Assessment of Biomedical Knowledge According to Confidence Criteria

Ines JILANI[a], Natalia GRABAR[a], Pierre MENETON[a] and Marie-Christine JAULENT[a]
[a] *INSERM, UMR S 872, Éq. 20, Les Cordeliers, Paris, F-75006 France ; Université Pierre et Marie Curie-Paris6, UMR S 872, Paris, F-75006 France ; Université Paris Descartes, UMR S 872, Paris, F-75006 France*

Abstract. The characterisation of biomedical knowledge taking into account the degree of confidence expressed in texts, is an important issue in the biomedical domain. The authors of scientific texts use grammatical and lexical devices to qualify their assertions. We named these markers of qualification "confidence markers". We present here the results of our efforts to collect confidence markers from full texts and abstracts, to classify them on the basis of semantics, and their use within our knowledge extraction system. We propose in this study, an implementation of these confidence markers for functional annotation of the human gene *Apolipoprotein (APOE)* thought to be involved in Alzheimer's disease. As a result, we obtain, through the extraction system, triplets: (G, F, PMID), in which G is the gene *APOE*, F is its function found in texts and the PMID of the article from which this knowledge was extracted. Moreover, a spatial 3D of the triplets, relative to each other, is proposed depending on their respective confidence degree.

Keywords. Data acquisition-data capture, Data analysis-extraction tools, Modeling

Introduction

Bibliographical databases, such as Pubmed[1], are indexing increasingly large numbers of biomedical articles. They are essentially consulted by the medical, biological and bioinformatics communities, and the number of searches last year (2006-2007) rose to over 82.3 million, with 423 million page views[2].

This study follows on from the automatic extraction of knowledge about genes from scientific articles indexed in Pubmed, based on a natural language processing method [1]. Loss of context limits the extraction of information about genes, such as their functions, the diseases associated with them and their interactions. For instance, if knowledge arises from an experimental procedure or constitutes a reference to another article, then the context allowing the recipient of the extracted knowledge to associate it with certain reliability and confidence is missing. There is an important issue to formalise this kind of additional information, to weight the extracted knowledge and use it more confidently.

Information on the validity of knowledge is often given by a specific linguistic device, sometimes called "hedge", "modifier" or "qualifier". These linguistic

[1] http://www.ncbi.nlm.nih.gov/sites/entrez

[22] NLM Technical Bulletin - June 18, 2007 MLA 2007: NLM Online Users' Meeting Remarks http://www.nlm.nih.gov/pubs/techbull/mj07/mj07_mla_dg.html (last visited on Oct, 19th 2007)

phenomena can be referred to as “confidence markers”. They belong to different grammatical categories - verbs, adverbs or adjectives - and qualify the author’s assertions in the article. Consider, for instance, the following sentences “Copper deficiency is a plausible cause of Alzheimer disease (AD). This hypothesis should be tested with a lengthy trial of copper supplementation” (from the abstract of the article with Pubmed Identifier 17928161). The terms underlined are markers, indicating qualifications used by the author for tacit weighting of the reliability of his claim. The word “hedge” was first used in this area in 1972 by Lakoff [2], who described hedges as “words whose job it is to make things more or less fuzzy”. Hyland [3], Light [4], Mercer [5] later carried out qualitative studies of these qualifiers. However, they have neither modelled them, nor integrated their use for weighting any kind of information in a knowledge extraction system.

In this study, we worked on the use of these confidence markers in scientific articles, their significance, their classification and their automatic detection in texts for knowledge weighting purposes. The main aim was to document the information so that it could be used confidently.

We first present the materials and methods used and implemented for our study, then the results obtained followed by their discussion and finally, a conclusion to place our results in their context and to describe the avenues for a further exploration.

1. Materials and Methods

Corpora. We used three corpora obtained by querying Pubmed [6]. CORP1 was obtained with a list of 160 human candidate genes thought to be related to AD [7]. This corpus is a collection of 355 abstracts, containing 817 sentences, 213,618 words and it is 1 MByte in size. CORP2 is composed of 68 full texts related to the abstract references in CORP1, available from Pubmed Central[3]. It contains 27,912 sentences, 1,123,873 words and is 2.4 Mbytes in size. CORP3 is a collection of 348 abstracts collected from Pubmed with a query containing a list of 160 nematode genes identified by biologists as potentially linked to AD. CORP3 contains 825 sentences, 201,753 words and is 1 MByte in size.

