemented such that has been found by considering concept-based semantics of résumés for English language in the literature. In fact, NLP algorithms without semantic consideration can be applied successfully however which can be more effective with semantic consideration if it also is applied on the résumé documents having semantic content.

On the database of Kariyer.net[1] company, there are more than 6,000,000 unstructured, written in both English and Turkish languages and free-style résumés as MS Word[2] documents. Therewithal the contents of the résumés are similar to their aims, the classification, the collection under the titles and the presentation of information can be totally different. Information gathering from each of these résumés to the existing system's database into a searchable form will be increasing the possible losses in terms of human factor. Therefore, there are many difficulties of résumé servicing by such governmental/commercial companies or unions, which consume too much their crucial resources such as time, capacity, human effort, money etc. Moreover, filtered résumés are the most important resource for employment in human resources departments of the governmental/commercial companies. Hundreds of applications can be made by mail online or fax and by hand to the ads of a member position that published by a company for its employees search. So, all the résumés other than online ones are filed and their chance of recall and being found becomes almost impossible. As the information contained in résumés is the only way for a company to employ suitable candidates, structurally organized and searchable résumés, regardless of the size of each are important for the company's human resources department. Other than the résumés uploaded by the candidates applying for a job, the request for evaluating the résumés accumulated in their hands over time and from different channels have emerged.

In the files and storage environments of companies making intensive recruitment, ten thousands of résumés are suspended without evaluation. In order for NLP algorithms to operate properly, intended scope must be determined in advance and the algorithm need to be developed specifically for this scope. With this study, development of a semantic-based information parsing system is considered which is first productized for the both languages English and Turkish language and which is targeted to have commercial value is presented. Therewithal formal analysis of the application prepared according to résumé scope, semantic analyzer part is designed in a way appropriate for the linguistics rules and structure for both languages separately and its adaptation with different scopes will be possible by this approach.

The rest of this paper is organized as follows: In Section 2 presents recent researches of IE applications on résumés in the literature. In Section 3 investigates the Ontology Knowledge bases used in the system. While Section 4 presents the architecture of the system through a case study, Section 5 is dedicated to conclusions.

[1] www.kariyer.net

[2] http://office.microsoft.com/tr-tr/word/

2 Related Works

Free format texts are being operated together with context interpretation and information debugging constitutes one of the major parts of information management systems. Semantic Web (SW)[3] provides to produce results appropriate with human communication from the information in the computer format or in contrast, the use of standardized methods for data generation from the resource documents written in free style. In fact, the SW is an extension of current Web technology that was developed in order to share and integrate information not only in natural language, but also by the associated software to be understood, interpreted and can be expressed in a way, so that makes easier to find the required data in the software. Additionally, ontologies are considered one of the pillars of the SW which can be developed through ontology languages that are a type of knowledge representation used for describing the concepts, properties and relationships among concepts of a domain.

Table 1. DOAC And Résumé RDF Comparison

Ontology Structure	*DOAC*	*Résumé RDF*
Classes/Concepts	15	16
Properties/Relationship s	17	73
Year	2005	2002

A SW based word treasure (vocabulary) can be considered as a special form of (usually light-weight) ontology, or sometimes also simply as a set of URIs with an (usually informally) described meaning. Recently, many ontology languages have been proposed and standardized such as RDF(S)[3] , Web Ontology Language (OWL)[4] and its new version OWL 2.0[5] etc. The OWL expresses the concepts in a specific space, terms and features in the form of ontology. In this way, it is possible to adapt the heterogeneous information from the distributed information systems. Additionally, each described concept in ontology encapsulates a subset of instance data from the domain of discourse [1].

A résumé, in general, consists of personal information, education information, work experience, qualifications and preferences parts and is kept somewhere and used during job interview. Scanning or making structural transformation of the millions of free-formatted résumés from the company/institution repositories with human factor will result in the loss of too much time and human effort. In addition, a limited number of studies have been done to change the résumés of the free-format to a structural format. Recently, most of the researches were designed for résumé ontology/vocabulary in English and some of them are presented below:

Bojārs and Breslin proposed Résumé RDF ontology in order to identify résumés semantically. The purpose of this ontology was designed for the systems which reveal

[3] `http://www.w3.org/RDF/`, 1999

[4] W3C Recommendation, `http://www.w3.org/tr/owl-features/`, 2004.

[5] W3C Recommendation, `http://www.w3.org/TR/owl2-overview/`, 2009.

the structure of “authority finder”. With this ontology structure, it is possible to describe information about people, features and skills etc. semantically [2].

Another similar study is Description of a Career (DOAC). In this study, a vocabulary is suggested by R. A. Parada to describe résumés [3]. In DOAC concepts about information, features, capabilities or skills of people were depicted. However, a limited number of concept descriptions were done. In the Résumé RDF ontology basic topics like, jobs, academic knowledge, experiences, skills, publications, certificates, references and other information were discussed. In the Résumé RDF two different namespaces were described that are:

```
http://purl.org/captsolo/résumé-rdf/0.2/cv
http://purl.org/captsolo/résumé-rdf/0.2/base
```

In the first namespace, there are a variety of concepts about résumés (such as work experience, education etc.). In the second namespace (base), the values of the general characteristics of résumés are defined (such as BS, MS, PhD or foreign language level etc.). However, the DOAC is in the form of word treasure (vocabulary) rather than ontology. With this vocabulary generally “professional skills” features about people are defined that has a namespace and is given in the below:

```
http://ramonantonio.net/doac/0.1/personal
```

Both of them can define a person's résumés and can describe experience of the capabilities of that person and similar information such as personal information of that person can be described with the help of common facilities or important differences. However, as in the Résumé RDF a greater number of properties (73) are defined, in information extraction semantically, more information can be obtained from queries. In contrast with the Résumé RDF, in the DOAC less semantic feedback can be achieved as a smaller number of properties (17) are defined.

Another research has been proposed by E. Karaman and S. Akyokuş [4] that is also IE based system model that can do the finding dismantling operation from résumés in four different sequential steps. These are called as *Text Segmentation, Scanning and Identifying Name Property, Classifying Name Property,* and *Text Normalization*. In the proposed system by E. Karaman and S. Akyokuş, a set of résumés written in the English language are considered to parse through a syntactic-based matching algorithm instead of the semantic matching [5-8]. While matching, the system matches the extracted information from a document with the predefined résumé vocabulary. For example, the system can see that the abbreviation *"IAU"* is *"Istanbul Aydin University"* but it can’t understand the semantics of ‘It is a university’, in other words "IAU is a university".

