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/

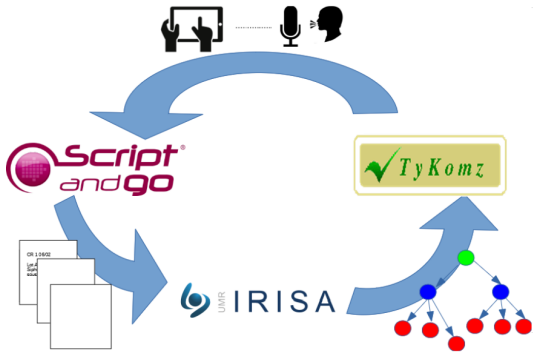

Fig. 1. figure describing the context of the project

powered by a domain-specific vocabulary. For example, in the sentence "there is a stain of paint in the kitchen", the system must understand that it is a stain of paint and that the kitchen is a room. To our knowledge, there is no ontology or taxonomy specific to the construction industry in French. A version is under development by [2] but ontology is in English and very generic. We therefore choose to extract useful knowledge from textual data, and then, in a second step, to organize them.

III. Related works

The goal of ontology learning (OL) is to build knowledge models from text. OL use NLP knowledge extraction tools to extract terminology (concepts) and links between them (relationships). The main approaches found in the literature are rule-based systems, statistical or learning based approaches.

The reference in the field of rule-based systems was developed by [3]. General Architecture for Text Engineering (GATE) is a Java collection of tools initially developed at the University of Sheffield since 1995. An alternative is offered by the existing semi-automatic ontology learning text2onto [4]. More recently, [5] developed UIMA, a system that can be positioned as an alternative to GATE. Amongst other things, UIMA makes possible to generate rules from a collection of annotated documents. Exit, introduced by [6] is an iterative approach that finds the terms in an incremental way.

[7] with TERMINAE is certainly the oldest statistic approach. Developed for French and based on lexical frequencies, it requires pre-processing with TermoStat [8] and ANNIE (GATE). [9] presents a method for extracting terminology specific to a domain from a corpus of domain-specific text, where no external general domain reference corpus is required. They present an adaptation of the classic *tf-idf* as ranking measure and use different filters to produce a specific terminology. More recently, the efficiency of ranking measure like mutual information developed for statistical approach is discussed in [10] and [11]. [12] proposes Termolator a terminology extraction system using a chunking procedure, and using internet queries for ranking candidate terms. Approach is interesting but the authors emphasize the fact that the runtime for each queries is a limiting factor to produce a relevant ranking.

Closer to our work, [13] presents an approach combining linguistic pattern and Z-score to extract terminology in the field of nanotechnology. [14] propose TAXI, which combines statistics and learning approach. TAXI is a system for building a taxonomy using 2 corpora, a generic, the other specific. It ranks the relevance of candidates by measure (frequency-based), and by learning with SVM. [15], [16] present TexSIS, a bilingual terminology extraction system with chunk-based alignment method for the generation of candidate terms. After the corpus alignment step, they use an approach combining log likelihood measure and Mutual Expectation measure [17] to rank candidate terms in each language. In the same order, [18], [19] present an approach to extract grammatical terminology from linguistic corpora. They compare a series of well-established statistical measures that have been used in similar automatic term extraction tasks and conclude that corpus-comparing methods perform better than metrics that are not based on corpus comparison. [20] and [21] present methods with word embeddings. With a small data set for learning phase, they improve the term extraction results in n-gram terms. However, these papers involve labelled data sets for learning phase, which is the main difference with our proposed approach. The originality of our approach is to combine a lexico-syntactical and a statistical approach while using external resources.

IV. Resources and statistics

First experiments were carried out using technical reports collected from some customers from our industrial partners who will be called as NC collection thereafter. Each document contains all the non-compliance that was found on one work site and describe solutions to resolve it. However heterogeneity of the formats as well as the artificial repetition of the information between two reports found in the same construction site made the term extraction quite difficult. An insightful analysis of those reports reveals vocabulary richness, however, difficult to exploit given numerous misspellings, typing shortcuts, very "telegraphic" style with verbs in the infinitive, little punctuation, few determinants, etc. As a consequence, we used a collection of technical specifications called CCTP[2]. CCTPs are available online on public sector websites[3]. Several thousand documents were collected by our industrial partner using an automatic web collecting process. Figure 2 presents some key descriptive statistics of theses collections.

