UNICOMP: Identification of Enterprise Competencies to Build Collaborative Networks

Kafil Hajlaoui[1], Xavier Boucher[2], Omar Boussaid[1]
[1]Université de Lyon (ERIC Lyon 2)
5 Avenue Pierre Mendès-France, 69676 Bron, France
{kafil.hajlaoui, omar.boussaid}@univ-lyon2.fr
[2]Ecole des Mines de Saint Etienne
158 cours Fauriel, centre G2I, 42023 Saint Etienne, France
boucher@emse.fr

Abstract. In a context of decision-aid to support the identification of collaborative networks, this paper focuses on extracting essential facets of firm competencies. Due to the complexity of the notion of competence, this contribution is based on a semantic representation of information using semantic ontology, bonds and a linguistic treatment based on the utilization of syntactic patterns. To identify enterprise competencies, the UNICOMP system uses company web sites as information source, as well as a general ontology of competencies as semantic resource.

Keywords: *information extraction, ontology, patterns, enterprise networks.*

1 Introduction

The approach presented below is part of a research to provide a decision support to support the identification of potential Virtual Breeding Environments (VBE). This work takes as a hypothesis the existence of an "open universe" of potential partners to build VBEs where any company can participate. Typically, this issue often appears when one has to analyse a regional business area to identify potential collaborative networks among firms. As a consequence, the identification of potential collaborative partners will be based on the use of public available information.
The web is considered as the fist information on business partners. In previous publications [1] [2], the authors explained the decision aid approach based on the extraction and use of key characteristic information on companies, concerning their activity fields and internal competencies. This information on company competencies will be further used, at a second stage, to generate new knowledge on the potential structure of VBEs. This assumption induces the use of specific information extraction mechanisms. The objective of this paper is to present an information extraction approach to extract company competence traces, through the utilization of public information available on company websites.

Some research on competencies identification techniques are based on the analysis of texts presenting rather homogeneous and structured data. In such approaches, the characterisation of competencies requires private information provided by the

companies [3] [4]. Non structured documents, like interviews on competence description realized with company experts can be analyzed to further generate a structured description of company competencies. Other authors identify company competencies by inferring expert rules. [5] and [6] employ "expert rules" based on similarities among individual competencies to generate a systematic identification of available competencies. [7] applies other kinds "of expert rules" to individual competencies, based on deductive relationship and not similarity. Another example is provided by [8] where the semantic annotation is also employed on the documents produced by the employees. This competence identification technique is primarily based on rules manually made by an expert of the domain.

However, such approaches can not be applied to our context where, by hypothesis, only non-structured public information on companies is available. The basic information to characterize competencies is available through company public websites: this assumption induces the use of specific information extraction mechanisms, requiring to cope with difficult semantic issues.

Section 2 presents the extraction approach based on the use of ontology and lexical patterns. Since the ontology has been presented in previous works [1], the paper focuses on the formalisation of lexical patterns (section 3). Section 4 describes the implementation of the extraction system and discusses the performance results, followed by conclusion in section 5.

2 Competence identification: use of ontology and lexical patterns

Due to the complexity of the notion of competence, the identification of competencies through website information requires implementing semantic mechanisms [2]. The approach presented here uses ontology and lexical patterns applied to the specific area of mechanical industry. The ontology provides a generic description of this domain and the projection of patterns on the corpus is used to activate the ontology classes for each company, thus providing a company competence trace.

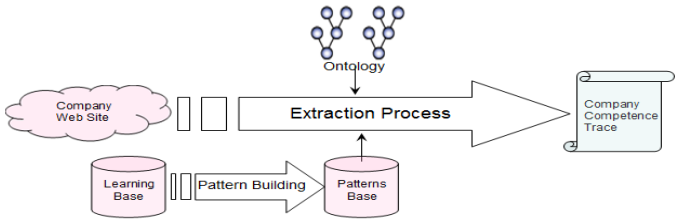

Fig. 1. Extraction process using Ontology and Patterns

The formalization of the ontology has been described in [1] [2]. Consequently, we just provide here a brief summary and the rest of the paper focuses on formalization and use of patterns for competence identification.

The ontology requires rigorous conceptualization of the notions related to firm competencies. ARCHONTE ontology building methodology [9] was chosen, because it defines precise cognitive principles and mechanisms for each step of ontology

formalization. This method makes possible a good semantic representation of the conceptual network of the ontology through a « semantic normalisation » mechanism.