To combine several different methods as a hybrid approach from some studies are also discussed in the literature [9-12]. For example, in K. Yu, G. Guan and M. Zhou [13] studies HMM modeling is presented with a statistical approach, also, J. Piskorski and his group’s [14], SVM modeling is presented with learning based approach [15].

In this study, the proposed ontology driven information extraction system that is called ***Ontology-based Résumé Parser (ORP)*** will be operated on few millions English language and Turkish language résumés to convert them ontological format. The system will also assist to perform the expert finding/discovery and aggregation of skill information among résumé repository through its involved semantic approach.

The *Ontology Knowledge Bases (OKBs)* of the ORP system contain many domain ontologies each that contain concepts, properties and relationship among them for each type of résumé segments (Section 4.1). Moreover, OKBs is generated in English but also contains `<Literal xml:lang="tr">` declaration that is created for indicating the Turkish equivalent concept of each English concept in the OKBs that is shown in below OWL example.

```
<Declaration>
<Class IRI="#LanguageSkill"/>
<Literal xml:lang="en">Language Skill</Literal>
<Literal xml:lang="tr">Dil Becerileri</Literal>
</Declaration>
```

The OKBs provide to associate along concepts, relations and properties that may profit to solve some special cases of relationships among concepts (i.e. a Résumé Ontology in the OKBs is required for the both segments: work experiences and education segments, since a person's work place or company/institution information can be a university who works for and also studies in the same university).

3 Ontology Knowledge Bases (OKBs)

The system uses its own ontologies that are *Education, Location, Abbreviations, Occupations, Organizations, Concepts* and *Résumés Ontologies* in its OKBs. The functionality and the purpose of OKBs with its ontologies are described below and also the parentheses are used to keep Turkish meanings of the ontologies;

- *Education Ontology (Eğitim Ontolojisi –EO):* It keeps the concepts of comprehending the words of the education domain in Turkish such as 'University' (Üniversite), 'High School' (Lise) etc. and some properties among concepts such as '*hasDegree'* i.e. 'Honour degree' (Onur Derecesi) that is related for the education domain. The ontology also keeps the individuals that are instances of entire education institutes such as *'Istanbul Aydin University'*. The IAU is an individual of the 'University' concept in the ontology. Beside this, each concepts of this ontology indicates a relational concept in the Résumé Ontology. i.e. the system can understand the mean of the term 'University' in a candidate's résumé that is parsed from the education/work/other segments through the pre-declared concept 'University' in the Education Ontology.
- *Location Ontology (Konum Ontolojisi –LO):* It keeps the concepts of the location domain in Turkish such as 'Country' (Ülke), 'City' (Şehir), 'Village' (Köy) etc also some properties are declared such as '*hasPostalCode*' property of a city. The ontology is also associated some other ontologies through URIs such as Résumé Ontology, Concepts Ontology so on.
- *Abbreviations Ontology (Dil Kısaltmaları Ontolojisi –AO):* The ontology for abbreviations of certain word groups in the English or Turkish Languages such as *'Istanbul Aydin University'* as a single concept and also keeps its properties such as

'hasAbbreviation' property value is IAU. The ontology is also associated the Education Ontology, Concepts Ontology, Occupations Ontology so on.

- *Occupations Ontology (Meslekler Ontolojisi –OCCO):* It contains entire concepts of the used terms for the types of occupations in Turkish such as 'Doctor' (Doktor), 'Chef Assistant' (Şef Yardımcısı), 'Academician' (Akademisyen) and also includes their relations such as *'subClass'* i.e. 'Chef Assistant' is a subclass of the concept 'Chef'.
- *Organizations Ontology (Organizasyonlar Ontolojisi –OO):* contains entire concepts for the types of Companies/Institutions in Turkish language such as 'Pastry-shop' (Pastane), 'University' (Üniversite) or 'Hospital' (Hastane) so on.
- *Concepts Ontology (Kavramlar Ontolojisi –CO):* It contains common concepts in general such as 'Date' (Tarih), 'Year' (Yıl), 'Month' (Ay), 'Day' (Gün) or 'Currency' (Para Birimi) and their relationships. The ontology almost has association entire OKBs in the system.
- *Résumés Ontology (Özgeçmiş Ontolojisi –RO):* While there is no single correct format or style for writing a résumé in English or Turkish as well as other languages, therefore we can summarize the following types of information that are generally used terms by a résumé owner, and typically organized through ontologies for processing and understanding by machines in the following way as listed in Section 4.1.

Effective detailing of the OKBs will help understanding of the mechanics of the system. The most important one is Resume Ontology that is only explained with detail in the following section and also discussed the association ship among others.

3.1 Résumé (Özgeçmiş) Ontology

The Résumé (Özgeçmiş) ontology, the RO, is developed in order to express on the semantic data contained in a résumé, such as personal information, business and academic experience, skills, publications, certifications, etc. A résumé individual is described in ontology form via a résumé upper ontology. It is possible to annotate semantically the information of personal, work experiences, academic or educational life, skills, taken courses, certificates, publications, personal/professional references and other information in the résumé of the person.

Personal information keeps many concepts such as 'Name', 'Address', 'Email', 'Mobile', 'Home Phone', 'Military State' etc of a person. The 'Education' concepts contains the sub concepts such as 'dissertation', 'certificates', 'fellowships/awards', 'areas of specialization', 'areas of research', 'teaching interests', 'teaching experience', 'research experience', 'publications/presentations', 'related professional experience' and their properties. The education segment in the RO is mostly associated with the Education Ontology –EO.

The ***'Work Experiences'*** involves the semantic annotations for current work, previous works and goal works, namely, the information personal work preferences for future. Furthermore, the RO is designed with querying in mind and is able to extract a better semantic information from résumés e.g. 'Company' is a concept in the RO, may related to the person's current work, previous work or targeted work that are annotated via a ‘ro:employedIn’ (ro:Gecmis_is) property used for work history, ‘ro:isCurrentWork’ (ro:halen_is) used for the company information of current work, and ‘ro:isGoalWork’ (ro:Hedef_is) used for information of the person's target.

Additionally, the ***Skill*** is also considered in the RO and that is designed as in Resume RDF [5]. The skill data can be described semantically by ‘ro:skillName’ (ro:yetenekAdı) concept in RO. In addition, skill levels is also semantically described through owl:objectProperty that is named as ‘ro:skillLevel’ (ro:yetenekSeviye) from ‘bad’ (kötü) to ‘excellent’ (mükemel). Moreover, the ‘ro:skillLastUsed’ (ro:yetenekSonKullanım), ‘ro:skillHasCertificate’ (ro:yetenekSertifika) and ‘ro:skillYearsExperience’ (ro:yetenekYılSayısı) are used since it allows to quantify of skill levels particularly on foreign languages or used software tools.

