

# Information Extraction from CV

Tomasz Kaczmarek  
The Poznan University of Economics  
t.kaczmarek@kie.ae.poznan.pl

Marek Kowalkiewicz  
The Poznan University of Economics  
m.kowalkiewicz@kie.ae.poznan.pl

Jakub Piskorski  
German Research Center for Artificial Intelligence  
piskorski@dfki.de

## Abstract

The paper gives an outlook of an ongoing project on deploying information extraction techniques in the process of converting any kind of raw application documents written in Polish, such as CVs, motivation letters or application forms into compact and highly-structured data. We pinpoint the challenging issues to be faced and potential benefits in the area of learning systems, HR and recruitment modules of information systems.

## 1. Introduction

Recent advances in information technology such as Information Extraction (IE) [Appelt and Israel 1999] provide dramatic improvements in conversion of the overflow of raw textual information into structured data which constitute the input for discovering more complex patterns in textual data collections. To be more precise, the task of IE is to identify predefined set of concepts in a specific domain and ignoring other irrelevant information. Although considerable work on utilizing IE technique in real-world business application has been reported, research on this topic in the context of processing Polish texts has been almost neglected. On the other side, the need to process textual information is increasing. We decided to launch a project that is aiming at applying modern IE tools for processing CVs' and motivation letters written in Polish. There were three incentives for choosing this topic:

- this type of documents is semi-structured and usually contains numerous named entities, that can be identified with IE tools
- the complexity of Polish language makes applications of IE in this language particularly interesting and challenging as a research topic,

while choosing semi-structured documents gives hope for achieving results,

- such results could be directly applicable in several areas. A tool to extract relevant information from CVs' could be beneficial for HR departments, personal advisory companies and e-learning systems. It could ease and speed up the process of data extraction to the structured form, which allows for further data mining and retrieval tools. It would enable creation of navigational structure between such data to ease search of potential employees. For e-learning systems, or any other systems that require rich user information, it would enable creation of initial user profile without limited possibilities of obtaining it through user survey.

This paper describes the early stage work in the depicted area, the challenges to be met and the aims of the project.

## **2. Project description**

### **2.1. Project goals**

The most general project goal is to extract any information that could be relevant for systems that require information about potential employees / users / customers, their skills, job and education history.

This was divided into several sub goals, namely:

- identifying information to be extracted,
- preparation of linguistic and ontological resources for applying IE tools in Polish language,
- preparing tools to extract relevant information from texts,
- applying selected tools on the test document collection,
- assessment of the results.

The following section is devoted to the description of proposed solutions to achieve these targets.

### **2.2. Proposed solution description**

#### **2.2.1. Information to be extracted**

Preliminarily we have selected the following information to be extracted from the CVs' and motivation letters:

- first name, last name of the letter / CV author,
- his address, email, phone contact, homepage or other contact information,

- current and past job titles and the period of work, the names of organizations that an author worked for, to reconstruct his job history in a structured way,
- education, college / university names, postgraduate studies, name of speciality,
- language proficiency levels,
- information about other skills – to be extracted from education, work, leisure / hobby and other included references. This could be mapped to the given set of skills (see ontologies for skills or skill specifications).
- if there is a list of publications included (in the case of researcher CV) it could be used to obtain additional information about speciality – the use of publication titles, books, conference names etc.

### 2.2.2. NLP resources and tools

In order to accomplish the goals of the project we use SProUT – a novel multilingual IE platform [Drożdżyński et al., 2004] which has been adopted to the processing of Polish [Piskorski et al, 2004]. Furthermore, we intend to utilize Polish named-entity grammars developed within the highly declarative grammar paradigm of SProUT [Piskorski 2004]. These grammars are dedicated for recognition of standard named entities like persons, organizations, locations, temporal expressions and quantities. Obviously, in order to meet project requirements some work on adapting available linguistic resources and additional grammar writing is envisaged.

Although the aforementioned resources constitute a backbone of our extraction engine, a great amount of work will focus on exploring various merging strategies for assembling the information recognized by NE (named entity recognition component) and small-scale structure grammars since relevant pieces of information may be distributed across the whole document. As a matter of fact, merging small-scale structures is a lesser studied area of IE [Kehler 1998], and there is still a lot of space for improvement.

Furthermore, since the input data contains partially text fragments in English (e.g., English names in the publication titles, software names, certification and training titles, scientific grades and titles etc.) the tool should be able to apply corresponding grammars for English in the areas of text that will be recognized as written in this language. In a preprocessing phase a language recognition tool will split the document into language specific fragments to which an appropriate grammar is applied. For processing English we intend to deploy and adapt the available SproUTs NE-grammar for English.

