Large-Scale Knowledge Acquisition from botanical texts

François Role[1], Milagros Fernadez Gavilanes[2], and Éric Villemonte de la Clergerie[3]

[1] L3i, Université de La Rochelle, France
Francois.Role@univ-lr.fr
[2] University of Vigo, Spain
mfgavilanes@uvigo.es
[3] INRIA Rocquencourt, France
Eric.De_La_Clergerie@inria.fr

Abstract. Free text botanical descriptions contained in printed floras can provide a wealth of valuable scientific information. In spite of this richness, these texts have seldom been analyzed on a large scale using NLP techniques. To fill this gap, we describe how we managed to extract a set of terminological resources by parsing a large corpus of botanical texts. The tools and techniques used are presented as well as the rationale for favoring a deep parsing approach coupled with error mining methods over a simple pattern matching approach.

Introduction

In this paper we are concerned with new methods for making more readily accessible the large amount of information contained in existing printed floras. Floras are books that contain descriptions of taxa (plant species, genus, etc.) occurring in a particular geographic area. A taxon description usually includes short sections relating to nomenclature, ecology, geographical distribution, and a free-text account of the plant morphology. This information is crucial to scientists in the field of botany. Traditional printed floras published before the computer age are rich documents containing a wealth of information which is still of great value but is difficult to search and exploit. In spite of this fact, not many investigations address the problem of the digitization of these legacy texts. Among the few projects that resemble our endeavor is the digitization of Flora Zambeziaca conducted at the herbarium of the Royal Botanic Garden Kew. Flora Zambeziaca is a large African flora covering about 8500 species and available as a set of printed volumes published from 1960 onwards [1]. Other similar projects include Flora of Australia online and Flora of North America. All these systems use the information contained in the floras only to implement online databases that can be searched by traditional access points (scientific name, synonyms, geographical location, etc.) However, the morphological description sections are left unanalyzed and can only be searched using basic full-text techniques.

In contrast to these approaches, we propose to mine these textual portions in order to extract terminological resources using statistical and NLP techniques.

These terminological resources will be used to help users make more precise queries against digitized botanical texts. They could also be used to serve as a starting point for the creation of domain ontologies, a future goal of our research.

This research was conducted as part of the "Biotim" project [2] on processing botanical corpora. Work concentrated on the *Flore du Cameroun* [**FdC**], an African flora comprising 40 volumes, each volume running to about 300 hundred pages. Although published between 1963 and 2001, the **FdC** exhibits a relatively regular structure. Each taxon description comprises a set of paragraphs. Most of these paragraphs (sections relating to author and collector, bibliography, ecology, distribution, etc.) are short, stereotyped fields which are in principle easy to recognize and analyze using pattern matching. One exception, however, is the description section which provides a detailed free-text account of the main morphological features of a taxon. This section is a (possibly) long text, which consists of several sentences which may have a complex syntactic structure. Analyzing such textual content is a task that requires the use of linguistic resources and tools. However it is a necessary step to fully exploit the richness of the information contained in the Flora.

The paper is organized as follows : In section 1 we describe how we were able to derive the logical structure from the digitized text. Section 2 discusses the various approaches we developed to better exploit the complex content found in the free-text sections of taxonomic descriptions. Finally, we present a comparison with related work and conclude.

1 Capturing the Logical Structure

Starting from the digitized printed pages, the logical structure of the 37 volumes of **FdC** was retrieved using Perl regular expressions. The different sections of the plant descriptions (author, name, bibliography, type, distribution, ecology, material, morphological description) were recognized and marked-up in XML.

The documents have been stored in an online repository. They can be browsed via an interface where XML elements can be expanded and collapsed by clicking on icons. We also experimented with shredding and storing the XML documents into the columns of an object-relational database which could then be searched using XML query languages such as XQuery. Another benefit of capturing the logical structure is that we are then in a position to easily derive a semantic web-compatible representation of hierarchical relationships between taxa. Once converted into an XML document, the marked-up text so obtained can be fed into a simple XSLT program which generates a hierarchy of OWL classes mirroring the taxonomic hierarchy. This can be seen as the starting point for the construction of a domain ontology.

2 Analyzing the Textual Content

Having captured the logical structure we are then able to identify and target for analysis the free text sections that describe the plants in terms of their physical features. As said in the introduction, these free-text descriptions are often poorly exploited. They usually contain a separate sentence for each major plant organ. Marking up this implicit structure provides a context to identify adjectives specific to each organ. For example, adjectives like "*ligulée-lancéolée*" (ligulate-lanceolate) or "*bilobée*" (bilobate) are suitable for describing a leaf while "*multiflore*" (multiflora) is appropriate for describing an inflorescence.

Having available a list of the adjectives that are used to describe specific parts of a plant may help users formulate more precise queries against the free-text description section. In order to create this terminology we performed a morpho-syntactic analysis of the sentences in the description section. We used the tools developed by the INRIA ATOLL team to generate morpho-syntactic annotations in an XML format compliant with the MAF proposal. The tagging tools proved very reliable in unambiguously detecting punctuation marks, thus allowing us to segment each description section into sentences relating to a certain organ. Using simple XPath expressions, we then searched each sentence for all adjectives or past participles that agree in gender, in case and number with the noun at the beginning of the sentence. We used this technique on three volumes of the flora dealing with the orchid family, which accounts for about 10% of the taxa described in the Flore du Cameroun.