Lexical resource. WordNet® [8] is a large lexical database of English: nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept. This database was used in our study to enrich the extracted confidence markers, by identifying their synonyms.

Our method consists in collecting the largest possible number of confidence markers (from CORP1, CORP2, CORP3), enriching them using WordNet®, classifying confidence markers and defining other confidence criteria. This enabled us to establish a confidence marker model, which was applied to our system for extracting knowledge concerning genes.

Collection of confidence markers. We collected confidence markers manually, retaining every linguistic device suggesting that the authors were placing a qualification on their claims, the presence of the author in the text, hints in the text relating to the context of the information presented, and so on.

Enrichment of the list of confidence markers. The confidence markers collected provide a partial view of possible linguistic devices used by authors in scanned

[3] http://www.pubmedcentral.nih.gov/

documents. However, it is useful to extend these devices to morphosyntactic variations and their synonyms. We used WordNet® resources for synonymic extension.

Classification of confidence markers and definition of other confidence criteria. The markers must be classified into homogeneous subcategories for use within an automatic extraction tool and for knowledge weighting purpose. These subcategories are further graduated according to the confidence or discredit they add to the extracted knowledge. Other criteria should also be taken into account to weight the extracted information: the impact factor (IF) is one of these. It is a measure of the frequency with which the "average article" in a journal is cited over a given period of time. The IF for a journal is calculated over a three-year period, and can be considered to be the average number of times published papers are cited up to two years after publication.

Modeling confidence criteria. The last step of the method involved establishing a model of the final set of markers. This confidence marker model must be useable within an automatic tool for extracting functional knowledge from texts, together with related information about confidence. This method is based on regular expressions (or graphs) constructed manually, as described elsewhere [1].

Finally, we applied this model to the context of AD and, more precisely, to searches for knowledge relaying to the apolipoprotein E gene *(APOE)*, which is thought to be linked to AD in humans. We built a specific corpus (CORP_APOE) for the application of the model. This corpus was obtained by querying Pubmed with *APOE* gene and its synonyms. We considered all biological functions of *APOE*, extracted them with the knowledge extracting system to get triplets (*G*, *F*, *PMID*), where *G* is the gene *APOE*, *F* is its function and *PMID* is the Pubmed identifier of the article from which the knowledge was obtained. Afterwards, we detected all confidence markers potentially qualifying the knowledge extracted, weighted and placed the triplets in a 3D space according to the criteria described above.

2. Results and Discussion

In this section, we detail the results obtained for each step of the proposed method.

Collection of confidence markers. A list of 250 manually collected confidence markers was generated.

Enrichment of the list of confidence markers. Once collected, we extended the markers, using their morphological variants. We do that by searching lemmas of the confidence markers. For instance, we extended the identified confidence marker “previous study” to “previous studies” (adding plural or singular forms), and “remain unknown” to “remains unknown” (inflectional form of verb). This process increased the number of confidence markers listed to 478.

We also carried out a second type of extension in which synonyms of the extracted confidence markers were taken into account. For instance, we extended the confidence marker “we anticipate” to “we expect” (*expect* being a synonym of *anticipate* in WordNet®), “previous study” to “previous work” and “previous report” (*work* and *report* are synonyms of *study*). There is some debate concerning the use, sufficiency, accuracy and linguistic relevance of WordNet® [9] but it was considered appropriate for the context in which we were working. Indeed, the process of extension based on synonyms used here did not modify outcomes; it simply increased the size of the set of

confidence markers to 700. Notice that ignoring this step of acquiring variants and synonyms of confidence markers would have resulted in a loss of weighting information, as our list of markers would not have been sufficiently exhaustive.

