Linguistic information extraction for job ads (SIRE project)

Romain Loth, Delphine Battistelli, François-Régis Chaumartin, Hugues de Mazancourt, Jean-Luc Minel, Axelle Vinckx

Linguistic information extraction for job ads (SIRE project)

Romain Loth*, Delphine Battistelli*, François-Régis Chaumartin***,
Hugues de Mazancourt**, Jean-Luc Minel*, Axelle Vinckx**,

(*) MoDyCo - UMR 7114 CNRS
Univ Paris Ouest Nanterre La Défense
FR-92001 Nanterre
+33 140 977 431
rloth@u-paris10.fr

(**) Lingway Labs
18 rue Pasteur
FR-94270 Le Kremlin-Bicêtre
+33 158 461 240
hugues.de-mazancourt@lingway.com

(***) Proxem
7 impasse Dumur
FR-92110 Clichy
+33 612 165 923
frc@proxem.com

ABSTRACT
As a text, each job advertisement expresses rich information about the occupation at hand, such as competence needs (i.e. required degrees, field knowledge, task expertise or technical skills). To facilitate the access to this information, the SIRE project conducted a corpus based study of how to articulate HR expert ontologies with modern semi-supervised information extraction techniques. An adaptive semantic labeling framework is developed through a parallel work on retrieval rules and on latent semantic lexicons of terms and jargon phrases. In its operational stage, our prototype will collect online job ads and index their content into detailed RDF triples compatible with applications ranging from enhanced job search to automated labor-market analysis.

Categories and Subject Descriptors
H.3.1 [**Information storage and retrieval**]: Content Analysis and Indexing – *abstracting methods, dictionaries, indexing methods, linguistic processing.*

General Terms: Languages.

Keywords
Information extraction, human resources, job ads, natural language processing.

1. INTRODUCTION
The project is based on the idea that applying state-of-the-art information extraction techniques to web-published job ads could help increase the fluidity of information on the labor market, through a framework of adaptative semantic labeling. On the existing job-boards, 'help wanted' ads are mainly indexed according to geographical region, occupational branch or business sector. The information pertaining to skills and missions is not thoroughly picked up. Once analyzed and categorized, this information could help in finding a better match and provide a wider range of job search queries. In addition, the corpus of online job offers as a whole is a good indicator of the underlying trends on the labor market. Its analysis can help in assessing job availability for a given profession or business sector and in understanding the fast-paced changes of the required skill sets across professions.

The SIRE project stems from a partnership between the R&D teams of the two companies Lingway and Proxem and a research group from MoDyCo, an academic research laboratory. The first phase (18 months) focused on identifying available resources, analyzing the lexicon used in French job ads, writing extraction rules and exploring different indexing techniques for the (wide) set of related tasks. Presently the project enters its practical phase, with the gathering of a public interest group and the planned release of a first prototype.

2. RELATED ISSUES

2.1 Theoretical Background
As text genres go, job advertisement combines several specificities (short canonical structure, homogeneity of topic, organized articulation of lexical registers). That directed our research towards relevant concepts and methods within linguistics. On a more practical level, the desired depth and scale of automated analysis made us borrow procedures and data models from knowledge engineering, machine learning and information extraction. A job posting constitutes a specific communication act between an organization and an unknown candidate/reader. This communication act, repeated again and again, determines its own written register. It is a descriptive, public text which is subjected to regulations and restricted to a single topic: the open position. Job ads are also routinely copied by their redactors and compared by their readers, which tends to contrive the writing into a conventional format (both visual and redactional). These phenomena relate our research to the corpus study of written text registers as a constructed functional form [5].

Since [10] and related works, natural language processing has witnessed significant advances in the unsupervised modeling of word meaning. The general idea is to implement a similarity metric on a lexical space, where the distances are learned from word distribution and co-occurrence frequencies in relevant contexts. These lexical spaces take the mathematical form of dimensionality-reduced vector spaces as exemplified in the methodology referred to as LSA (Latent Semantic Analysis). Lexical analysis is in turn directly connected to academic work on knowledge engineering. Ontologies were shown to be adequate formal models for the semantic description of technical, specialized registers [2]. In particular, the ontology model provides a shell to integrate the collected lexical material within the numerous nomenclatures to be found in the HR field (occupational classifications, business sectors, contract types, skills classifications).

Finally, classical machine learning techniques are also used in the process, among which several clustering algorithms, sample-based classification and duplicate removal algorithms. One of the challenges of the project is to combine machine-learning based

mining with symbolic methods (pattern-based retrieval component and ontology-based inferences).

2.2 Domain-related Works

The literature on HR ontologies is rapidly growing and already includes very elaborate models (see [4] and [8], among others). In their construction process, the authors distinguish inter-related but separate job features. Knowledge engineering successfully converged on that point with the typologies used by job-boards and above all with those used by HR experts themselves [9].