Table 2. A Small Portion of the Résumé (Özgeçmiş) Ontology-RO

Personal Information
Education
Dissertation
Fellowships/Awards
Areas of Specialization
Research and Teaching Interests
Teaching Experience
Research Experience
Publications/Presentations
Certificates
Related Professional Experience
Work Experiences
Current Work
Previous Works
Skill
Language
Used Computer Software
Driving License
Activities
References
Other
.
.

```
<!—A portion of the Résumé Ontology in English Language-->
<owl:Class rdf:ID="Resume"/>
<owl:Class rdf:ID="Employee"/>
<owl:Class rdf:ID="Company"/>
<owl:Class rdf:ID="Skill"/>

<owl:Class rdf:ID="Software_Skill">
    <rdfs:subClassOf rdf:resource="#Skill"/>
</owl:Class>
<owl:Class rdf:ID="Driving_Skill">
    <rdfs:subClassOf rdf:resource="#Skill"/>
</owl:Class>
<owl:Class rdf:ID="Language_Skill">
    <rdfs:subClassOf rdf:resource="#Skill"/>
</owl:Class>

<owl:ObjectProperty rdf:about="#hasSoftware_Skill">
    <rdfs:range rdf:resource="#Resume"/>
    <rdfs:domain rdf:resource="#Tool"/>
      <owl:inverseOf>
        <owl:ObjectProperty rdf:about="#isSoftware_SkillOf"/>
      </owl:inverseOf>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:ID="#hasDriversLicense">
     <rdfs:domain rdf:resource="#Resume"/>
     <rdfs:range  rdf:resource="&xsd:boolean"/>
</owl:DatatypeProperty>

<owl:ObjectProperty rdf:ID="WorksInCompany">
    <rdfs:domain rdf:resource="#Employee"/>
    <rdfs:range rdf:resource="#Company"/>
        <owl:inverseOf rdf:resource="#CompanyMembers"/>
</owl:ObjectProperty>
```

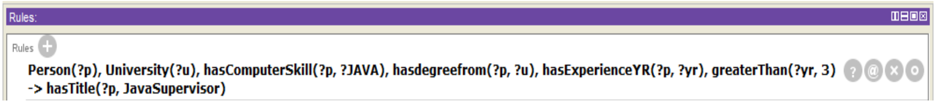
Furthermore, the ontology uses literals to describe skills in the form of ***owl:datatypeProperty*** since avoid uncertainty of skill identification and enabling straight skill matching. The concept ***‘skill’*** has many subclasses for language skill, driving skill, software skill, tool or machine skill etc. and allows to semantically describing if a person has foreign language ability, a driver license or used softwares/tools/machines and their levels respectively. The semantic declaration by the

RO will assist to perform the expert finding/discovery and aggregation of skill information among résumé repository. As shown in Table 2, a portion of the upper ontology of Résumé Ontology is depicted.

In fact, a résumé is characterized by its skills. The hasSoftware_Skill and isSoftware_SkillOf relations join the two classes together through a bidirectional link. They are inverseof one another, so they have inverted domain and range. The owl:inverseOf construction can be used to define such an inverse connection between relations through the *owl:ObjectProperty* (Lines 15-21, Table 2). Skill assertion should be considered since people mostly have mentioned their software/language/driving skill information on their résumés.

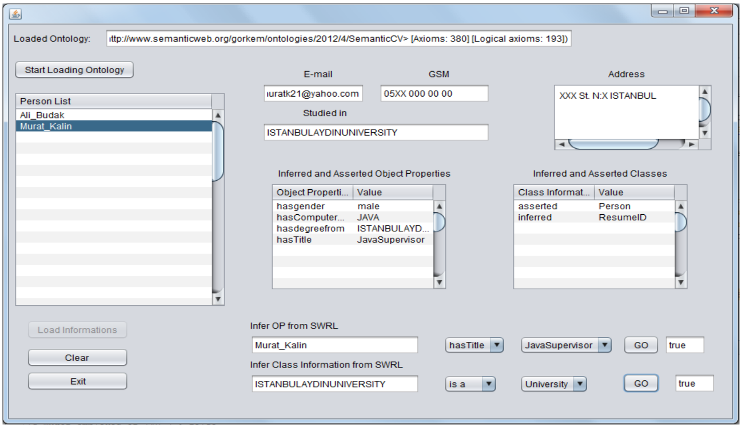
To the addition of the contributions of the RO, these kind of semantic declarations provide to infer another meaningful data from pre-asserted data for a résumé through an inferencing mechanism. The mechanism is able to give an opportunity to discover an appropriate résumé for a specific job position of a company.

Through semantic based inferencing rules declaration based on the resume ontology, a person with a university degree who has *Java knowledge* and has experience more than *three years* will be suitable candidate for some companies that are looking for a *'Java Supervisor'*. The conceptualizations of a person and a university can be captured from OWL classes called *'Person' (p)* and *'University' (u)*. The experience year, degree and computer skill conditions can be expressed from *hasExperienceYR, hasdegreefrom (if p has degree from university) and hasComputerSkill (if p has JAVA knowledge)* object properties. This rule could be written as:



Person(?p), University(?u), hasComputerSkill(?p, ?JAVA), hasdegreefrom(?p, ?u), hasExperienceYR(?p, ?yr), greaterThan(?yr, 3)
-> hasTitle(?p, JavaSupervisor)

Fig. 1. Rule syntax form semantic rule tab of Protégé ontology tool [16]



Loaded Ontology: ıttp://www.semanticweb.org/gorkem/ontologies/2012/4/SemanticCV> [Axioms: 380] [Logical axioms: 193])

Start Loading Ontology

Person List
Ali_Budak
Murat_Kalin

E-mail: ıuratk21@yahoo.com
GSM: 05XX 000 00 00
Address: XXX St. N:X ISTANBUL
Studied in: ISTANBULAYDINUNIVERSITY

Inferred and Asserted Object Properties

Object Properti...	Value
hasgender	male
hasComputer...	JAVA
hasdegreefrom	ISTANBULAYD...
hasTitle	JavaSupervisor

Inferred and Asserted Classes

Class Informat...	Value
asserted	Person
inferred	ResumeID

Infer OP from SWRL
Murat_Kalin | hasTitle | JavaSupervisor | GO | true

Infer Class Information from SWRL
ISTANBULAYDINUNIVERSITY | is a | University | GO | true

Fig. 2. System finds a person with a university degree who has *Java knowledge* and has experience more than *three years*. Also it infers Istanbul Aydin University is a University.