The novelty here would be applying SProUT IE tool for Polish language with a view of mixed nature of the processed documents.

### **2.2.3. Tests on document collection**

We plan to gather test collection of about hundred CVs' and motivation letters. The collection will be indexed manually and then it will serve as a test bed for SProUT engine tuned to process this type of documents. Manual document index will specify the structures that contain relevant information (see above) that should be retrieved. We expect the outcome of each test run to be a set of (possibly partially) filled templates that represent entities that were recognized by a tool.

### **2.2.4. Solution evaluation**

For evaluation purposes we intend to manually create a list of expected output structures (for each strata separately – named-entity boundaries, fully instantiated templates, etc.) and evaluate the system according to the standard recall-precision metrics. The evaluation phase serves also for the improvement of a processing engine. We hope that the comparison of manual and automatically generated structures allows us to advance both the linguistic resources and the engine itself (without losing the generality of extraction process).

## **2.3. Possible project extensions and continuation**

The project is in fact an initial, important stage for several other possible projects. Having a tool that extracts education, job and skill related information with reasonable precision and recall allows for further projects basing on utilization of this information. The following paragraphs briefly sketch the possible continuations.

### **2.3.1. Semi-automatic ontology building**

The IE tool prepared in the project could serve as a basis for an automatic ontology building tool. Knowledge mining is currently popular and widely pursued research topic, with numerous applications of IE and NE technologies. Our project outcome could serve as basis for extraction of new ontological entities and relationships in the area of personal information, skills, education, job and training. The scenario for this would be as follows: SProUT would have an initial set of ontological rules for the domain of education and employment, described as grammar dependencies. The example of such rules would be: educational institution that is associated with time period could indicate studies or school for given person, the name of job position accompanied by company name and possibly time period suggests a working period in person's job history. As SProUT is capable of recognizing new entities, it could also identify their context and connect them with those already known. This would lead to ontology development as new rules and new entities could be extracted in the process similar to machine learning.

### 2.3.2. User profiles for retrieval and filtering systems

One of initial project motivations and aims was to provide basis for constructing user profiles for information retrieval and filtering systems that would be founded not only on the survey of user (temporary) interests but also on his skills and abilities. Information filtering system could benefit greatly having such profiles available, as it could pass information that is potentially relevant to the user according to his skills. The same applies to the retrieval system that could present the results of query processing in a way that takes into account user's skills not to give him to "complicated" or to "simple" documents.

### 2.3.3. Web crawler for potential job candidates

There are a number of web pages that contain CVs' or other personal information, which could be potentially relevant for personal advisory companies or job-seeking agencies. With our tool a web crawler for potential candidates could be envisaged and easily achievable. Such technology could change the paradigm of job searching – instead of sending multiple copies of CV to various institutions while looking for job, it would be enough to place it on the Web and have it indexed to be placed in the right database. Whether this would be beneficial for job seeking process remains to be found.

## 3. Bibliography

1. [Appelt and Israel 1999] D. Appelt and D. Israel. 1999. *An introduction to information extraction technology*. A Tutorial prepared for IJCAI-99 Conference.
2. [Drozdzyński et al, 2004] W. Drozdzyński, H-U. Krieger, J. Piskorski, U. Schäfer, F. Xu. 2004. *Shallow Processing with Unification and Typed Feature Structures – Foundations and Applications*. In German AI Journal *KI-Zeitschrift*, Vol. 01/04, Gesellschaft für Informatik e.V.
3. [Kehler 1998] *Andrew Kehler: Learning Embedded Discourse Mechanisms for Information Extraction*, in the Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing, Stanford, CA, March 1998.
4. [Piskorski et. al 2004] J. Piskorski, P. Homola, M. Marciniak, A. Mykowiecka, A. Przepiórkowski, M. Woliński. *Information Extraction for Polish Using the SProUT Platform*. In Proceedings of International Conference on Intelligent Information Systems - New Trends in Intelligent Information Processing and Web Mining, in Zakopane, Poland, May 2004
5. [Piskorski 2004] J. Piskorski. *Rule-based Named-Entity Recognition for Polish*. In the Proceeding of the Workshop on Named-Entity Recognition for NLP Applications held in conjunction with the *1st International Joint Conference on NLP*, Sanya, Hainan Island, China, March 2004