Corpus study of kidney-related experimental data in scientific papers

Brigitte Grau, Anne-Laure Ligozat, Anne-Lyse Minard
IBISC, Tour Evry 2
91000 Evry, France
grau, ligozat, minard @ensiie.fr

Abstract

The Quantitative Kidney DataBase (QKDB) is a relational database that was created in order to centralize kidney-related experimental results. Each result is characterized by different attributes and the scientific paper from which it was extracted. Currently, this database is populated by hand by experts of the domain. We present a corpus study of some papers that have already been analyzed in order to exhibit the specificities and difficulties of the extraction process; then we propose a first solution to extract automatically the desired knowledge from papers.

Keywords

Information extraction from scientific papers, database populating, kidney experimental results

1 Introduction

The goal of the Quantitative Kidney DataBase (QKDB) project, as described on the web site[1], is to make kidney-related physiological data easily available to the scientific community. The emphasis is on experimental results relevant to quantitative renal physiology, with a particular focus on data relevant for evaluation of parameters in mathematical models of renal function. The vast collection of heterogeneous experimental data is necessary not only for evaluation of the many parameter values but also for experimental and clinical validation of the simulation results.

QKDB thus contains experimental results extracted from scientific articles in renal physiology. Currently, these experimental results are manually introduced in the database. Each result is described by several attributes, whose values are found in the text. Thus, the manual process consists in finding all the relevant results and their characteristics in a paper, by highlighting them in the analyzed paper, and then entering them in the database. Table 1 presents QKDB records for the following experimental results:

Mean arterial blood pressure of the anesthetized mice was 99.3±5.4 mmHg in wild-type, 90.5±2.9 mmHg in heterozygous, and 79.5±5.9 mmHg in homozygous mice.

In addition, a curator, that is an expert of the domain, has to verify the validity and the coherence of the data. In particular, the different values given to features must be chosen at the right level of granularity and must not be present with different forms (synonyms or acronyms for example).

This process is a heavy task, and currently only 300 papers have been processed, although there are thousands of relevant articles.

Our project aims at providing a tool that will help the expert when processing a text [4, 2]. Even if there are many works in information extraction, few of them are dedicated to designing an assistant tool. This purpose leads us to always keep a link between the information extracted and the source text, in order to navigate easily from database to text and conversely from text to database.

Our tool will propose the curator expert each result given in the paper, with its contextual description, and either the expert will validate the data, or he will enter other values. Thus, the problem can be decomposed into two tasks:

- selecting relevant results;
- highlighting the including passages and the values of the descriptors for a selected result.

The information we look for can be modelled by a template that represents the description of an experimentation in kidney studies. Even if many systems apply IE techniques to scientific papers, they are generally dedicated to the domain of molecular biology and they often look for specific entities and some relations between these entities and not for a complex template. We can find such a problem in systems (see for example [3]) issued from MUC evaluations [6], in which most entities were named entities such as *person*, *organization* or *location* names, that can be recognized using gazetteers and rules relying on linguistic features. In our case, if the result value corresponds to a named entity, other descriptors are domain-specific terms, whose recognition would require to refer to an ontology dedicated to this domain that does not exist currently. Furthermore, it also requires the modelling of the relations between an experimentation and each of its descriptors.

Most systems only use abstracts for the extraction task; only few of them analyse full-length papers [2, 5]. One of the reasons is that corpora are difficult to convert into usable plain text format. However, the systems analyzing full-length papers obtain better results than by using only the abstract. This is confirmed by

[1] QKDB website: http://physiome.ibisc.fr/qkdb/

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 21–26

Paper id	Qualitative data	Value	Parameter	Species	Organ	Region	Comment
124	Mean arterial plasma	99.3 ± 5.4mmHg	blood pressure	mouse	kidney	arterial plasma	wild-type mice
124	Mean arterial plasma	90.5 ± 2.9mmHg	blood pressure	mouse	kidney	arterial plasma	heterozygous mice
124	Mean arterial plasma	79.5 ± 5.9mmHg	blood pressure	mouse	kidney	arterial plasma	ACE KO mice

Table 1: *Examples of QKDB records (as displayed in QKDB web interface)*

Shah's results [8]. The authors made some measures of keywords in papers according to the section they belong to. They show that although the abstract contains many keywords, other sections of the papers are better sources of biologically relevant data. Nevertheless, these sytems look for singular relations, described by patterns, and do not aim at filling complex templates. So, they do not have to gather the right information from the whole text. In our case, abstracts cannot be used at all, because they do not convey results, and we must search for them in the whole text.