Collection	NC	CCTP
Total number of documents	58 402	3665
Without pre-processing		
Total number of words	130 309	230 962 734
Total number of different words	93 000	20 6264
Average words/document	125.3	63 018.48

Fig. 2. statistics of the collection.

[2]The technical specifications book (CCTP in French) is a contractual document that gathers the technical clauses of a public contract in the field of construction.

[3]For example, https://www.marches-publics.gouv.fr/ or http://marchespublics.normandie.fr/.

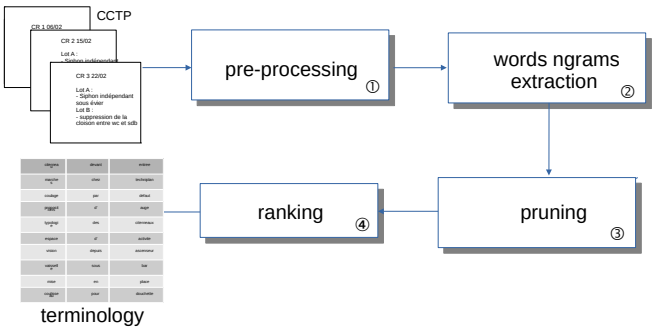

Fig. 3. System overview

V. Methodology

A. System Overview

Figure 3 presents an overview of the system designed and implemented: steps are further explained in Sections V-B to V-E. In Step 1 pre-processing of raw information extracted from CCTP collection takes place; this is required for normalizing the entire set of documents. In Step 2 n-grams are extracted (by using measures). 1,2,3 grams are extracted. In Step 3, n-grams are filtered by using linguistic patterns and Internet queries. Finally, in Step 4 a ranking is applied to the filtered n-grams.

B. Normalization, pre-processing and word ngrams extraction

In Step 1, a text normalization is performed to improve the quality of the process. We remove special characters such as "/" or "()". Different pretreatments are done to reduce noise in the model: we remove numbers (numeric and/or textual), special symbols. "." are tagged with a special character to not create artificial n-grams. Specific words (including named entities) like company names, dates, etc. are normalized and will be removed in the next module. We do not include a stop list to keep n-grams with prepositions, for the purpose described in the remainder. Then, we tokenize the entire collection before using TreeTagger [22] to get the part-of-speech tags and lemmas of each word. After this step we transform all vocabulary from the CCTP collection into 1-grams, 2-grams and 3-grams. Special characters or normalized words resulting from the previous processing are discarded. N-grams with a very low frequency (2) are also discarded.

C. Linguistic patterns module

We use grammatical labels generated in the previous step (section V-B) and linguistic patterns to retrieve collocations such as NOUN-NOUN, NOUN-PREP-NOUN. These patterns are frequently found in the literature [6] to capture specific words in French like "carte de crédit"(*credit card*) and discard 3-gram like 'créditer sa carte' (*credit his card*) with the pattern VERB-PREP-NOUN. Among frequent patterns found in literature, those patterns have been selected according to the statistics obtained from a knowledge model of another field (agriculture), given by one of our industrial partners. Figure 4 presents main patterns we selected using this knowledge model. The sum of the percentages is less than 100% because

	Number	Percentage
1-grams	1360	65.24%
NOUN	1037	76.25%
VERB	194	14.26%
ADJ	120	8.82%
2-grams	390	19.57%
NOUN-NOUN	346	88.72%
ADJ-NOUN	11	2.82%
PREP-NOUN	7	1,79%
VERB-NOUN	5	1,28%
3-grams	188	9.43%
NOUN-NOUN-NOUN	150	79.79%
PREP-NOUN-NOUN	15	7.98%
NOUN-PREP-NOUN	6	3,19%
VERB-NOUN-NOUN	6	3,19%

Fig. 4. Distribution of linguistic patterns according to the knowledge model.

patterns with a very low frequency are not included in the table. We observed that the noun based patterns are the most frequent patterns, whatever the size of the n-gram. The other selected patterns also contain nouns, but they are n-grams with verbs, adjectives or prepositions. Therefore, we have configured our system to keep only the ngrams corresponding to these patterns.

D. Pruning step

This step uses the Internet to prune n-grams for which no information is returned after querying Bing[4] search engine. We count the number of links in the result pages that contain exactly the ngram. We save the number of exact matches between the ngram and the title and snippet of each result. We keep only the n-grams whose number of matches exceeds a defined threshold. We varied this threshold between 1 and 50 and results presented in Section VI-B have been obtained with a threshold of 10.