From a company website, it is often impossible to extract a weel structured and exhaustive identification of firm competencies. However it remains possible to extract a sufficient set of competence "traces", which makes possible to assess with a good degree of precision an indicia of competence similarity among firms. To implement the extraction mechanism, an ontology of "competence traces" has been formalized. This ontology is composed of two sub-ontologies: a *generic ontology* and a *domain ontology*. The former provides a very generic conceptual model of "competence traces", which is independent of any business sector. However because of this high generality level, the generic ontology is not directly applicable for information extraction. Then, it is complemented by a *domain ontology,* which constitutes an extension of the generic ontology, specific to some selected business sector (mechanical industry in our case). The "competence traces" concepts of this domain ontology are directly linked to the applied field of the mechanical industry. Furthermore, at the lower level of the domain ontology, each ontological concept is associated to specific terms of the domain, further used in the extraction mechanism as "competence trace *identifiers*". The *identifiers* will be used to detect the presence of a competence trace concept within document corpora (extracted from company websites). This detection of "competence traces" is called the activation process applied to the ontology conceptual classes.

To fulfill the objective of extraction information, the direct detection of terms in a corpus is not a sufficient condition. To confirm the presence of an ontology concept (associated to one or several *identifiers*) in the corpus, we have to cope with context-dependent linguistic phenomena which can change the semantic of the *identifiers*: synonyms, antonyms and, more generally, semantic similarity among terms make necessary to use specific semantic techniques. To deal with such ambiguity phenomena we propose the use of linguistic patterns.

3 Lexical patterns

Lexical patterns aim at formalising a contextual signature of an expression (*identifiers* in our case). Patterns are based on principles of distributive semantics which states that the meaning of a word is strongly correlated to its context. More formally, lexical patterns identify and formalize linguistic relationships, by defining syntactic constraints on the context of the terms [10]. Patterns are built by schematizing the lexical and syntactic context which appears common to several text fragments (corresponding to various occurrences of the linguistic relationship considered). This scheme constitutes the lexical pattern, which can be used to extract other text fragments from a textual corpus [11]. In linguistics approaches for information extraction, patterns are used to reduce ambiguity, when associating structural regularities with semantic information.

In our case, the patterns generalize expressions identified in the texts and provide a generic formalization of particular lexical relations identified in the corpora between the *identifiers* of the "competence traces" ontology and additional contextual terms.

For each lexical relation, several occurrences from the corpora are analyzed, to build a generic abstraction, at the linguistic level. Later on, the objective is to re-use these patterns to extract new occurrences of the identifiers from new corpora without semantic ambiguity. In fact, the lexical patterns are built to deal with identified situations of semantic ambiguity linked to the *identifiers* of the domain ontology.

The 3 phases below briefly describe the procedure followed to identify and formalize patterns.

3.1 Corpus normalization and filtering of relevant sentences

The corpora considered is constituted by text extracted from websites. This first stage consists in normalizing the corpora by substituting some domain expressions (with several semantically equivalent alternatives) by a normalized term. This normalization makes easier later linguistic treatments. For example, the identification of a company can be expressed by various terms: “our company”, “our corporation”, “we”, “the name of the company”, etc. These expressions are replaced by the normalized expression “Company Represent”. Two brief examples are illustrated below (original expression then normalization):

ATTAX designs, industrializes and markets fastenings for all industries.
<Company Represent> designs, industrializes and markets fastenings for all industries
MECADEX Company is specialized in the undercutting of precision.
<Company Represent> is specialized in the undercutting of precision.

During the normalization stage, the *identifiers* of the domain ontology are also identified in the corpora. In the examples, the identifiers “Design” and “specialized” are identified using an automatic search of their lemma. For instance the lemma <specialize> recognizes all words including “specialize” in their canonic form.

After this normalization and *identifiers* localization, human experts are necessary to filter some relevant text fragments (sentences) linked to the *identifiers*, which appear relevant to identify and to formalize semantic patterns as shown in the two following steps.

3.2 Identification of representative and relevant examples

This second step consists in using the relevant text fragments previously filtered to identify semantic ambiguity situations linked to the *identifiers*. Then, examples of representative syntagms are associated with each ambiguity situation. They can help understanding the relevant meaning. Such syntagms are linguistic expressions which can and/or must be correlated with the identifier to define a relevant semantic interpretation. Only the expert is able to select and evaluate the relevance of a syntagm. Example of ambiguity recorded in the corpora:

<Company Represent> industry, based in Genas (Lyon France), is specialized in the **design and the realization of assembly machines**.

One of the main activities of *<Company Represent>* is the **design and the manufacture of telpher carriers**.

Ambiguity: how to distinguish between the design of final mechanical product and the design production equipments? When identifying firm competence traces, 2 different semantic orientations have to be considered linked to 2 distinct ontology classes.

When such ambiguity is identified for a concept, the expert extracts from the text fragments several expressions using this concept without ambiguity. As a result of this second step, each *identifier* is characterized by a set of structural linguistic diagrams, aiming at dealing with semantic ambiguity.