Thus the realization of an assistant for extracting information that combines search in full length papers and the filling of complex templates appears to be a complex task that presents several difficulties such as the recognition of the relevant terminology, the retrieval of pieces of information in the whole text, and, given that papers describe several experimentations that share common descriptors and differ with other ones, the selection of the relevant information according to a specific result.

So, we first conducted a corpus study in order to specify the extraction task and pinpoint its specificities and its difficulties. We developed our corpus study based on the existence of the QKDB database (see section 2). As we possessed the data in the database on one hand, and the papers from which they were extracted on the other hand, we developed a tool that automatically projects the data into the texts, i.e. that retrieves and annotates QKDB values linked to each experimental result. This first step allowed us to realize a study concerning the ambiguities of the terms and their variations between the database and the texts either by visualizing the data or by computing quantitative criteria we have defined for characterizing the task (see section 3). This projection was developed on a subset of the database papers.

The second step consisted in developing a first extraction system with extraction rules, based on the projection tool and our study. This extraction process was evaluated on another subset of papers of the QKDB database, and this provides us a baseline for evaluating further developments (see section 4).

2 Description of the corpus

2.1 Database

QKDB contains around 300 scientific articles concerning kidney-related physiology from several journals (such as the American Journal of Physiology - Renal Physiology or the Journal of Clinical Investigation). More than 8000 experimental results were manually extracted from these articles by biologists. Each result is described by several parameters: quantitative value, species, organ... In QKDB, four main tables represent these results :

- the table *source* represents an article with its authors, title, publication year...
- each result is stored as a tuple of the table *record*, which contains the result value, the unit, experimental conditions...
- table *field_type* contains the link between a *field_type* number and its description: for example, species correspond to *field_type* 1, while organs correspond to *field_type* 2.
- the other parameters describing the result are stored in the *field* table: *mouse* is an instance of a *field* with *field_type* 1 (species), as well as *arterial plasma*, with its acronym *AP*, and its *field_type* 7 corresponding to a region. Several fields are associated to each result to describe the experimental conditions in which it was obtained.

2.2 Articles

The articles are stored in a PDF format in QKDB. Each article is generally composed of several sections: title, authors, abstract, methods, results, discussion and references. The results can be given either in the body of the article, or in tables. For our study, we needed to process the articles in a plain-text format, while keeping their logical structure that we represent with an XML structure. The conversion of PDF articles to an XML format is being studied, but some elements are difficult to extract from PDF, such as tables or special characters (for instance, ±). Thus, XHTML versions of articles in QKDB were retrieved from the web, which is possible for some of the most recent articles. This sub-corpus is presently composed of 20 articles.

The articles were transformed into an XML format, which contains the following tags: title, authors, body of the article, paragraphs, tables (with rows and columns), and footnotes. This corpus contains about 933 QKDB records.

2.3 Description of the experiments

We look for experimental data in the articles, which can be composed of the following information:

- a result value, which is the numerical value measured in the experiment;
- a unit of measurement, which qualifies the numerical value;
- a precision, which usually indicates the standard error of the measure;
- the number of animals on which the experiment was performed;
- qualitative data, which describe qualitatively the result;
- a comment, which gives additional information, for example about the species or their treatments;

These are all attributes of the table *record* in QKDB; they do not have predetermined values. The following characteristics on the contrary have fixed values; they correspond to the tuples of the *field_type* table.

- the species on which the experiment was performed;
- the organ, region, tube segment and epithelial compartment (and possibly the cell type), which are the locations of the experiment;
- a parameter, which indicates what property was measured (weight, permeability, inner diameter, concentration...);
- the solute, which indicates what was measured (for example *HCO3-* if its concentration was measured).