Through the inferencing mechanism of the system, a significant information stack is obtained from a résumé that can be stored separately in OWL form through the RO upper ontology model. Therefore, with a résumé scanning module, it can be possible to get information whether the résumé is eligible for a required criterion. A semantic search agent can easily access in its semantic descriptions and examine the suitability the résumé. In next section, the working mechanism of the system on a sample free formatted résumé is analyzed as a case study.

4 Working Mechanism of ORP

The system detects and extracts the desired information from the unstructured résumé after using its ontology parsing mechanism applied on the terms of the résumé and the concepts in OKBs. After that, the system transfers the parsed unstructured data from the focused résumé to the suitable place of the system's database and also structures it in the OWL form to a separate personal résumé ontology file.

The ORP system performs five major steps that are *converting an input résumé*, *dividing the résumé to some specific segments*, *parsing meaningful data from the input résumé*, *normalize it* and finally *applying classification and clustering task* to structure the résumé. The steps are described through a case study in below:

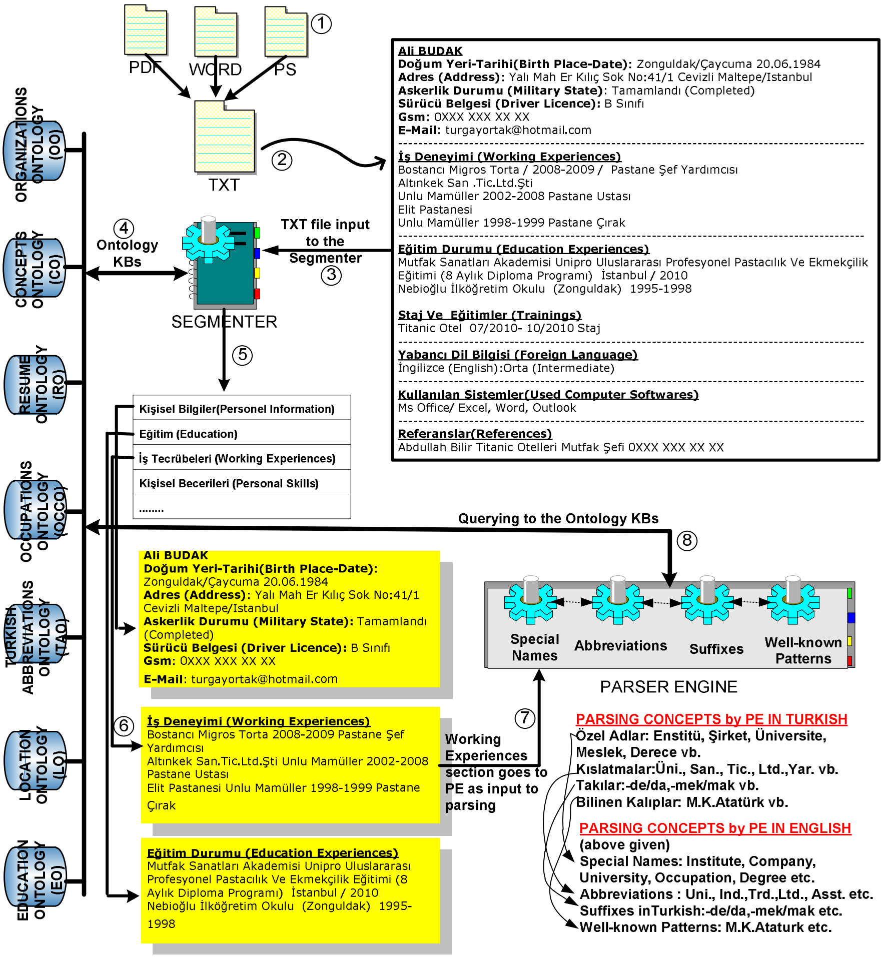
-Converter: It will carry out the process of converting for the given as .doc, .docx, .pdf, .ps etc. form of résumés into plain text. In case study, as a first step a free formatted sample résumé (i.e. Mr. Ali Budak résumé on the right in Figure 2) is presented to the system as an input (.doc, .pdf, .txt etc.) which will be converted to plain text format through a converter (for example: .txt, Figure 2 step 1 to 2).

-Segmenter: With the help of a Segmenter, the produced plain text will be segmented into parts like personal information, education, work experience, personal experience etc. Then, the necessary parts will be cut out and send to the ORP Parser Engine. The semantic-based segmentation process of the working experiences segment of the considered case study is shown in the figure below (Figure 2 step 3 to 4).

Table 3. Eng/Tr Translatıon of the Given Résumé Above (For Fıgure 1)

SEGMENTS	*ONROLOGY KNOWLEDGE BASES*	*A SAMPLE RÉSUMÉ GIVEN ABOVE*
Tur ish/Eng ish	**Turkish/English**	**Turkish/English**
Kişisel Bilgiler/Personal Information	Eğitim Ont./Education Ont.	'Bostancı' is a location in Istanbul city.
Eğitim/Education	Yer Ont./Location Ont.	'Migros' is a market company in Turkey.
İş Tecrübeleri/Working Experience	Kıslatmalar Ont./Abbreviation Ont.	'Torta' is an ignored unnecessary data.
Kişisel Becerileri/Personal Skills	Meslekler Ont./Occupations Ont.	'Pastane' is an organization that is pastry-shop.
	Özgeçmiş Ont./Resume Ont.	'Şef Yardımcısı' is a title of occupation that is called Assistant of Chef.
	Kavramlar Ont./Concepts Ont.	'Altınkek' is a name of a company.
	Organizasyonlar Ont./Organizations Ont.	'San.' is abbreviation of the industry that same as 'Ind.' in English and same for others: 'Tic.' is Trade 'Trd.', 'Ltd.' is Limited 'Ltd.' and 'Şti.' is company 'Co.' so on...

As shown in the Figure 2, the input résumé (belong to Mr. Ali Budak) is transferred to the Segmenter in the format of plain text (Step 3), the system separates the segments (such as personal information, work experiences, education etc.) by using its OKBs (Step 4 and 5) that are shown above in yellow boxes (Step 6). During segmentation, the ORP Segmenter takes a number of sample terms from the résumé to differentiate particular segments of the résumé.


