Biological Information Extraction of Scientific Articles

Joana Gomes

Instituto Superior Técnico

Abstract During the years, many scientific documents have been submitted in the area of Biology in order to understand and anticipate the effects of global warming in drastically reducing the biodiversity of Earth. Besides this, there is a huge dispersion of knowledge and it becomes difficult to deeply study each species as the information is usually spread over different articles. With the evolution of technologies, text mining techniques have been developed and used in order to extract automatically relevant data from texts, images and charts. In this work the main objective is to extract information on birds that are present in scientific articles trying to answer the question: "Is it possible to build a system that may extract automatically data regarding specific bird species from scientific articles?" To develop our solution, we created a system that analyses text through a combination of techniques of natural language processing, regular expressions and machine learning algorithms. The system receives, as input, the set of documents to analyze and as a result it presents the possible values to the characteristics of the species that we want to analyze (body temperature, body mass, among others). As main conclusion to this work, we demonstrated that it is possible to build a system that extracts data from scientific documents in the Biology domain. However, it is not yet possible to have a fully automatic process, being relevant to have a human user that may solve the result ambiguity.

Keywords: Knowledge Bases, Slot Filling, Text Mining, Information Extraction, Machine Learning, Natural Language Processing, Regular Expressions

1 Introduction

On Earth there are different forms of life: humans, animals, plants, microorganisms and the multitude of genes that make up and distinguish them. All these beings make up the biodiversity of the planet and, over the centuries, the existence of some species has been threatened due to various factors, including, environmental (pollution, fires) and hunting. The extinction of several species has increased the concern about nature and its biological diversity, becoming evident the need to develop models to understand and anticipate the effects of global change on biodiversity [24].

The scientific community has conducted many field studies and many articles have been published on the various characteristics of biological beings on the planet. Being Biology the science responsible for this field of study, it splits into several branches (for example, zoology, microbiology, genetics, physiology, biotechnology, etc.) what contributes to an enormous dispersion of information relating to a particular species [7].

Given the existence of several scientific articles in various branches of biology, it has become difficult and increasingly long to study at the level of taxonomic group and also to gather and aggregate information of scientific review articles or databases, as it would be necessary for experts to read a lot of papers. To meet the biology information needs, there have been set up systems for

information extraction and databases have been created. On the Internet we can find several databases dedicated to various branches of biology. However, taking as example the taxonomic group of birds, most databases aggregates information of their physical form (mass, length), being rare the data of their physiology (body average temperature or water content).

For the physiological data of the taxonomic group of birds, we propose to address the problem that can be formulated in the following question: "Is it possible to build a system that can extract data from certain species of birds from scientific articles?". In response to the question asked, we define as the main objective of this work to build a system to extract information about birds of physiological data from specialized scientific articles.

2 Concepts

Knowledge bases are designed to facilitate access to information inherent to knowledge and, being knowledge a complex concept that aggregates various information, existing systems used to create these structures are based on processes that handle large amounts of data from various sources (texts, tables, images or diagrams), sizes and types [15].

In order to obtain the relevant data for the knowledge bases it becomes imperative to refer to the concept of text mining is the process for obtaining information from unstructured data, implying identification techniques, extraction, management, integration and interpretation of the data automatically [4]. A Text Mining system begins by gathering all relevant texts to its purpose (with the application of information retrieval techniques [6]), identifies and extracts the desired information (with the application of techniques and algorithms for information extraction [18]) and finally, the data is interpreted and associations are found between them so they can be used to assist in decision making [4] (with the application of data mining techniques [12]).

2.1 Information Extraction

The area of Information Extraction appears in Natural Language Processing community, specifically in two competitions of great importance, Message Understanding Conference[1] and Automatic Content Extraction[2] , in order to automatically extract structured information from unstructured documents [19].

In an information extraction system, the documents for which you want to extract information correspond to the input data. These documents must be in digital format so that the information and data they contain may be analyzed. To do this, you must use Character Optical Recognition technologies to perform the conversion [8]. To ensure the system knows what to extract it is also necessary to provide a model with the entities and fields to be extracted. Optionally one can use knowledge bases, dictionaries, glossaries and ontologies to identify entities. Upon completion of the information extraction process, data is obtained in a structured format that can be later analyzed [23].