In a noticeable contrast, text-mining approaches like [1] have proceeded in a bag-of-words fashion, by mapping all words from the text to a common classifying space. The work presented in [7] goes into more detail and takes advantage of the conventional sections of the job ad to harvest separate groups of words or character n-grams. Although text-mining approaches allow adaptative clustering of the documents, they're insufficient on their own because they mix distinct information (i.e. degree with personality requirements). For our purposes, this impedes the use of symbolic heuristics and linkage with classifications.

3. METHODOLOGY

3.1 A Corpus-Based Typology

Our method consisted in manually tagging different kinds of text extracts from the ads in the hope of relating them to corresponding 'ontological' features of the job. During the first stage of the project, we thus performed detailed annotation of a representative sample of 200 job ads. It allowed us to delineate a typology of 13 commonly encountered information types (referred to as i-types, see ***table 1***).

Table 1. Encountered i-types in the text of a French job ad (typology and average values)

Information type	Subtypes	Freq. per ad	Length (words)
Occupation	position title, profession keywords	2.36	3.98
Mobility	location, travel requirements, commercial zone	0.85	2.67
Company	name, group, website, quantitative data (turnover, workforce)	2.43	3.15
Recruiter	name, specialization, website	0.23	3.27
Sector	activity details, trade field, sold product or service	2.06	5.02
Team and conditions	team or work unit, supervisor, work environment	2.09	4.63
Missions	role of the job, tasks, objectives	**7.99**	8.06
Contract	contract type, duration, salary, working hours	1.10	3.41
Expertise	composite expertise phrase (time + kind of experience)	1.29	9.45
Competence	field knowledge, technical skills, languages, software	2.04	3.74
Personality	*(no emerging subtypes)*	3.25	2.63
Education	degree, title, field of studies	0.86	5.26
Contact info	e-mail, contact name, address, ref no, contact procedure	0.79	7.20

What we obtain is a set of empirical labels, relevant for the segmentation of the discourse structure of a typical French job ad. These labels are used as a bridge between the text-retrieval component and the ontology.

The study of occurrence contexts for each i-type in the annotated results showed that pattern-based extraction could be used in most cases to tell apart these different pieces of information. Lingway developed dedicated rule-based grammars to automatically extract some of these information types, such as missions and competencies (the same grammars can be reused to extract this information from different HR documents).

3.2 Ontology Construction

Additionally, we posit that the final {document+tags} output should be available at different levels of granularity. Indeed, as underlined in [3], various use cases have to be considered for 'tasks & competencies' indexes of the ads (e.g. job search, team staffing, corporate strategy information, market monitoring). Some cases will require detailed data, whereas some other cases call for a normalized reduced set of categories. The lower part of the ontology (interpretation grid) is being built with a nested classification structure, where the same text data can be described at different levels.

As a helping tool during development, we build a latent semantic vector lexicon (or v-lexicon) to check distance between instantiated concepts. We borrow techniques from the LSA model to develop our own semantic metrics, specifically dedicated to the representation of occupational terminology. This v-lexicon is constructed with words and terms extracted from 1081 job ads. Each word is POS-tagged, lemmatized and associated with a feature vector, where the dimensions are a composite of its phrase and document occurrence contexts. A similar work is done using the multi-word expressions (MWE) from tasks and skills inventories instead of words for the entries. The approximate quality of the semantic distances obtained on MWE is compensated by the convenience of unsupervised training and the all-purposeness of the tool.

Our resulting metric is useful both for ontology engineering in the development phase (measuring the distance between tentative concepts or categories) and for classification tasks in the operational phase (picking a tag from a predefined set or categorizing the document as a whole). We are also experimenting on specialized versions of the lexicon, with dimensions trained only for one i-type. The lexicon is continually enhanced by adding new relevant contexts as new dimensions (see [6] on that process). In a simplified view, the enriching of the ontology provides new dimensions to rework the vector lexicon, which in turn refines term extraction and helps enriching the ontology. These iterations are semi-supervised: hand-crafted heuristics and manual sorting fills the gaps and errors of the automatic process. The v-lexicon is a shared resource in the development work and allows us to echo the improvements of each component on the rest of the linguistic processing chain.

4. PROCESSING WORKFLOW

When SIRE reaches its operational stage, the components will run on a server dedicated to crawling job ads and indexing them into a detailed database structure based on RDF triples (see ***figure 1***).

The first processing step crawls job ads from the Web. Depending on the application, it could either target specialized websites from a given professional branch or be set on major job-boards.

The job-ad pages are cleaned from their boilerplate text using a variant of classical heuristics: tag/text ratio and a diff between groups of pages from the same location. The next task is to spot duplicate job offers appearing in the collected corpus. A shingling algorithm will mark duplicate candidates, although the final check must actually wait for the indexing stages.