All these characteristics form slots of a complex template for an experiment. They may not be filled in for some records; only the result value is mandatory.

Here is an example of a sentence containing results:
(...) serum osmolality increased to 517 mOsm compared with 311-325 mOsm in wild-type and heterozygous mice.

It can be noticed that some of the information entered in QKDB comes from the sentence itself (value, parameter), but some of it is also inferred from the rest of the article (species, organ...).

The objective of our project is to be able to automatically annotate such an experiment result in the texts, that is to extract fillers for each slot of the experiment template. Yet, a first step consisted in studying the type of information to annotate and its expression in the articles. Since the links between the records and their expression in the texts was lost when the database was populated, we had to recreate them by projecting QKDB records into the texts (see Fig. 1) in order to create an annotated corpus.

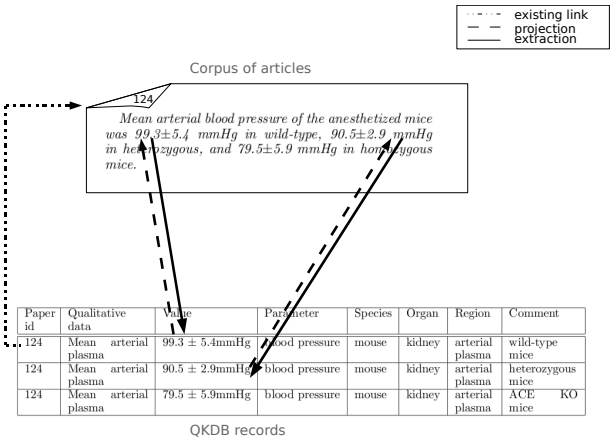

Fig. 1: *Projection and extraction of QKDB records*

3 Projection of QKDB tuples

As was shown before, a QKDB record is composed of a result value, and several other slots which describe this value: species, organ, parameter... The values of these slots can be far from the result value in the article. The objective here was thus to project QKDB records values in the articles and to annotate the words expressing them, in order to study the way results are expressed and the way their description is given. The value of the result is often expressed the same way in the article and in the database, since numerical values present few variations. Yet, there might be ambiguities when projecting the numerical value on its own to the text: a figure such as 2 might have several occurrences in the text, aside from being a result value. Thus, some filters have been applied and the measurement unit has to be found to disambiguate such numerical values. However, these units are not always present next to the experimental result, for example in tables, where the unit can be expressed in a column title, and the value in the same column, but a different row.

The projection script first detects result values from QKDB records in the text bodies. Then, it retrieves the values of the attributes associated to them, with QKDB fields that act as gazetteers and we only take into account variations in number. The following example shows an excerpt of annotated text:

serum <results tuple="3" type="parameter"> osmolality </results> increased to <results tuple="3" type="result_value"> 517 </results> <results tuple="3" type="unit"> mOsm </results> compared with 311 mOsm in wild-type mice.

With this basic projection, several studies have been conducted by projecting either attribute values linked to one measure at a time in an article, or the whole database values in the article. The goals were to:

- show the type of information which was considered relevant for the specialists who populated the database (in which sections are the results and their descriptors found?);
- detect the records that were not projected into the article because a different form was used (silence), and conversely the records that were wrongly recognized (ambiguities);
- indicate the position of the records describing a result with respect to the value of the result (in the same sentence, the same paragraph or in the rest of the text);

- estimate the quantity of noise, i.e. QKDB values that are in the article, but are not linked to a record extracted from this article.

3.1 Evaluation

The study has been conducted on 5 papers, that represent 95 templates. The results of the projection were studied quantitatively, based on the XML formats. In order to make manual and qualitative studies, an easily-readable format was also constructed: the XML formats were translated into XHTML formats, with a CSS stylesheet and an XSLT processor. The result is a XHTML file which highlights the results and their descriptors, with different colors depending on the type of descriptor, as Fig. 2 shows.