PDF
WORD
PS
①
TXT
②
ORGANIZATIONS ONTOLOGY (OO)
CONCEPTS ONTOLOGY (CO)
RESUME ONTOLOGY (RO)
OCCUPATIONS ONTOLOGY (OCCO)
TURKISH ABBREVIATIONS ONTOLOGY (TAO)
LOCATION ONTOLOGY (LO)
EDUCATION ONTOLOGY (EO)
④ Ontology KBs
TXT file input to the Segmenter ③
SEGMENTER
⑤
Kişisel Bilgiler(Personel Information)
Eğitim (Education)
İş Tecrübeleri (Working Experiences)
Kişisel Becerileri (Personal Skills)
....
Querying to the Ontology KBs
⑧
⑥
Special Names
Abbreviations
Suffixes
Well-known Patterns
PARSER ENGINE
⑦
Working Experiences section goes to PE as input to parsing

Ali BUDAK
Doğum Yeri-Tarihi(Birth Place-Date): Zonguldak/Çaycuma 20.06.1984
Adres (Address): Yalı Mah Er Kılıç Sok No:41/1 Cevizli Maltepe/Istanbul
Askerlik Durumu (Military State): Tamamlandı (Completed)
Sürücü Belgesi (Driver Licence): B Sınıfı
Gsm: 0XXX XXX XX XX
E-Mail: turgayortak@hotmail.com

İş Deneyimi (Working Experiences)
Bostancı Migros Torta / 2008-2009 / Pastane Şef Yardımcısı
Altınkek San .Tic.Ltd.Şti
Unlu Mamüller 2002-2008 Pastane Ustası
Elit Pastanesi
Unlu Mamüller 1998-1999 Pastane Çırak

Eğitim Durumu (Education Experiences)
Mutfak Sanatları Akademisi Unipro Uluslararası Profesyonel Pastacılık Ve Ekmekçilik Eğitimi (8 Aylık Diploma Programı) İstanbul / 2010
Nebioğlu İlköğretim Okulu (Zonguldak) 1995-1998

Staj Ve Eğitimler (Trainings)
Titanic Otel 07/2010- 10/2010 Staj

Yabancı Dil Bilgisi (Foreign Language)
İngilizce (English):Orta (Intermediate)

Kullanılan Sistemler(Used Computer Softwares)
Ms Office/ Excel, Word, Outlook

Referanslar(References)
Abdullah Bilir Titanic Otelleri Mutfak Şefi 0XXX XXX XX XX

Ali BUDAK
Doğum Yeri-Tarihi(Birth Place-Date): Zonguldak/Çaycuma 20.06.1984
Adres (Address): Yalı Mah Er Kılıç Sok No:41/1 Cevizli Maltepe/Istanbul
Askerlik Durumu (Military State): Tamamlandı (Completed)
Sürücü Belgesi (Driver Licence): B Sınıfı
Gsm: 0XXX XXX XX XX
E-Mail: turgayortak@hotmail.com

İş Deneyimi (Working Experiences)
Bostancı Migros Torta 2008-2009 Pastane Şef Yardımcısı
Altınkek San.Tic.Ltd.Şti Unlu Mamüller 2002-2008 Pastane Ustası
Elit Pastanesi Unlu Mamüller 1998-1999 Pastane Çırak

Eğitim Durumu (Education Experiences)
Mutfak Sanatları Akademisi Unipro Uluslararası Profesyonel Pastacılık Ve Ekmekçilik Eğitimi (8 Aylık Diploma Programı) İstanbul / 2010
Nebioğlu İlköğretim Okulu (Zonguldak) 1995-1998

PARSING CONCEPTS by PE IN TURKISH
Özel Adlar: Enstitü, Şirket, Üniversite, Meslek, Derece vb.
Kıslatmalar:Üni., San., Tic., Ltd.,Yar. vb.
Takılar:-de/da,-mek/mak vb.
Bilinen Kalıplar: M.K.Atatürk vb.

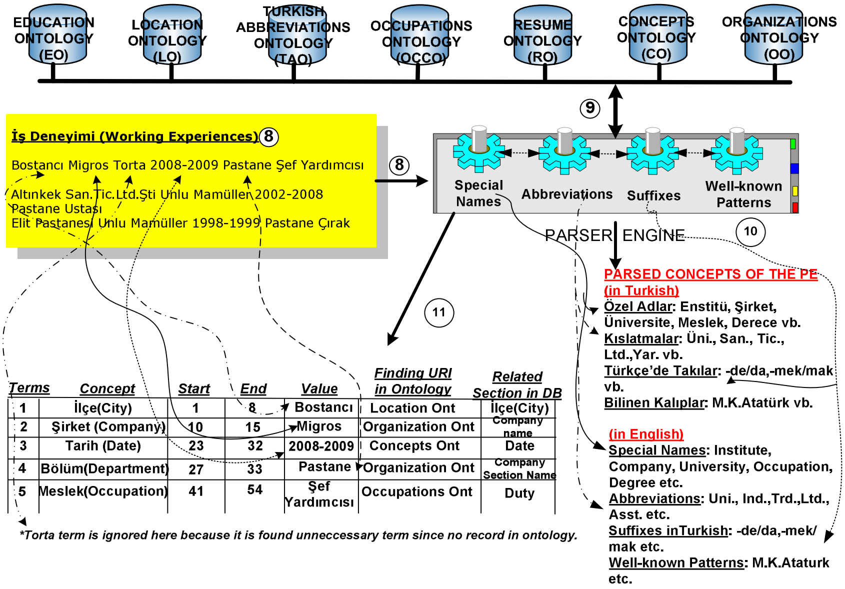
PARSING CONCEPTS by PE IN ENGLISH
(above given)
Special Names: Institute, Company, University, Occupation, Degree etc.
Abbreviations : Uni., Ind.,Trd.,Ltd., Asst. etc.
Suffixes inTurkish:-de/da,-mek/mak etc.
Well-known Patterns: M.K.Ataturk etc.

Fig. 3. A sample CV in Turkish and the system's *Segmenter* and *Parser Engine* working on it

-Parser Engine: At this step, the system does parsing process of proper names/ concepts, the abbreviations, suffixes and prefixes well-known patterns in sentence. During this decomposition process, as shown above, the OKB will be used. OKB of the proposed system is consisted of education, place / space, abbreviations, personal information, concepts, and companies' ontology. In the example below, "work experience" section is sent to a Parser Engine as an input (Figure 2 Step 7 or Figure 3 Step

8). One by one, each section will be separated and turned in to formatted form in Parser Engine. The obtained output from the Parser Engine is shown in Figure 3. The system can infer which concept is the piece scanned in a sentence from ontologies, and also can keep the information of start and finish lines for the next sentence.

In Figure 3, a yellow box contains the working experience segment of the sample résumé that is analyzed by the parser engine of the system. The segment contains the sentence ***'Bostancı Migros Torta / 2008-2009 / Pastane Şef Yardımcısı'***. The system starts to detect each terms of the sentence in concept base and tries to find each concept's URI information (ontology location) and appropriate place in the system's database (Figure 3 Step 8 to 11).