Each Information Extraction system is built to answer questions from different domains. However, although they may have some significant differences, there are certain common components to all systems, namely, metadata analysis, tokenization, morphological analysis, detection of phrases, named entities detection, recognition and parsing sentences.

[1] http://www-nlpir.nist.gov/related_projects/muc/index.html
[2] http://www.itl.nist.gov/iad/mig/tests/ace/

Regarding the language components that are dependent of the domain, these may vary depending on the application requirements. In this group we find the four main tasks of the Information Extraction in particular Name Entity Recognition, Coreference Resolution, Relation Extraction and Extraction.

2.1.1 Information Extraction Methodologies

There are several approaches to develop an Information Extraction system. According to Krallinger, Erhardt and Valencia [9], to obtain data on biology branch it is common to follow one out of four approaches: dictionary based, rules based, Machine Learning algorithms or hybrid approaches.

- **Dictionary Based Approach:** Using a list of names for the domain in question, one carries out a search of these names in free text obtaining as a result the names that are present in both the list and the documents. The main disadvantage of this method is the need to have an exhaustive list of the names and terms to look for, including misspellings, name variants, abbreviations, synonyms and ambiguous names. In the case of the Biology area it is difficult to maintain exhaustive lists with all the necessary information due to, among other factors, the constant updating and discovery of new terms [23].

- **Rules Based Approach** Rules are defined (e.g. regular expressions [11]) consisting of a pattern and an action to develop. The rules are applied to the free text and when a given pattern is identified, the action is performed [15]. Rules can be prepared according to one of two methods: by hand coding or using Machine Learning algorithms. At first it is necessary the existence of human experts in the knowledge domain to define the rules, unlike the latter, in that from existing structured examples, the system learns the extraction rules [15].In this method the main drawback is the rules created according to the manual coding method, as they are so specific for the domain that they are rarely applicable to other domains. On the other hand it makes the construction of the system take very long and can delete important terms that do not correspond exactly to the predefined standards [23].

- **Machine Learning Based Approach:** Automatic Learning aims to develop algorithms to automatically detect patterns in data [10]. We can divide the Machine Learning algorithms depending on type of learning, namely supervised learning, semi-supervised or unsupervised [23]. Supervised learning is normally performed by induction of a model able to predict future events based on a large set of training data [15]. Algorithms that follow a supervised learning are difficult to apply in the field of biology due to the complex task of compiling a large set of training data. In order to overcome the high cost of the preparation examples, arises semi-supervised learning which differs from the previous in that it requires a much lower quantity of examples. In unsupervised learning there is no distinction between training data and test data, with all incoming data being processed in order to create a sort of summary or agglomeration [10]. Usually they are clustering or dimensionality reduction algorithms [23].

- **Hybrid Based Approach** Each domain in which Information Extraction techniques are applied have different specificities which leads to the need to adapt the solution. Thus, sometimes a hybrid approach is used to combine the advantages of the different approaches outlined above.

2.2 Feature Selection

In classification problems, the presence of a large number of features used to characterize the data that is processed by the classifiers, may result in situations of overfit, having consequences for its performance. In order to solve this problem some dimensionality reduction techniques have been created [1].

The Feature Selection technique aims to choose a subset of the most relevant features from the original set according to certain criteria. The Feature Selection mainly affects the training phase of the classifier. First the features are selected and then the data is processed.

2.3 Knowledge Base Population

The question of knowledge bases population came in 2009 with the aim of promoting research in automated systems that discover information about entities from a large volume of documents and, in a post phase, use this data to develop knowledge bases [20]. There are several groups participating in this competition that includes several subtasks. In this case, it is interesting to highlight the Filling task step.

In Slot Filling task it is necessary to create an information extraction system to obtain certain attributes of individuals and organizations from a large number of documents. Participants in this task have access to the document where you need to extract information (usually data from web pages and discussion forums) and a model with a list of itemized fields that must be filled. Therefore, the ultimate goal is the population of several Knowledge Base fields using the approach that they think may best suit the problem [20].