Functional analysis of AQP1 in kidney was done using purified apical membrane vesicles from proximal tubule epithelium. Osmotic water permeability (P_f) was measured from the time course of 90° scattered light intensity in response to a rapidly imposed osmotic gradient. P_f was decreased strongly in membranes from knockout versus wild-type mice with intermediate P_f for heterozygotes (Fig. 2A, P_f ~ 0.018 ± 0.002 cm/s wild-type; 0.010 ± 0.001 cm/s heterozygous; 0.002 ± 0.0003 cm/s knockout). After inhibition by 0.2 m M HgCl 2 , P_f in membranes from wild-type and heterozygous mice was similar to P f in knockout mice (0.001–0.002 cm/s, Fig. 2B). This low P_f is typical of lipid bilayer membranes not containing water channels, indicating that AQP1 is the principal functional water transporter of proximal tubule apical membrane.

Fig. 2: *Article showing QKDB records*

In order to make systematic studies, we also developed an evaluation script. A reference corpus was created (based on a first version of the articles annotated by the projection tool), in which all QKDB records are annotated in the texts. A script then evaluates the difference between the results obtained by the projection (or later the extraction) and the corpus in which all records are annotated.

3.2 Results of the projection

3.2.1 Location of relevant information

We expected the results to be mostly in the Results or Discussion sections, and indeed, 94% of the results are. The content of the comment slot is usually in the Methods section. This will help detect the relevant information, by giving priorities to the parts of the article that should be annotated.

Another issue was the part of QKDB records coming from tables: tables are a kind of structured information but, since the presentation of results in tables varies, they are difficult to process automatically. 60% of QKDB records come from tables. This proportion being quite high, special processes will have to be conceived for extracting results in tables.

3.2.2 Silence

A reason for silence is term variations. Derivational forms are found: the region *urine* can be mentioned with the term *urinary*.

Many abbreviations were also found: *mOsm* is generally used for *milliosmole*, but only one form is present in QKDB.

Even for numerical values, small variations can be observed: *0.1* can also be entered in the database as *0.10*, or as *10-1*. Finally, solute names can be written in many different ways: *water* can also be written *H20*, *H₂O*, *H 2 0*...

All these variations have to be considered to be able to retrieve QKDB values into texts. Another reason for silence is that some QKDB records are erroneous; eventually, some numerical values are the average of two values of the text, in which case it is difficult to find the source of the QKDB value in the text.

		correct	ambig.	silence	# meas.
Region	S	9	20	72	46
	P	14	27	59	
Tube segment	S	0	24	76	51
	P	20	41	39	
Solute	S	28	0	72	57
	P	14	14	72	
Parameter	S	42	19	39	95
	P	40	27	33	
Species	S	79	0	21	95
	P	67	13	20	

Table 2: *Ambiguity and Silence (%)*

3.2.3 Ambiguity

We also wanted to study ambiguities: for each measure, are there several different instances of a slot in the same sentence or paragraph (which would make the detection of the right instance harder)? See for instance the following sentence in which there are three values for the field tube segment: In controls versus PAN rats, Na+/K+ATPase activities were (pmol ATP/mm/h): *proximal convoluted tubule*, 2954±369 vs 2769±230; *thick ascending limb*, 5352±711 vs 5239±803; and *cortical collecting duct*, 363±96 vs 848±194 ($P<0.01$), respectively. These values will have to be desambiguated to be associated to their own result. In this case, either punctuation or proximity to the result would constitute good criteria.

Table 2 presents results for some slots with their degrees of correctness, ambiguity and silence, in the context of a sentence (S) and of a paragraph (P). The last column precises the number of records in the reference corpus for each slot. We can see that even in a single sentence, ambiguity remains high enough.

3.2.4 Position of the slot instances

One of the goals was to study which slot instances could generally be found close to the result value, and when they were remote, where they could be found. For the parameter slot, for 65% of the measures, the parameter name is in the same sentence as the value, which makes it easier to detect. For the solute slot, its name appears in the same sentence as the value in 90% of the cases. On the contrary, information given in the comment slot, which gives details about

the experiment that could not be inserted into other slots, such as additional information about the studied species, is in the same sentence as the value for only 35% of the records.

As the domain is renal physiology, the organ slot value is quite always *kidney*, but some articles refer to experiments for other organs, which are always given right before the result.