İş Deneyimi (Working Experiences) 8

Bostancı Migros Torta 2008-2009 Pastane Şef Yardımcısı

Altınkek San.Tic.Ltd.Şti Unlu Mamüller 2002-2008
Pastane Ustası
Elit Pastanesi Unlu Mamüller 1998-1999 Pastane Çırak

Terms	*Concept*	*Start*	*End*	*Value*	*Finding URI in Ontology*	*Related Section in DB*
1	İlçe(City)	1	8	Bostancı	Location Ont	İlçe(City)
2	Şirket (Company)	10	15	Migros	Organization Ont	Company name
3	Tarih (Date)	23	32	2008-2009	Concepts Ont	Date
4	Bölüm(Department)	27	33	Pastane	Organization Ont	Company Section Name
5	Meslek(Occupation)	41	54	Şef Yardımcısı	Occupations Ont	Duty

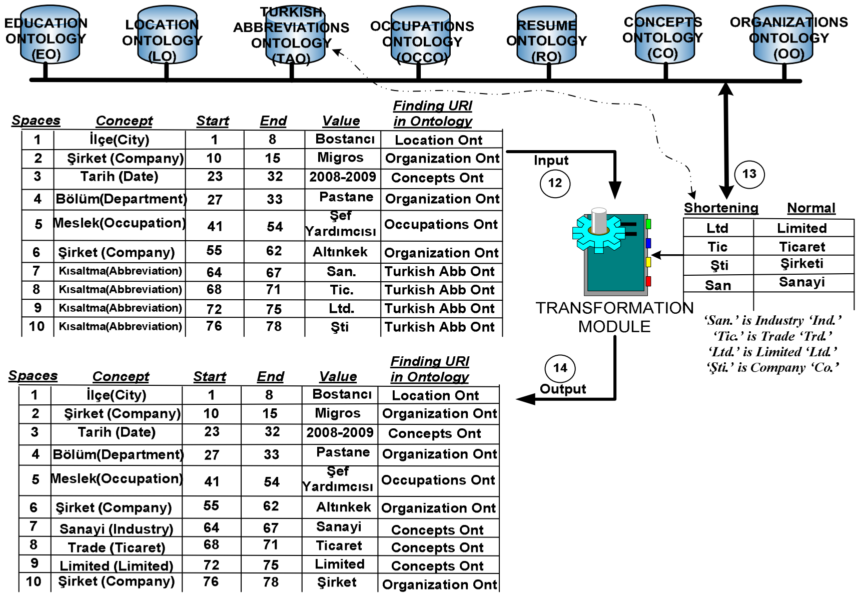
**Torta term is ignored here because it is found unneccessary term since no record in ontology.*

PARSED CONCEPTS OF THE PE
(in Turkish)
Özel Adlar: Enstitü, Şirket, Üniversite, Meslek, Derece vb.
Kıslatmalar: Üni., San., Tic., Ltd.,Yar. vb.
Türkçe'de Takılar: -de/da,-mek/mak vb.
Bilinen Kalıplar: M.K.Atatürk vb.

(in English)
Special Names: Institute, Company, University, Occupation, Degree etc.
Abbreviations: Uni., Ind.,Trd.,Ltd., Asst. etc.
Suffixes inTurkish: -de/da,-mek/mak etc.
Well-known Patterns: M.K.Ataturk etc.

Fig. 4. The system's *Parser* continues to analyze *Working Experiences* section of the same résumé according to *Special Names*, *Abbreviations*, *Suffixes*, and *Well-known patterns*

-The next process is text normalization though a Transformation Module. The system will scan the abbreviations in the latest result table and turn them into standard usage (Figure 4 Step 12). For example, above in the 2nd row in the working experience segment in Figure 3, the work experience knowledge includes many abbreviations such as San. Tic. etc. The system converts these abbreviations into normal state (Figure 4 Step 11 to 13). The transformation module detects the abbreviated concepts and retransforms to normal form and assigns them in to another table that does not involve any abbreviation concept. For instance, table may include 'Ltd.' abbreviation then the transformation module will convert it into normal state 'Limited' with its indicated concept in OKB. In final table, the transformation module always keeps the URIs of the fitting full concepts of abbreviations.



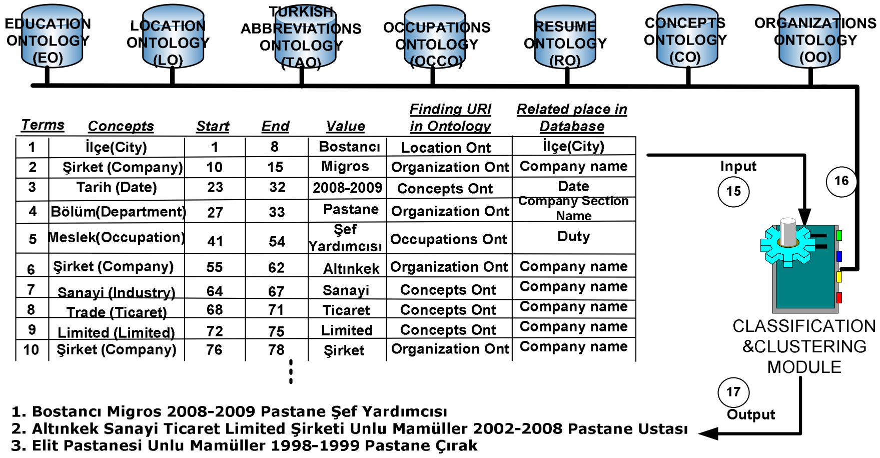
Spaces	Concept	Start	End	Value	Finding URI in Ontology
1	İlçe(City)	1	8	Bostancı	Location Ont
2	Şirket (Company)	10	15	Migros	Organization Ont
3	Tarih (Date)	23	32	2008-2009	Concepts Ont
4	Bölüm(Department)	27	33	Pastane	Organization Ont
5	Meslek(Occupation)	41	54	Şef Yardımcısı	Occupations Ont
6	Şirket (Company)	55	62	Altınkek	Organization Ont
7	Kısaltma(Abbreviation)	64	67	San.	Turkish Abb Ont
8	Kısaltma(Abbreviation)	68	71	Tic.	Turkish Abb Ont
9	Kısaltma(Abbreviation)	72	75	Ltd.	Turkish Abb Ont
10	Kısaltma(Abbreviation)	76	78	Şti	Turkish Abb Ont

Shortening	Normal
Ltd	Limited
Tic	Ticaret
Şti	Şirketi
San	Sanayi

'San.' is Industry 'Ind.'
'Tic.' is Trade 'Trd.'
'Ltd.' is Limited 'Ltd.'
'Şti.' is Company 'Co.'