2.3 Evaluation Metrics

To carry out the performance evaluation of information extraction systems, four metrics are used: Accuracy (evaluates the percentage of all correct forecasts), Precision (measures the proportion of extracted fields that were returned correctly compared with all extracted values, whether they be correct or incorrect), Recall (measures the proportion of correct fields that were correctly returned taking into account the reference fields forecasted to be filled in) and F-Measure (It combines the two previous metrics, precision and recall. To have an acceptable F-Measure it is necessary to have a balance between the metrics).

3 Related Work

In order to obtain information from scientific articles we are faced with some specific characteristics of Biology area that hinder the extraction of information, namely:

- **Specialized language:** The language used in biology articles is constantly changing due to changes in our understanding of the branch. New terms are created and others are removed, sometimes quite ambiguous [13].

- **Syntax difference:** There is no standard syntax descriptions of the different taxonomic groups, or even within the same taxonomic group. There are descriptions written in English, other languages and there are several names and abbreviations for the same entity [23].

The systems can be implemented following different approaches: dictionary, rules and Machine Learning. However it is noted that systems are not very common to use only one of the approaches and therefore there are systems which follow a hybrid approach which combine more than one of these approaches.

Systems using the studied dictionary revealed to have as primary goal the extraction and recognition of names of entities, which in the case of biology, are taxonomic names. In the area of biology, taxonomic names are frequently updated, which converges to a possibility of existing ambiguity in biological words. Thus, the TaxonFinder[3] system, using a dictionary approach, identifies scientific names in a text by comparing several lists of data. The various lists were built manually by domain experts, and each list contained, respectively, species names, genera names, family names and words from the common lexicon [23].

For approaches according to the rules, these are usually used to improve the accuracy of information extraction in systems using other approaches. In the literature you can find the Protein Active System Site Template Acquisition (PASTA) was implemented in order to extract information about the functions of the amino acids in protein molecules, using as source scientific journals and complete articles. The final product would be the creation of a knowledge base of the active sites of proteins. To achieve this, it was used the analysis of the text through regular expressions and use of inference rules to fill in the required fields of the Knowledge Base [14].

Systems using only resources of Machine Learning are not many, yet NetiNeti system (Name Extraction from Textual Information-Name Extraction for Taxonomic Indexing) is based on Machine Learning in order to recognize and discover scientific names in text taking into account errors from Optical Character Recognition, misspellings and variations of names. Initially they conducted the tasks of linguistic analysis independent of the domain and then they got the candidates' names for use in classification, using the classifier of automatic supervised learning, the Naïve Bayes and Maximum Entropy [2].

The hybrid approach is the most chosen technique to implement the Information Extraction systems due to the possibility of combining the advantages of several approaches. TaxonGrab is an example of a system designed to identify taxonomic names using a combination of a list of terms in English, non taxonomic (the dictionary), and rules of binomial nomenclature of Linnaeus. Believing that most taxonomic names are not used in common language, the approach means that if a given word does not exist in the dictionary it is because it can be a taxonomic name. To confirm that the extracted terms are taxonomic names, these are compared to rules created from the binomial nomenclature of Linnaeus. This system is not very accurate, because it does not include misspellings problems and the words in a language other than English, however, it has the advantage of not requiring a complete list of taxonomic names [16].

Systems for Knowledge Base Population

Since 2009, several groups of researchers are participating in the sub-task Slot Filling for the Knowledge Base population competition. According to Surdeanu and Ji [22], the best systems have 52% human intervention and the best strategy for Information Extraction was based on Distant Supervision for conducting the training of the system.

The system that showed the best results was proposed by Angeli et al. from Stanford University[5]. This system was based on a framework developed by the authors: DeepDive. This framework was

[3] http://taxonfinder.org

implemented to assist in building systems in order to make it easier to integrate the knowledge of a particular domain without the concern of the programming of the entire process.

Indeed, the system receives as input data unstructured data and then three major phases are performed in the information extraction process: 1) Feature extraction; 2) Probabilistic Engineering; and 3) Inference and learning. Finally, the model calculates automatically the probability of each candidate belonging to certain classes and produces the output database together with all different values of the calculated probabilities. Finally, it builds the Knowledge Base with the identified cases.