E. Ranking step

We tested several measures as provided in [6], [16] like maximum likelihood estimation or mutual information in order to rank selected n-grams by quality but results were disappointing. We finally use classical *Z score* [23] with twenty years of the French newspaper Le Monde[5] as generic collection. This metric considers word frequencies weighted over two different corpora, in order to assign high values to words having much higher or lower frequencies than expected in a generic collection. We defined it as follows :

$$p_1 = a_0/b_0 \quad (1)$$

$$p_2 = a_1/b_1 \quad (2)$$

[4]https://www.bing.com/

[5]http://www.islrn.org/resources/421-401-527-366-2/

$$p = (a_0 + a_1)/(b_0 + b_1) \quad (3)$$

$$Z_Score = \frac{p_1 - p_2}{\sqrt{(p * (1-p) * (\frac{1}{b_0} + \frac{1}{b_1})}} \quad (4)$$

Where a is the lexical unit considered (1-gram, 2-gram or 3-gram), a_0 the frequency of a in the CCTP collection, b_0 the total size in words of CCTP collection, b_1 the frequency of a in the collection Le Monde.

VI. Experiments and results

A. Experimental protocol

We have made a manual evaluation on all 3 grams retained by the system. Manual evaluation was realized by 6 specialists in the field of construction. Each specialist evaluating one third of the results. 5144 3-grams were evaluated with this method and each n-gram was evaluated by 2 different specialists. For each n-grams, the specialist can choose between three possibilities:

- 0 3-gram is irrelevant
- 1 3-gram is relevant but does not belong to the domain
- 2 3-gram is relevant and belongs to the domain

Evaluation was done in two steps and we use Kappa measure[6] [24] and inter-annotator agreement at the end of the first step to show the difficulty of the task. At the end of the first step, we obtained a Kappa score of 0.62 and a global inter-annotator agreement of 0.74, which is quite good as explained in [25]. The difficulty of the task was to distinguish the domain-specific vocabulary from the generic vocabulary used in the field of construction. Each disagreement was re-evaluated in the second step by a pair of experts. Figure 5 shows the final results of the evaluation.

B. Results

In this section, we present the results obtained during the manual evaluation of the 3-grams retained by the system. We only compute the accuracy and the error rate, because we are not able to compute the recall for this collection[7]. We have merged the assessments of each expert using two different evaluation rules:

- a strict evaluation where a n-gram is considered correct if both experts have rated it relevant and in the domain .
- a flexible evaluation where a n-gram is considered correct if both experts consider it relevant and at least one of the experts consider it in the domain.

	strict evaluation	flexible evaluation
accuracy	0.77	0.91
error rate	0.23	0.09

Fig. 5. Results of manual evaluation on the 3-grams.

[6]We use general formula as follows : $\kappa = \frac{A_0 - A_e}{1 - A_e}$ where A_0 = observed agreement and A_e = expected (chance) agreement.

[7]Indeed, we do not know every the relevant terms existing in the corpus, so we cannot estimate the recall for the collection of terms we automatically extract.

Strict evaluation shows good quality results (0.77). Analysis of the results shows that the main error is related to "incomplete n-grams". For example, the 3-gram "personne à mobilité" (person with mobility) is not relevant while the 4-gram "personne à mobilité réduite" (person with reduced mobility) can belong to the field of construction. Some errors can also be traced back to the CCTP documents. For example, "engin de guerre" (war machine) is a term which does not belong to the field but a law relating to the presence of war machine on the building sites is reported in every CCTP. The flexible evaluation shows very good results (0.91) and the difficulty of assessing class of some terms such as "absence de remise" which has 2 distinct meanings in French (no outhouse and no discount). The first meaning is relevant in the field of construction but not the second.

VII. Conclusion and perspectives

The paper reports our experiments and results for building a precise and large terminology for the construction domain. Collecting terminology is indeed the first step towards a complete knowledge model containing both concepts and relationships. During our work we were faced to several problems: finding resources and selecting them for building an appropriate corpus, thinking and developing pre-processing for cleaning those resources, experimenting distinct measures for n-grams and selecting the most appropriate, improving results by adding linguistic patterns and Internet queries. The current results are quite promising according to the evaluation of the extracted terminology carried out by 6 experts in the field. As a perspective, we will develop generic modules and guidelines for adapting these pre-processing modules to other languages. Most importantly, the results of our work are useful for extracting taxonomical and non-taxonomical relationships. For the both purposes, we are currently working on SemEval collection [26]. Applying our method to other domain corpora and datasets is another future direction for this research.