3.2.5 Noise

Finally, all QKDB records were considered, and projected in each article, in order to evaluate the quantity of noise in each article, i.e. the number of QKDB values that were in an article, but were not related to an experimental result (and thus not entered in a QKDB record associated to this article). On 5 articles, only one case of noise was found: an article refers to a *pH* measure that is not linked to a result. That means that most QKDB values that will be detected in the texts should be connected to an experimental result.

3.3 Synthesis by slot

This projection enabled us to determine the most frequent kinds of expressions of QKDB values in the texts. These observations were later used to develop an automatic annotation tool.

Numerical slots (result value, precision of the value, and number of animals) required writting rules. The result value is a number, which can contain exponents, in which case the exponents are after the possible precision ($9.37 \pm 0.77\ e^{-4}$).

The precision is an integer or decimal number always preceded by $\pm$.

The number of animals can be preceded by *n=*, or followed by a name of species.

The unit is composed of one or several base units (such as *g*, *m*, *mol*), which can be preceded by prefixes from the International System of Units (such as *m*, *d*, *h*, μ). The units are joined with dots, slashs or spaces, and can be followed by the exponent *-1*. They can immediately follow the result, or be given with the parameter studied as in *Apical membrane Pf averaged (in cm/s)* $9.37 \pm 0.77\ e^{-4}$

Parameters have predefined values in the database, which should be completed when new articles will be processed. Their proximity to the result value will be a criterion to help to detect them.

Species also have predefined values in QKDB, and their list can be easily enlarged with lexicons for example. In 90% of the articles, only one species is mentioned in the article, which helps detecting the species of an experiment, and its occurrence number is rather high.

The comment slot contains additional information about the species, about their treatments... The result value and the comments associated are often in a different part of the article, since the comments usually are in the Methods section. The values of this slot vary, so their automatic detection may be harder than for the other slots.

For the organ slot, when it is not *kidney*, the organ is specified before the result.

For other slots (such as tube segment or cell type), the terms entered in the database are not necessarily those of the texts: in articles, the terms can be more specific or on the contrary more generic than in QKDB. For example, *proximal tubule segments* becomes *Proximal Straight Tubule* in the database.

4 Extraction phase

4.1 Rules

In a first step, result values have to be detected in the texts.

A numerical value was considered as a result value if it is in the Results or Discussion sections, and is either followed by a $\pm$ character and another numerical value, or followed by a unit, or in a table.

Then, each result value has to be linked to the terms describing it. To do this, we have to explore the contexts of the results.

For each slot describing an experiment, a strategy was developed. Patterns expressing the context of a value were written. A result value can for example be searched with the following pattern: [numerical_value] $\pm$ [numerical_value] [unit] meaning that the result will be followed its precision and unit.

To detect the units, we look for base units (such as *m*, *g* or *mol*), with potential prefixes (such as *k*, *n* or *d*) and postfixes (such as *-1*), and separated by dots, slashs, or spaces.

The number of animal studied in the experiment can usually be found after the phrase *n=*.

These are the slots with no predefined values in the database. For other QKDB slots, we look for instances of them in the sentence containing the result value, or the previous one, except for the species, which can be found with a simple strategy: the most frequent species (or even word) can be selected, unless another species was mentioned right before.

The objective was to build an evaluation framework, with a first extraction system that will constitute a baseline, before introducing variations and more complex selection strategies.

4.2 Results

The study was made on the rest of the corpus, thus 15 articles containing around 840 measures. First, the result values were annotated, then the other slot instances were selected either in the same sentence as the result value, or in the same sentence and the previous one, or in the same paragraph. For tables, these 3 contexts are identical. Precision and recall values are shown in Table 3. The precision corresponds to the number of slots correctly annotated divided by the number of slots annotated. The recall corresponds to the number of slots correctly annotated divided by the number of slots that should be annotated (those in QKDB). When results that have not been inserted in QKDB are annotated, their describing slots are counted as erroneous. Thus, in order to give a more precise idea of the extraction results for these slots, the table also shows precision values only calculated for slots linked to a result value of QKDB.