Classification of confidence markers and definition of other confidence criteria. The necessity to classify the confidence markers according to their semantics has emerged in order to give different weights to the knowledge they characterise. We highlighted four different categories, described below in ascending order of confidence:

1. **Interrogation or trial and error of the author.** Knowledge that remains unproven and requires demonstration. This knowledge may also correspond to the author's interpretation in the absence of conclusive proof within the article.

e.g.: *"remain to be confirmed", "has yet to be identified", "?"*

2. **Distance suggested by the author compared to his assertions or the knowledge presented in the text.** This fine distinction was suggested by Hyland [3]. It may also correspond to a restriction of the knowledge concerned to a specific context (e.g.: the context of the article or experiment).

e.g.: *"our findings suggest that", "in this case we conclude that", "it is possible that"*

3. **Studies by other researchers, references to other works, articles or methods.** These elements are associated with a high level of confidence in the knowledge communicated, as we assume that if an article is cited, the information is assumed, or at worst simply believed to be true. This information therefore does not require further demonstration in the article concerned.

e.g.: *"previous observation", "it is now believed that", "it has been proposed that"*

4. **Demonstration or proof given by the author.** This corresponds to work carried out by the author and presented in the article concerned. It involves the use of markers implying that the knowledge is highly reliable and based on demonstrations presented within the article. These markers are associated with the highest level of confidence in knowledge.

e.g.: *"we reveal that", "we show here that", "our results indicate that"*

These four categories correspond to *Type 1* confidence markers.

We also distinguished *qualifiers*, modifying confidence levels within the four categories described above. These qualifiers characterise subjectivity in texts, describing things or events that are possible but not observed or not certain. These qualifiers are *Type 2* confidence markers. They are represented in Fig. 1, from negation to affirmation, i.e. from the least probable to the most probable[4].

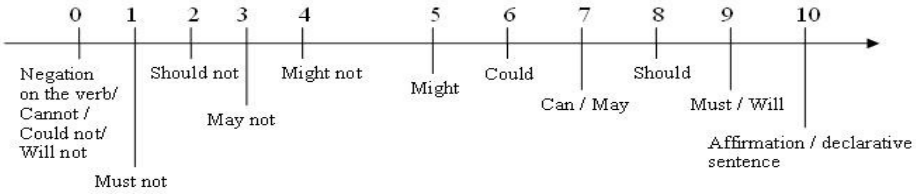

Figure 1. Graduation of *Type 2* confidence markers

Negation indicates that the author presents a negative assertion, with the author considering the knowledge negated to be false. This knowledge is therefore of the lowest confidence (*"unable", "not", "neither…nor"…). Must not, should not, may not*

[4] Ian Jacobs. 1995. English Modal Verbs, available at www.w3.org/People/Jacobs/modals.ps

and *might not* are closer to negation than the other markers to the right of the arrow (*might, could,…*) and convey more confidence than a clear negation. *Might, could, can/ may* and *should* express increasing probabilities. Finally, *must/will* and affirmative sentences expresses the highest positive degree of certainty.

Through the use of IF criterion, we hypothesised that the IF of a journal is related to the reliability of the biological and medical information published. This does not imply that journals with a low IF publish erroneous biological experiments, but simply that biologists are used to give more confidence weight to information derived from journals with a high IF. The publication of an article in a journal can be readily found using the ISSN (international standard serial number) of the journal, available in the XML format of the result of a Pubmed search, in the tag *<ISSN IssnType="Print">**0003-2697**</ISSN>* (the number in bold corresponds to the ISSN of the journal). A listing of the IF of journals can be used to weight the knowledge extracted from these journals.

Modelling confidence criteria for their automatic extraction. We used regular expressions (or graphs) [1] to retrieve information from texts and therefore modelled all the confidence markers in these terms. First, lemmas of each word in confidence phrases were defined, making it possible to detect all forms of a word, rather than just a single form. Second, synonyms of the markers, defined in the enrichment step, were collected into the same expression. For instance, confidence markers such as "we anticipate" and "we expect" form the expression *we<have>*(<anticipate>+<expect>)*, where words between < > correspond to the canonical form of the word, + corresponds to logic *OR* and * means *0 or n* occurrences. When isolated markers were found to be synonyms within WordNet®, we included them in the same expression. This was the case for "we hypothesise" and "we suspect", and, with extension for synonym, this created the following regular expression:

we<have>(<hypothesise>+ <speculate>+ <expect>+<predict>+<suspect>)*.