Spaces	Concept	Start	End	Value	Finding URI in Ontology
1	İlçe(City)	1	8	Bostancı	Location Ont
2	Şirket (Company)	10	15	Migros	Organization Ont
3	Tarih (Date)	23	32	2008-2009	Concepts Ont
4	Bölüm(Department)	27	33	Pastane	Organization Ont
5	Meslek(Occupation)	41	54	Şef Yardımcısı	Occupations Ont
6	Şirket (Company)	55	62	Altınkek	Organization Ont
7	Sanayi (Industry)	64	67	Sanayi	Concepts Ont
8	Trade (Ticaret)	68	71	Ticaret	Concepts Ont
9	Limited (Limited)	72	75	Limited	Concepts Ont
10	Şirket (Company)	76	78	Şirket	Organization Ont

Fig. 5. The system's *Transformation Module* converts involved abbreviations in the *Working Experiences* segment


EDUCATION ONTOLOGY (EO)
LOCATION ONTOLOGY (LO)
TURKISH ABBREVIATIONS ONTOLOGY (TAO)
OCCUPATIONS ONTOLOGY (OCCO)
RESUME ONTOLOGY (RO)
CONCEPTS ONTOLOGY (CO)
ORGANIZATIONS ONTOLOGY (OO)
Input
15
16
CLASSIFICATION &CLUSTERING MODULE
17
Output

Terms	Concepts	Start	End	Value	Finding URI in Ontology	Related place in Database
1	İlçe(City)	1	8	Bostancı	Location Ont	İlçe(City)
2	Şirket (Company)	10	15	Migros	Organization Ont	Company name
3	Tarih (Date)	23	32	2008-2009	Concepts Ont	Date
4	Bölüm(Department)	27	33	Pastane	Organization Ont	Company Section Name
5	Meslek(Occupation)	41	54	Şef Yardımcısı	Occupations Ont	Duty
6	Şirket (Company)	55	62	Altınkek	Organization Ont	Company name
7	Sanayi (Industry)	64	67	Sanayi	Concepts Ont	Company name
8	Trade (Ticaret)	68	71	Ticaret	Concepts Ont	Company name
9	Limited (Limited)	72	75	Limited	Concepts Ont	Company name
10	Şirket (Company)	76	78	Şirket	Organization Ont	Company name

1. Bostancı Migros 2008-2009 Pastane Şef Yardımcısı
2. Altınkek Sanayi Ticaret Limited Şirketi Unlu Mamüller 2002-2008 Pastane Ustası
3. Elit Pastanesi Unlu Mamüller 1998-1999 Pastane Çırak

Fig. 6. The system's *Classification and Clustering Module* detects involved meaningful sentences in the *Working Experiences* segment of the résumé

-Classification and Clustering of Concepts: In this step, the combination of obtained concepts from the concept stack returned from the Parser Engine (as shown in the table) operation will be carried out. As shown in the figure above, three different phrases exist in the business experiences have been shown in section of business.

The system uses the concepts of found suitable OKBs at that moment to carry out meaningful sentences during its classifications processes (Figure 5 Step 15 to 17). The Classification and Clustering module detects each individual sentence in the working experiences segment of the résumé according to the generated table of transformation module in previous step. The Classification and Clustering module starts to detect start and stop point of each sentence according to the generated table. Each typical working experience sentence may contains some possible concepts a ‘City’, a ‘Date’, a ‘Company’, a ‘Department name’, an ‘Occupation’, and some abbreviations etc. The Classification and Clustering module will able to define start and stop points in a typical working experience sentence. In Figure 5, the system finds three different working experience sentences and then the system performs structuring its OWL ontology that is keeping reuse during expert finder searching in future (Table 4).

4.1 Generated Personal Résumé Ontology for Individuals

End of the *Classification and Clustering of Concepts* step, the system is able to generate its OWL form of the input résumé. For instance, personal information section in a résumé may involve *hasBirthCity, hasBirthTown, hasBirthDate, hasCurrentAddress, hasCurrentTown, hasEMail* etc. Similarly, the past working experience section is able to declare through some crucial properties such as *WorkingExperience, hasWorkCity, hasWorkCompany, hasWorkDate, hasStartWorkDate, hasEndWorkDate, hasWorkDuty* so on. The values of these properties are kept in the OWL form for the person (Mr.Ali Budak) that is shown in Table 4. For our example, first working experience is depicted in OWL form through the ***WorkingExperience property*** *(WorkingExperience_1, Line 17-24 in Table 4)*. The property contains ***hasWorkCity*** property that indicates a city concept. The city concept keeps a value that is ***"Bostancı"*** under the ***"&Location"*** ontology.

Table 4. A Sample of a Person Résumé (Özgeçmiş)

Resume No: R-0000001		
Name: Ali Budak	1	<!— A portion of the Résumé (Özgeçmiş) Ontology in English Language -->
Doğum Yeri-Tarihi(Birth Place-	2	<Resume rdf:ID=" R-0000001">
Date): Zonguldak/Çaycuma	3	<hasBirthCity rdf:resource="&Location;Zonguldak"/>
20.06.1984	4	<hasBirthTown rdf:resource="&Location;Çaycuma"/>
Adres (Address): Yalı Mah Er Kılıç	5	<hasBirthDate rdf:datatype="&Concept;Date">20.06.1984
Sok No:41/1 Cevizli	6	</hasBirthDate>
Maltepe/Istanbul	7	<hasCurrentAddress rdf:resource="&Location;Street"/>Yalı Mah Er
Askerlik Durumu (Military State):	8	Kılıç Sok
Tamamlandı (Completed)	9	No:41/1 Cevizli</ hasCurrentAddress>
Sürücü Belgesi (Driver Licence): B	10	< hasCurrentTown rdf:resource="&Location; Maltepe"/>
Sınıfı(B Class)	11	<hasCurrentCity rdf:resource="&Location;İstanbul"/>
Gsm: 0XXX XXX XX XX	12	<hasMilitaryState rdf:resource=”#Completed”/>
E-Mail: turgayortak@hotmail.com	13	<hasDirverLicence rdf:resource=”#B Class”/>
İş Deneyimi (Working Expe-	14	<hasCurrentGSM rdf:datatype="&Concept;GSMNumber">0XXX XXX
riences)	15	XX XX
-Bostancı Migros Torta / 2008-	16	</hasCurrentGSM>
2009 / Pastane Şef Yardımcısı	17	<hasEMail rdf:datatype="&Concept;EMail">turgayortak@hotmail.com
-Altınkek San.Tic.Ltd.Şti Unlu	18	</hasEMail>
Mamuller 2002-2008 Pastane	19	<WorkingExperience rdf:about=”WorkingExperience_1”>
Ustası	20	<hasWorkCity rdf:resource="&Location;Bostancı"/>