3.3 Pattern generation

In this 3rd step, the examples of syntagms associated to ambiguity concerning the *identifiers* are used to generate generic linguistic structures (patterns) semantically equivalent to the syntagms examples. This formalization stage is based on the expertise of a cognitive engineer, and the patterns are expressed in a formal language for further use. The application to the case study (corpora from mechanical industry) generated 35 patterns dealing with ambiguity, complemented with 100 simple patterns (no ambiguity). Below we underline 3 typical use cases of patterns for information extraction:

Simple detection of a concept: patterns constituted by a simple expression used to confirm the presence of an *identifier*. This pattern category is only applied for *identifiers* without ambiguity.

Semantic clarification of a concept: this type of pattern is constituted by a more complex linguistic scheme. The linguistic scheme makes possible to deal with ambiguity situations by checking contextual information which confirms the semantic interpretation of a text fragment.

Extraction of additional information attached to a concept: some patterns are used to extract additional relevant information linked to an i*dentifier*. For instance, with the *identifier* "specialization", we need to extract the speciality of the company and not only to detect that the company has a speciality.

As explained in section 4, these 135 patterns are first used to confirm the presence of *identifiers* (even in ambiguity situations) throughout the corpora extracted from a company website. Then, depending on the detected *identifiers*, an activation procedure will confirm or not the activation of classes of the "competence traces" ontology. This competence traces will characterise a company.

4 UNICOMP system, for competence identification

4.1 Architecture of the system

UNICOMP system is the implementation of our approach which aims at extracting company competence traces, using public information available on websites. It is constituted of four modules.

Pre-Processing: The pre-processing module extracts text from HTML pages constituting the company websites and realises a first filtering and cleaning of the text obtained. This pre-treatment has been explained in [12].

Acquisition and transcription of patterns: As described in section 3, the patterns are generated through various steps (normalisation, filtering, identification of examples and the pattern formalisation). Then, the semi-formal patterns are transformed into automatic comprehensible formal diagrams, which can be formally manipulated by a pattern matching system called UNITEX. The result of this transformation is a set of syntactic graphs (see example on the annex), constituting the pattern base associated to the identifiers, and later use to extract information on competencies.

Pattern localization: In this step UNITEX[1] is utilized to locate occurrences of patterns throughout the corpora constituted by the company websites. The result of the pattern projection on the company text is a set of occurrences. An occurrence is a valid syntactic diagram, which will be further use to validate the presence of an elaborate competence *identifier*, correlated with relevant information. However, a single pattern is not sufficient alone to confirm an *identifier* as underlined below.

Class activation: the module of class activation uses the various occurrences of patterns, located by the precedent module, to validate or not the presence of identifiers from the "competence traces" ontology. This validation called "activation" is realised by the algorithm PCA (Pattern and Classes Activation) which uses a specific protocol presented in [2]. Each class activation is generally based on the use of several patterns identified previously. The algorithm PCA exploits deterministic rules defined to confirm the effective activation of some classes depending on all the patterns identified.

The current version of UNITEX system generates syntactic patterns, via a semi-automated procedure, still requiring an expert contribution. It would be interesting to build a fully-automated method, making possible to generate patterns related to every identifier linked to the ontology concepts.

4.2 Results of the competencies extraction

Applying pattern localization and class activation, several terminal classes of the domain ontology will be considered activated by a company website. An activated class corresponds to a competence concept, the presence of which has been confirmed by the patterns. The result of the full activation process constitutes a "company competence trace" constituted by a sub-ontological tree. Fig. 2 provides an example of a competence trace structure. This figure provides a schematic representation of a part of the whole competence trace ontology. Each conceptual class is only represented by a circle. The activated conceptual classes are shown with black circles and the non-activated with white circles. The ontology has 4 conceptual levels, among which the 2 last ones (3^{rd} and 4^{th}) correspond to the domain ontology. The activation process, based on pattern localization (with patterns linked to the *identifiers*), activates the terminal classes situated on the bottom levels of the ontology (domain ontology). Then the activation is propagated upwards in the tree (each node is linked

[1] Unitex is an application which uses a set of software to treat texts in natural language by using linguistic tools

to its children by an "OR" activation link: the activation of 1 child class is sufficient to activate the parent class).