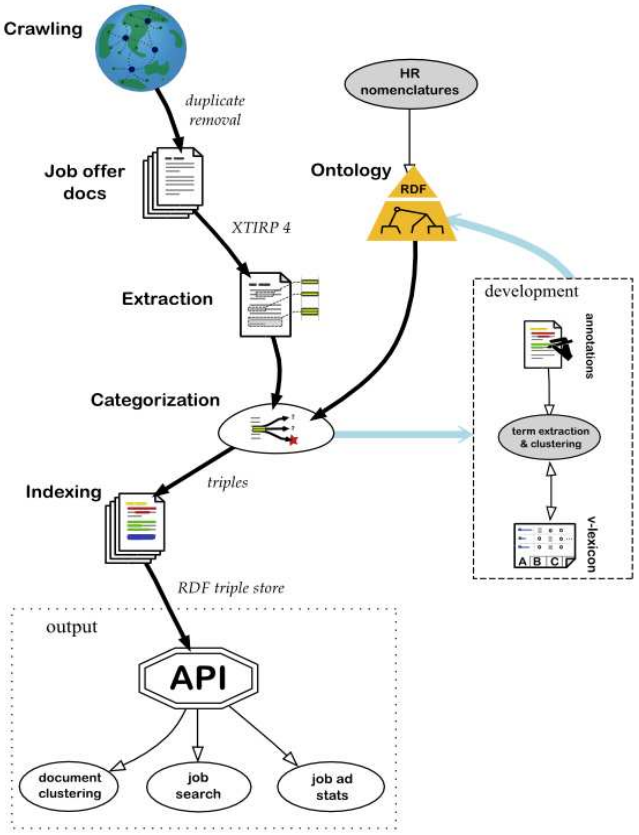

Figure 1. SIRE Linguistic Processing Workflow

For each offer, the text retrieval component XTIRP runs pattern-extraction rules to fill an XML structure tree with typed fragments. These extracts are then annotated by an array of categorizers that use the text to choose corresponding tags from the nomenclatures in the ontology.

Each categorizer outputs an RDF triple built on a subject-predicate-object format. The subject is the current job ad and the i-types are taken as predicates: e.g. <current job ad ID> 'has for a profession', 'names as a task', 'originates from the business sector'… The role of the categorizer is to find the object of the predication: 'accountant' , 'budget balancing', 'fishing industry'. The triples are then persisted in an RDF store and accessed through a SPARQL endpoint by API queries. The first prototype of this platform will be completed around June 2010.

5. CONCLUSION

The retrieval of relevant information concerning job descriptions and competencies is not a trivial task. They are formulated through phrasal expression which associate an impersonal HR register with professional jargons. To facilitate access to this information, we manually studied its wording in a sample corpus of 200 job ads, thus defining intermediate text-labels. In a next step we combine the meticulous descriptions of HR ontologies and nomenclatures with the more flexible but less detailed methods from the field of text-mining. Associated together, these techniques pave the way for a text-extracting and understanding platform dedicated to the interoperable and adaptive indexation of job ads.

6. ACKNOWLEDGMENTS

SIRE is an acronym for 'Sémantique, Internet, Recrutement et Emploi'. This project has received the Cap Digital label, thus joining the largest business cluster dedicated to digital content and services in the Paris region. It is granted by the FEDER European Regional Development Fund.

7. REFERENCES

[1] Aureli, E. et Iezzi, D.F. 2006. Recruitment via web and information technology. Proceedings of JADT 06 (Besançon, France, 2006).

[2] Bourigault, D., Aussenac-Gilles, N. et al. 2004. Construction de ressources terminologiques ou ontologiques à partir de textes. Revue d'Intelligence Artificielle. 18, 4 (2004), 24.

[3] Braun, S., Kunzmann, C. et al. 2010. People Tagging & Ontology Maturing: Towards Collaborative Competence Management. *From CSCW to Web 2.0*. Randall and P. Salembier (eds), Springer.

[4] Gómez-Pérez, A., Ramírez, J. et al. 2007. Reusing Human Resources Management Standards for Employment Services. Proceedings of FIRST 07 (Busan, Korea, 2007), 28-41.

[5] Halliday, M.A. and Hasan, R. 1989. Language, context, and text: Aspects of language in a social-semiotic perspective. Oxford University Press.

[6] Karov, Y. and Edelman, S. 1998. Similarity-based word sense disambiguation. Computational linguistics. 24, 1 (1998), 41-59.

[7] Kessler, R. 2009. Traitement automatique d'informations appliqué aux ressources humaines. *PhD thesis.* Université d'Avignon et des Pays de Vaucluse.

[8] Laclavik, M., Seleng, M. et al. 2007. Ontology based text annotation. Information modelling and knowledge bases XVIII. (2007), 311-315.

[9] Marchal, E., Mellet, K. et al. 2007. Job board toolkits: Internet matchmaking and changes in job advertisements. Human Relations. 60, 7 (2007), 1091-1113.

[10] Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd meeting of ACL (1995), 189-196.