Overall, more than 400 adjectives appropriate for describing the physical features of a tropical orchid were identified and validated. Again we encoded the obtained resource in OWL as a potential starting point to develop a domain ontology derived from the flora. This shallow parsing strategy enables us to identify dependencies between nouns and adjectives that are adjacent or very near to each other within the text. However, it fails in modeling long-range dependencies. By way of an example, in the case of a sentence such as "*feuille charnue à nervure étroite*" (fleshy leaf with narrow vein) the shallow parsing strategy described above does not allow to know if "*étroite*" (narrow) relates to "*feuille*" (leaf) or to "*nervure*" (vein). In fact, this approach even fails to detect that "*nervure*" is a component of "*feuille*" and more generally is not appropriate for dealing with suborgans names that are deeply nested within the sentences. Thus the surface form of the text directly governs our ability to pinpoint occurrences of organ names, a situation which is not desirable. We also experimented with third-party tools (ACABIT, FASTR) but still failed to detect long-range dependencies. Finally, besides not being adequate for detecting long-range dependencies, shallow parsing techniques for information extraction generally rely on hand-crafted extracting patterns. Designing these patterns is a costly task which often requires the help of a domain expert and has to be redone for each new domain and often tailored for slightly different corpora in a same domain (style variations, different authors, ...)

We have considered that these issues could be solved by adapting a generic French deep parser while trying to minimize the amount of required human intervention. The tuning of the parser was facilitated by both the use of Meta-Grammars and of error mining techniques. Indeed, our French grammar, named FRMG, is compiled from a source Meta-Grammar, which is modular and hierarchically organized. It is trivial to deactivate phenomena that are not present in the corpora (for instance clefted constructions, questions, imperative, . . .).

On the other hand, error mining techniques have been used to quickly spot the main sources of errors and then fix the grammar accordingly. The idea behind this is to track the words that occur more often that expected in sentences whose analysis failed. Those words may be obtained by using a fix-point algorithm and generally indicate some kind of problem, often related to lexical entries but sometimes to segmentation or grammatical issues. The algorithm is as follows. We note p_i the i-th sentence in the corpus and $O_{i,j}$ the j-th word in the i-th sentence. $F(O_{i,j}) = f$ denotes the form of occurrence $O_{i,j}$. Let $S_{i,j}$ be the probability that word occurrence $O_{i,j}$ was the reason why the analysis of a given sentence p_i failed. We first estimate $S_{i,j}$ using the formula $S_{i,j} = \mathrm{error}(p_i)/|p_i|$ where $\mathrm{error}(p_i)$ returns 1 if the analysis failed and 0 else. It means that, for a sentence whose analysis failed, we first assume that all its word occurrences have the same probability to be the cause of the failure. We then use these (local) estimations of $S_{i,j}$ to compute $S_f = \frac{1}{|O_f|} \cdot \sum_{O_{i,j} \in O_f} S_{i,j}$ where $O_f = \{O_{i,j} | F(O_{i,j}) = f\}$. It is an estimation of the average failure rate of the form $f = F(O_{i,j})$. We then use this (global) estimation of S_f to refine the initial estimation of $S_{i,j}$ and so on until convergence. Considering the sentences which receive at least one full parse, FRMG usually achieves a coverage 40 to 50% percent on general corpora. Without adaptation, FRMG only attained an initial 36% coverage on **FdC**. Nevertheless, by parsing the whole corpus (around 80 000 sentences depending on sentence filtering) 14 times and exploiting the feedback provided at each round using the error mining techniques, we were eventually able to improve the overall performance of the parsing from 36% to 67%. Each round took around a night processing on a local cluster of 6-7 PCs.

Once the grammar and vocabulary have been adequately tailored, our parser returns a shared forest of dependencies for each successfully parsed sentence. A dependency is a triple relating a source governor term to a target governed one through some syntactic construction provided as a label. According to Harris distributional hypothesis that semantically related terms tend to occur in similar syntactic contexts, examining the dependency edges that enter or leave nodes in the graph should allow us to extract terms that are semantically related. The problem is that even in a successfully parsed sentence, ambiguous dependencies may remain, due to the fact that several derivations may be possible and several syntactic categories may still be in competition for assignment to a word. At first glance, it seems that some of these ambiguities can be eliminated by using linguistic markers such as range constructions "*X à Y*" [X to Y], where both X and Y are adjectives (such as in the phrase "yellow to orange"), which implies that X and Y belong to the same semantic class, or very explicit linguistic markers such as "*coloré en X*" (X-colored), "*en forme de X*" (in form of X). However the number of these markers is limited, not to mention that they may be ambiguous (for example the range construction in French in competition with prepositional attachments). Second, from the beginning of the project we have sought to develop solutions that are not too corpus specific.