References

[1] S. Meignier and T. Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *in CMU SPUD Workshop*, 2010.

[2] P. Pauwels and W. Terkaj, "Express to owl for construction industry: Towards a recommendable and usable ifcowl ontology," *Automation in Construction*, vol. 63, pp. 100–133, 03 2016.

[3] H. Cunningham, "Gate, a general architecture for text engineering," in *Computers and the Humanities*, vol. 36, 2002, pp. 223–254.

[4] P. Cimiano and J. Völker, "text2onto," in *International conference on application of natural language to information systems*. Springer, 2005, pp. 227–238.

[5] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, "Uima ruta: Rapid development of rule-based information extraction applications," *Natural Language Engineering*, vol. 22, no. 1, p. 1–40, 2016.

[6] M. Roche and Y. Kodratoff, "Exit: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés," in *Proceedings of JADT 4*, 2004, pp. 946–956.

[7] B. Biébow, S. Szulman, and A. J. B. Clément, "Terminae: A linguistics-based tool for the building of a domain ontology," in *Knowledge Acquisition, Modeling and Management*, D. Fensel and R. Studer, Eds., 1999, pp. 49–66.

[8] P. Drouin, "Term extraction using non technical corpora as point of leverage," in *John Benjamins Publishing Company: Amsterdam/Philadelphia*, n. Terminology, vol. 9, Ed., 2003, pp. 99–115.

[9] M. F. M. Chowdhury, A. M. Gliozzo, and S. M. Trewin, "Domain-specific terminology extraction by boosting frequency metrics," Sep. 27 2018, uS Patent App. 15/469,766.

[10] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, 2009.

[11] Y. Bestgen, "Evaluation de mesures d'association pour les bigrammes et les trigrammes au moyen du test exact de fisher," *Proceedings of TALN 2017*, pp. 10–19, 2017.

[12] A. L. Meyers, Y. He, Z. Glass, J. Ortega, S. Liao, A. Grieve-Smith, R. Grishman, and O. Babko-Malaya, "The termolator: Terminology recognition based on chunking, statistical and search-based scores," *Frontiers in Research Metrics and Analytics*, vol. 3, p. 19, 2018.

[13] L. Gillam, M. Tariq, and K. Ahmad, "Terminology and the construction of ontology," *TERMINOLOGY*, vol. 11, pp. 55–81, 2005.

[14] A. Panchenko, S. Faralli, E. Ruppert, S. Remus, H. Naets, C. Fairon, S. P. Ponzetto, and C. Biemann, "TAXI at semeval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling," in *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, 2016, pp. 1320–1327.

[15] E. Lefever, L. Macken, and V. Hoste, "Language-independent bilingual terminology extraction from a multilingual parallel corpus," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, ser. EACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 496–504. [Online]. Available: http://dl.acm.org/citation.cfm?id=1609067.1609122

[16] L. Macken, E. Lefever, and V. Hoste, "Texsis: bilingual terminology extraction from parallel corpora using chunk-based alignment," *Terminology*, vol. 19, no. 1, pp. 1–30, 2013. [Online]. Available: http://dx.doi.org/10.1075/term.19.1.01mac

[17] G. Dias and H.-J. Kaalep, "Automatic extraction of multiword units for estonian : Phrasal verbs," in *Languages in Development*, 2003, p. 41:81–91.

[18] B. Daille, "Study and implementation of combined techniques for automatic extraction of terminology," *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 12 2002.

[19] C. Lang, R. Schneider, and K. Suchowolec, "Extracting specialized terminology from linguistic corpora," *GRAMMAR AND CORPORA*, p. 425, 2018.

[20] E. Amjadian, D. Inkpen, T. S. Paribakht, and F. Faez, "Local-global vectors to improve unigram terminology extraction," in *Proceedings of the 5th International Workshop on Computational Terminology*, 2016.

[21] G. Wohlgenannt and F. Minic, "Using word2vec to build a simple ontology learning system." in *International Semantic Web Conference (Posters & Demos)*, 2016.

[22] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," 1994.

[23] E. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." in *The Journal of Finance, 23(4). doi:10.2307/2978933*, 1968, pp. 589–609.

[24] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960.

[25] M. L. McHugh, "Interrater reliability: the kappa statistic," in *Biochemia medica*, 2012.

[26] M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, and M. Carpuat, "Proceedings of the 12th international workshop on semantic evaluation," in *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2018. [Online]. Available: http://aclweb.org/anthology/S18-1000