Extraction of terminology in the field of construction

1st Rémy Kessler
Université Bretagne Sud
CNRS 6074A
56017 Vannes,France
remy.kessler@univ-ubs.fr

2nd Nicolas Béchet
Université Bretagne Sud
CNRS 6074A
56017 Vannes,France
nicolas.bechet@irisa.fr

3rd Giuseppe Berio
Université Bretagne Sud
CNRS 6074A
56017 Vannes,France
giuseppe.berio@univ-ubs.fr

***Abstract*—We describe a corpus analysis method to extract terminology from a collection of technical specifications in the field of construction. Using statistics and word n-grams analysis, we extract the terminology of the domain and then perform pruning steps with linguistic patterns and internet queries to improve the quality of the final terminology. Results are evaluated by using a manual evaluation carried out by 6 experts in the field.**

***Index Terms*—terminology extraction, Internet queries, linguistic patterns.**

I. Introduction

The current era is increasingly influenced by the prominence of smart data and mobile applications. The work presented in this paper has been carried out in one industrial project (VOCAGEN) aiming at automating the production of structured data from human machine dialogues. Specifically, the targeted application drives dialogues with people working in a construction area for populating a database reporting key data extracted from those dialogues. This application requires complex processing for both transcripting speeches but also for driving dialogues. The first process is required for good speech recognition in a noisy environment. The second processing is required because the database needs to be populated with both right and complete data; indeed, people tend to apply a broad (colloquial) vocabulary and the transcripted words need to be used for filling in the corresponding data. Additionally, if some data populate the database, additional data may be required for completeness, thus the dialogue should enable to get those additional data (e.g. if the word "room" is recognised and used to populate the database, the location of the room must also be got; this can be done by driving the dialogue).

The application provides people with "hand-free" device, enabling a complete, quick and standardized reporting. First usages of this application will be oriented to reporting failures and problems in constructions.

The two processing steps mentioned above require on the one side a "language model" (for transcripting the sentences) and on the other side a "knowledge model" for driving the dialogue and correctly understanding the meaning of the word. The knowledge model is mainly an ontology of the domain (in this case, the construction domain) providing the standardized concepts and their relationships. As well-known, building such knowledge models needs time and is costly; one of the earlier questions raised by our industrial partners has been about "how to build, as automaticaly as possible, such a knowledge model". This question is closely related to the interest of quickly adapting the application to other domains (than the construction one) for reaching new markets. We developed a complete methodology and system for partially answering the question, focusing on how to extract a relevant terminology from a collection of technical specifications.

The rest of the paper is organized as follow. Section II present context of the project. Related work are reviewed in Section III. Section IV presents collected resources and some statistics about them. Section V describes the methodology developed for extracting relevant terms from collected resources. The details about the evaluation are presented in Section VI-A and results obtained, are given in Section VI-B.

II. Industrial context

Figure 1 presents the context of this work in VOCAGEN project. Our industrial partner Script&Go[1] develop an application for the construction management dedicated to touch devices and wishes to set up an oral dialogue module to facilitate on construction site seizure. The second industrial partner (Tykomz) develops a vocal recognition suite based on toolkit sphynx 4 [1]. This toolkit includes hierarchical agglomerative clustering methods using well-known measures such as BIC and CLR and provides elementary tools, such as segment and cluster generators, decoder and model trainers. Fitting those elementary tools together is an easy way of developing a specific diarization system. To work, it is necessary to build a model of knowledge, i.e. a model describing the expressions that must be recognized by the program. To improve the performance of the system, this knowledge model must be

[1]http://www.scriptandgo.com/en/