Context	All descriptors		Result values		QKDB results
	R	P	R	P	P
Sentence	0.45	0.33	1	0.64	0.52
2 sentences	0.45	0.32	1	0.65	0.51
Paragraph	0.46	0.22	1	0.53	0.49

Table 3: *Precision and recall for the extraction process*

Recall of result values is 1: all the results are annotated; yet precision is around 0.5 so twice the right number of results are annotated. Some of the erroneous templates do not refer to an experiment result, while other correspond to results of experimentations even if they have not been inserted in QKDB: either the user has simply omitted it or he has judged it uninteresting, because it was known by the community.

We have previously said that some descriptors were extracted by rules. In the paragraph context of the reference corpus, 43% of attributes have to be extracted by rules. The extraction system extracts 70% of them. For descriptors whose extraction is based on QKDB lists (57%), 28% of them are extracted.

4.3 Discussion

Some results are wrongly annotated, as in the following sentence:

In these studies, apical membrane vesicles were enriched 10.5 ± 0.5-fold for the luminal marker-glutamyltransferase

The pattern [numerical_value] ± [numerical_value] is recognized, but here it is not the value of a result.

Other fields are not annotated, mostly due to variations of terms. Several types of variations were detected in the projection phase: inflections, derivations, acronyms, typographic variations. Lists are being constructed to detect them in the texts, and link them to QKDB values, mostly automatically, for example with WordNet (for some inflections and derivations). The different variations of a term will thus be normalized to a standard form.

Besides recognizing the terms of the domain, we will have to work on the selection of relevant results, so that an annotator who will use our assistant tool will not have too erroneous propositions to discard. We will have to define with experts where to draw the line between recall over precision.

5 Relevant work

Template based IE systems were developped during the MUC conferences (see [1] for MUC-7 definition task and [7] for MUC-7 results). In MUC7, the Template Relation Task was dedicated to extract relational information on employee_of, manufacture_of, and location_of relations as the Scenario Template Task consisted in extracting prespecified event information and relating the event information to particular organizations, persons, or artifact entities involved in the event.

Some of these systems as LaSIE [6] have been adapted to extract biological information, designing PASTA [3]. PASTA aim at extracting information about the roles of residues in protein molecules. The extraction task consists of filling a template defined by three template elements and two template relations from MEDLINE abstracts. It makes use of syntactic and semantic processing based on a domain model that consists of a concept hierarchy (an ontology).

The BioRAT system [2] was designed to extract information from full-length papers, when they are available and can be converted from PDF to TEXT format. The kind of information extracted is designed by patterns represented by regular expressions that link words related to protein expression and interaction found in gazetteers and protein names. The extracted information is located inside a sentence.

Pharmspresso [5] is also a tool for extracting information from full texts. Like BioRAT, it searches for relations between categories of biological entities represented by patterns that can be found in a sentence.

6 Conclusion

This paper presents a corpus study on scientific articles in renal physiology. The goal of the project is to automatically annotate experimental results in these articles to populate a database. These results can be represented as a template, with slots for the description of the measure and the experiment fields (unit of measurement, species, organ...).

In a first step, a tool was constructed to project QKDB records towards articles, in order to annotate a corpus of reference and to study the repartition and expression of the results in the articles.

Then, a baseline information extraction tool was created, which will now have to be completed to take into accounts term variations, complex relationship expressions, and qualitative information (such as the qualitative data and comment slots).

References

[1] N. Chinchor. Overview of MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[2] D. Corney, B. Buxton, W. Langdon, and D. Jones. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206, 2004.

[3] G. Demetriou and R. Gaizauskas. Utilizing text mining results: The PastaWeb system. In *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, pages 77–84, 2002.

[4] R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willett. Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, 19(1):135–143, 2003.

[5] Y. Garten and R. Altman. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC bioinformatics*, 10(Suppl 2):S6, 2009.

[6] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conferences (MUC-7)*. Citeseer, 1998.

[7] E. Marsh and D. Perzanowski. MUC-7 evaluation of IE technology: Overview of results. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[8] P. Shah, C. Perez-Iratxeta, P. Bork, and M. Andrade. Information extraction from full text scientific articles: Where are the keywords? *BMC bioinformatics*, 4(1):20, 2003.