We therefore came upon the idea of developing a statistical iterative algorithm inspired from our error mining techniques to converge toward a better classification. The main idea is to locate the dependencies that are the most frequent at the corpus level. As for error mining, this "global knowledge" is then reused locally (at the sentence level) to reinforce the weight of some local dependencies at the expense of competing ones. The updated local weights can then be used to re-estimate the global weights and so on. After convergence, we globally get the most probable dependencies for high-frequency terms and, therefore, the most probable syntactic contexts for a term. Preliminary results have been obtained trying to classify terms into "organs", "properties" (in general), and "others", initiating the process with only 6 seed terms describing organs : "*feuille*" (leaf), "*pétale*" (petal), "*fruit*" (fruit), "*ovaire*" (ovary), "*rameau*" (small branch), and "*arbre*" (tree) and around ten properties. Actually, a weight $w_c(t)$ is attached to each term t for each class c, and we order following $w_c(t) * ln(\#t)$ to favor high frequency terms. Indeed, we are mostly interested in getting some knowledge about the most frequent terms, and, moreover, the results are statistically less significant for low frequency terms. Table 1 lists the 10 best terms for each category. The results seem satisfactory, but for two entries that reflect segmentation issues. It is worth observing that some of the best terms for the "other" category are good candidates to be property introducers ("*forme*"/shape, "*couleur*"/color, "*taille*"/size, "*hauteur*"/height, ...).

organs	**properties**	**other**
nervure (vein)	oblong (oblong)	diamètre (diameter)
fleur (flower)	ovale (oval)	longueur (length)
face (face)	ovoïde (ovoid)	hauteur (height)
feuille (leaf)	elliptique (elliptical)	largeur (width)
limbe (limb)	glabre (hairless)	taille (size)
rameau (small branch)	lancéolé (lanceolate)	forme (form)
sommet (apex)	ellipsoïde (ellipsoid)	forêt (forest)
sépale (sepal)	globuleux (globular)	* d
foliole (foliola)	floral (floral)	couleur (color)
base (base)	aigu (acute)	* mètre (meter)

Table 1. The best-ranked terms found by our method.

Overall the top 100 terms in each category seem to be correct, an intuition that we hope to confirm in the following way. In the first step, the automatically extracted list of terms will be compared to existing terminological resources. A particularly valuable resource for this purpose is the thesaurus of the French Society of Orchidophily recently put at our disposal. Terms not found in the thesaurus will be marked as such and sent to experts from two national herbaria (Cameroun and Senegal) for advice and validation, and standard statistical methods such as kappa will be used to quantify inter-rater agreement.

3 Comparison with Related Work

Most systems designed to derive clusters of related words mostly use shallow parsing [3,4,5,6]. Those relying on deep parsers usually adopt a sequential approach. First a sentence is parsed using a trained parser and then the head of each constituent of the sentence is identified using patterns [7]. As a variant, [8] uses a parser that directly generates dependency trees. In both cases it is assumed that the parsers have been accurately trained and can generate high accuracy parses, thus providing a sound basis for the dependency extraction.

In contrast, we accept that the parsing step returns errors, and we rely on statistical methods to progressively tune a generic untrained parser. This work also provided the opportunity to learn from non-verbal structures (noun phrases) whereas other research mainly focuses on dependencies between a verb and its arguments, to detect verbs that denote the same ontological relations. Last but not least, adapting the error mining techniques in order to exploit dependencies minimizes the need to design domain-dependent rules.

4 Conclusion

We have explored new ways for exploiting the rich scientific information found in botanical texts. To the best of our knowledge it is the first time that advanced linguistic tools have been applied to analyze the text descriptions found in printed floras. A combination of NLP and statistical tools allowed us to produce terminological resources. These resources can now be utilized for several purposes ranging from helping users make more precise queries against botanical databases to producing domain ontologies in the field of botany.

References

1. Kirkup, D., Malcolm, P., Christian, G., Paton, A.: Towards a digital african flora. Taxon **54**(2) (2005) 457–466
2. Rousse, G., Villemonte de La Clergerie, E.: Analyse automatique de documents botaniques: le projet Biotim. In: proc. of TIA'05, Rouen, France (April 2005) 95–104
3. Daille, B.: Terminology mining. In (ed), M.P., ed.: Information Extraction in the Web Era. Lectures Notes in Artifial Intelligence. Springer (2003) 29–44
4. Faure, D., Nédellec, C.: ASIUM: learning subcategorization frames and restrictions of selection. In: Proc. of the 10th Conference on Machine Learning (ECML 98) Workshop on Text Mining. (1998)
5. Grefenstette, G.: Explorations in Automatic Thesaurus Construction. Kluwer (1994)
6. Cimiano, P., Staab, S., Hotho., A.: Clustering ontologies from text. In: Proceedings of LREC'04. (2004) 1721–1724
7. de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proc. of LREC'06. (2006)
8. Lin, D., Pantel, P.: DIRT - discovery of inference rules from text. In: Proceedings of KDD-01, San Francisco, CA (2001) 323–328