Table 4. (*Continued.*)

-Elit Pastanesi Unlu Mamuller **1998-1999 Pastane Çırak** **……..** *In English:* **'Bostancı' is a location in Istanbul city.** **'Migros' is a market company in Turkey.** **'Torta' is an ignored unnecessary data.** **'Pastane' is an organization that is pastry-shop.** **'Şef Yardımcısı' is a title of *an* occupation type that is called Assistant of Chef.** **'Pastane Ustası' is a type of occupation that is called Pastry-shop Chef.** **'Çırak' is a title of *an* occupation type that is called footboy.** **'Altınkek' is a name of a company.** **'San.' is abbreviation of the industry that same as 'Ind.' in English and same for others: 'Tic.' is Trade 'Trd.', 'Ltd.' is Limited 'Ltd.' and 'Şti.' is company 'Co.' so on.** **'Elit Pastanesi Unlu Mamuller' is a name of a company.**	**<hasWorkCompany** **rdf:datatype="&Organization;SuperMarket">Migros** **</hasWorkCompany>** **<hasWorkDate rdf:resource="&Concepts;Date"/>2008-2009</hasWorkDate>** **<hasWorkDepartment rdf:resource="&Organization;Pastane"/>** **<hasWorkDuty rdf:resource="&Occupation;Şef_Yardımcısı"/>** **</WorkingExperience>** **<WorkingExperience rdf:about="WorkingExperience_2">** **<hasWorkCompany rdf:datatype="&Organization;Pastane">Altınkek San.** **Tic. Ltd. Şti Unlu Mamuller</ hasWorkCompany>** **<hasWorkDate rdf:resource="&Concepts;Date"/>2002-2008</hasWorkDate>** **<hasWorkDepartment rdf:resource="&Organization;Pastane"/>** **<hasWorkDuty rdf:resource="&Occupation;Pastane Ustası"/>** **</WorkingExperience>** **<WorkingExperience rdf:about="WorkingExperience_3">** **<hasWorkCompany rdf:datatype="&Organization;Pastane">Elit Pastanesi** **Unlu Mamuller </hasWorkCompany>** **<hasWorkDate rdf:resource="&Concepts;Date"/>1998-1999</hasWorkDate>** **<hasWorkDepartment rdf:resource="&Organization;Pastane"/>** **<hasWorkDuty rdf:resource="&Occupation;Çırak"/>** **</WorkingExperience>** **</Resume>** **</rdf:RDF>**

5 Conclusion

In this article, an ontology driven information extraction system is considered that is called Ontology-based Résumé Parser (ORP) is operated on few millions English language and Turkish language résumés to convert them automatically ontological format. The system is also assist to perform the expert finding/discovery and aggregation of skill information among résumé repository through its involved semantic approach. To do this, the system has its own ontological semantic based inferencing rules during its inferring mechanism. Therefore, the system has able to learn and keeps the inferred new relationships after inferring new information and uses them in the next steps when it requires again that makes the system is a learning-based system. As a conclusion, some of above mentioned properties of the ORP are not discussed here since taking too much space and are detailed separate studies as future studies of the project.

Acknowledgement. This project is supported by the Kariyer.net Company[1] that is titled as *"Document-based Semantic Information Extraction System from Turkish Résumés through Ontology"*.

References

1. Hu, B., Kalfoglou, Y., Alani, H., Dupplaw, D., Lewis, P., Shadbolt, N.: Semantic metrics. In: Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management, pp. 166–181 (2006)
2. Bojars, U., Breslin, J.G.: RésuméRDF: Expressing Skill Information on the Semantic Web. In: The 1st International Workshop on ExpertFinder, Berlin, Germany (January 2007)
3. Parada, R.A.: DOAC Vocabulary Specification (July 08, 2006), `http://ramonantonio.net/doac/0.1/`
4. Karamatlı, E., Akyokuş, S.: Résumé Information Extraction with Named Entity Clustering based on Relationships. In: INISTA 2010, Kayseri (2010) (last visited: April 12, 2010)
5. Paolucci, M., Kawamura, T., Payne, T.R., Sycara, K.: Semantic Matching of Web Services Capabilities. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 333–347. Springer, Heidelberg (2002)
6. Çelik, D., Elci, A.: Towards a semantic-based workflow model to the composition of OWL-S Atomic Processes-through process based similarity matching and inferencing Techniques. J. of Internet Technology, Taiwan Academic Network Executive Committee (2010) ISSN: 1607-9264, Published by: Taiwan Academic Network Executive Committee (Accepted, SCI-E)
7. Çelik, D., Elçi, A.: Ontology-Based Matchmaking and Composition of Business Processes. In: Elçi, A., Koné, M.T., Orgun, M.A. (eds.) Semantic Agent Systems. SCI, vol. 344, pp. 133–157. Springer, Heidelberg (2011)
8. Çelik, D., Elci, A.: OWL-S Semantic-Based Workflow Model Based on Atomic Processes Unification. Journal of the Science and Engineering University of Cankaya, Ankara, Turkey (submitted, June 2010)
9. Sovren Résumé / CV Parser, `http://www.sovren.com/` (last visited: April 12, 2010)
10. ALEX Résumé Parsing, `http://www.hireability.com/ALEX/` (last visited: April 12, 2010)
11. Résumé Grabber Suite, `http://www.egrabber.com/résumégrabbersuite/` (last visited: April 12, 2010)
12. Daxtra CVX, `http://www.daxtra.com/` (last visited: April 12, 2010)
13. Yu, K., Guan, G., Zhou, M.: Résumé information extraction with cascaded hybrid model. In: ACL 2005: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, pp. 499–506 (2005)
14. Piskorski, J., Kowalkiewicz, M., Kaczmarek, T.: Information Extraction from CV. Information Retrieval and Filtering, 185–192 (2005)
15. Chieu, H.L., Ng, H.T., Lee, Y.K.: Closing the Gap: Learning-Based Information Extraction Rivaling. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan (2003)
16. Protégé, OWL-S Ontology Editor CS / AI Department, University of Malta (2004), `http://owlseditor.semwebcentral.org/` (last visited: April 2009)
17. Gruber, T. (N.d.): What is Ontology? `http://wwwksl.stanford.edu/kst/what-is-an-ontology.html` (last visited: April 15, 2007)
18. Onder, P., Ozen, N., Unlu, S., Orhan, Z.: A Framework for Building a Turkish Lexicon and Knowledge Base. In: Proceedings of IKE 2008, The 2008 International Conference on Information and Knowledge Engineering, Monte Carlo Resort, Las Vegas, Nevada, USA (July 2008)