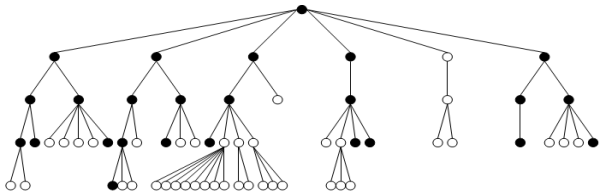

Fig. 2. Competence trace : an ontological sub-tree

4.3 UNICOMP performances

The performance evaluation of UNICOMP system is based on a comparison of the performance of the system with regards to the competence identification realized by a human expert. The human expert provides a reference of the expected result, and this reference makes possible to define the two indicators *precision* and *recall* (table 1). These 2 indicators are very common to evaluate information extraction mechanisms. Precision is defined by the capacity of the system to exclude the non-relevant classes. Recall is defined by the capacity of the system to activate the relevant classes The test was realized on a sub-collection composed by 10 companies in the mechanical industry domain.

Table 1. Indicators of precision and recall obtained by UNICOMP.

Companies	Precision	Recall
E1	0,81	0,56
E2	0,92	0,7
E3	0,87	0,5
E4	1	0,66
E5	0,75	0,54
E6	*1*	0,7
E7	*0,76*	0,83
E8	*0,8*	0,66
E9	*0,87*	0,77
E10	*0,88*	0,57
Average	***0,87***	**0,64**

The precision and the recall indicators show a good performance of UNICOMP system. They validate the pertinence of the current pattern base and ontology. However, these performances could still be improved both by a more exhaustive ontology and by a larger pattern base. A more exhaustive ontology would increase the representativity of the extraction process and thus ameliorate the recall. A higher number of linguistic patterns would improve ambiguity treatment and thus increase the precision.

5 Conclusion and perspectives

The identification of company competencies is a key factor for a decision support to build collaborative networks. The competence extraction approach adopted in this paper is based on a complex process of information extraction, using ontology and semantic patterns. This contribution has been implemented by the system UNICOMP, the performances of which have been studied. The coupling between ontology and patterns present a real added-value for the richness of information extraction. This research has already been further developed to integrate these results in a Decision Support System for collaboration Network Building using competence similarities measures [12]. As a perspective, an automatic or semi-automatic enrichment of the ontology and the pattern base, using case-learning techniques would be able to increase the overall performance of the system.

References

1. K. Hajlaoui, X. Boucher & M. Beigbeder. Construction et usage d'une ontologie de compétences pour l'identification de réseaux collaboratifs d'entreprises. Special issue "Ingénierie d'entreprise et des systèmes d'information" of the review ISI (Ingénierie des Systèmes d'Information), Edts. S. Nurcan, K. Benali, H. Pingaud, vol. 15, numéro 4, juillet-août 2010.
2. K. Hajlaoui, X. Boucher, and J.J Girardot. Competency ontology for network building. In 10th IFIP Working Conf on V. E. (PRO-VE'09). Thessaloniki, GREECE, 2009
3. E. Ermilova and H. Afsarmanesh. Modeling and management of profiles and competencies in VBEs. Journal of Intelligent Manufacturing 18, 561-586, 2007.
4. Vanderhaegen, and Loos. (2007). Distributed model management platform for cross-enterprise business process management in virtual enterprise networks. JIM 18:553-559.
5. E. Blanchard, M. Harzallah. Reasoning on competence management, Workshop on Knowledge Management and Organizational Memories of the 16th European conference on Artificial Intelligence (ECAI'04), Valence 22-27 août 2004.
6. M. Laukkanen and H. Helin, Competence management within and between organizations, Proceeding of the CAISE'05 Workshops, Enterprise Modelling and Ontologies for Interoperability Workshop, vol. 2 Porto, Portugal, June 13–17 (2005), pp. 359–362.
7. Y. Sure, A. Maedche and S. Staab. Leveraging corporate skill knowledge: from ProPer to OntoProPer, Proceedings of the 3rd International Conference P.A.K.M., Switzerland (2000).
8. O. Corby, R. Dieng-Kuntz and C. Faron-Zucker. Querying the Semantic Web with the CORESE search engine. In: R. Lopez de Mantaras and L. Saitta, Editors, Proceedings of the 16th European Conference on Artificial Intelligence (ECAI'2004), subconference PAIS'2004, Valencia, Spain, IOS Press (2004), pp. 705–709
9. B. Bachimont, A. Issac, and R. Troncy. Semantic commitment for designing ontologies. 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), LNAI 2473 :114–121, 2002.
10. N. Grabar and T. Hamon. Les relations dans les terminologies structurées : de la théorie à la pratique. Revue d'intelligence artificielle, 18(1) :57–85, 2004.
11 .M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. In In A. Zampolli, editor, Computational Linguistics (CoLing'1992), pages 539-545, Nantes, France,, 1992.
12. K. Hajlaoui, thèse, Dispositifs de recherche et de traitement de l'information en vue d'une aide à la constitution de réseaux d'entreprises. EMSE Saint Etienne, France, Dec 2009.