Nouns, adjectives, adverbs, as well as verbs, can be extended with WordNet® resource, as in the following expression:

<have><be>(previously+now)*(largely+widely+extensively+generally)*<confirm>* which picks up the sentences *"have been previously confirmed"* as well as *"is now largely confirmed"* or *"is widely confirmed"*. Through these examples, we can observe the capacity of regular expressions to identify the targeted linguistic events.

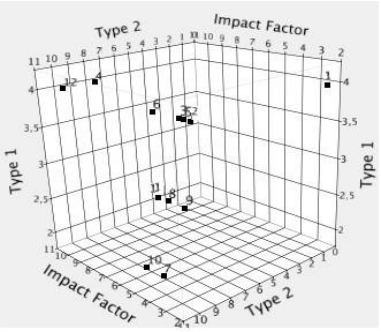

Figure 2. Graphical visualisation of the extracted triplets depending on their degree of confidence

Application in the context of apolipoprotein E gene. The application of such extended and precise regular expressions to CORP_APOE led to the extraction of 650

knowledge elements. These elements were extracted in the form of *(G, F, PMID)* triplets, where *G* is the gene, *F* its function and PMID the Pubmed identifier of the article from which the functional annotation was obtained.
This information can be visualised, as in fig. 2. The knowledge extracted is positioned in 3D space according to the *Type 1* confidence level of the biological knowledge, the *Type 2* confidence level and to the IF of the source journal.

3. Conclusion and Perspectives

This paper describes our experience with the attempt to integrate confidence markers into an automatic knowledge extraction system [1] and the application of these markers to documents in the biological and medical domains. The markers were collected manually from corpora of abstracts and full scientific texts extracted from Pubmed. They were extended with the WordNet® resource and classified into four categories of *Type 1* or ten categories of *Type2* depending on the semantics and strength with which they modify the knowledge reliability.

The extraction and characterisation of knowledge concerning the biological function of genes and their products, based on a system of this type, improves our understanding of the reliability of this information, particularly if the extracted knowledge is projected into a 3D space. We are currently working with biological experts to evaluate the efficiency of such a presentation.

We plan to work on the automatic detection of new confidence markers, to make the available list even more exhaustive. Such detection may be based on the syntactic category of elements within sentences and on their expected semantic role. Notice that the diverse writing styles of the authors are not taken into account in this study, how readers from different cultures intend markers introduced by writers in texts. We have also observed that these markers may be clustered together within a sentence, and we will try to exploit this observation for their automatic detection. In addition, we plan to take into account the type of the study presented in the article, i.e. if it is an observational study (epidemiological), a controlled experiment, or a clinical essay.

References

[1] Jilani I, Grabar N, Jaulent M-C. Fitting the finite-state automata platform for mining gene functions from biological scientific literature. 2006; In Semantic Mining in Biomedicine, Jena (Germany).
[2] Lakoff G. Hedges: A study of Meaning Criteria and the Logic of Fuzzy Concepts. 1972; Chicago Linguistic Society, 8, pp. 183-228.
[3] Hyland. K. The Author in the Text: Hedging Scientific Writing. 1995; Hong Kong Papers in LLT.
[4] Light M, Qiu X Y, Srinivasan P. The Language of Bioscience: Facts, Speculations, and Statements in Between. 2004; In BioLINK: Linking Biological Literature, Ontologies, and Databases.17–24.
[5] Mercer R E, Di Marco C. A design methodology for a biomedical literature indexing tool using the rhetoric of science. 2004; In BioLINK: Linking Biological Literature, Ontologies, and Databases, HLT-NAACL: Association for Computational Linguistics.
[6] Grabar N, Jaulent MC, Chambaz A, Lefebvre C, Neri C. Sifting abstracts from Medline and evaluating their relevance to molecular biology. Stud Health Technol Inform 2006;124:111-6.
[7] Lefebvre C, Aude JC, Glemet E, Neri C. Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates. Bioinformatics 2005;21(8):1550-8.
[8] WordNet, An Electronic Lexical Database, C. Fellbaum ed., 1998; The MIT Press, Cambridge, Mass.
[9] Slodzian M. WordNet: what about its linguistic relevancy? 2000; In EKAW2002.Juan-les-Pins (France).