3 Implementation

Our system follows a hybrid approach complementing the approach according to the rules of machine learning approach. Thus, for the analysis of texts Natural Language Processing techniques, regular expressions and various classifiers are used.
In the figure 1 you can check the overall architecture of our system, where we present the input documents into the system, the system and the final product.

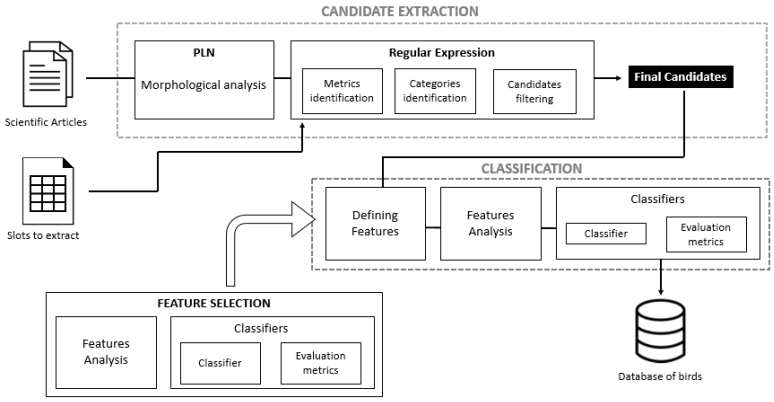

Figure 1: Overall Architecture of Information Extraction

3.1 Previous Knowledge and Input Documents

Before we build our system we had knowledge of what data is needed to get in the end to implement the knowledge base. Thus, through the analysis of scientific articles from Biology domain it was necessary to extract data from birds focusing on certain characteristics that we define as categories in this system. In Table 3.1 we present the categories for which the system is constructed and the words that we knew from the outset to be associated with each category. As we knew the desired categories referred to numerical values, we decided to go with the presented solution.

Table 3.1 Categories and related words

Category	Related Words
Body Mass	Wet weight; dry weight; wet mass; dry mass; at birth; hatching; hatchling, at fledging; fledging; adult; grams; kilograms
Body Temperature	Chick; adult; body; temperature; Celsius
Egg Temperature	Incubation; egg; temperature; Celsius
Fledging	Fledging; leaves the nest; days
Incubation	Incubation; hatching; days
Total Body Water	Total; body; water; content; percentage

Our system have three modules: the candidate extraction, classification and feature selection. Next we will explain each of the modules relating to its goals, development and products.

3.2 Candidate Extraction

Scientific papers are structured documents with fields in common with the text body sometimes presenting the text in multiple parallel columns. We used the tool named PdfMiner that allowed us to get the text that later we saved in documents with .txt format for future manipulation.

Next, in order to obtain the text separated by sentences we used the Natural Language Toolkit library (nltk) that allowed us to separate the sentences according to the punctuation. However, due to the columnar structure of the documents, there were some words that were separated by a hyphen. To correct this we analyzed all the sentences and eliminated this hyphen to force the word to be joined.

As we knew that the data to be extracted were just numbers, when we obtained all the sentences of the text, we then filtered out the phrases to consider only those containing numbers through the Part-of-Speech analysis inserted in Stanford CoreNLP tool. With it we scrolled through the sentences and selected only those containing numbers, either in the form of digits or in full.

In the following stages regular expressions were used to perform filtering to obtain candidates for each category. Thus, we began by looking for sentences that contained in its constitution certain measure which had been defined a posteriori according to the various categories. As we previously knew that certain words might be present in the sentence that would indicate that a given category exists, we proceeded to check the previous sentences validating that they contained such words for the category that we were analyzing. If confirmed, they would be saved as possible candidates to data to that category.

Having phrases already filtered by several categories, the only thing missing was to extract the values to be analyzed. Thus, with the use of regular expressions, we obtained the values present in sentences. These values could be presented as a singular value and order of magnitude or it could be a range of values where the order of magnitude could be in either end of the range or associated with each value. Obtained the values, it was necessary to transform the numbers that were presented in full towards digits and perform normalization.

3.3 Classification

The classification module had the objective of classify in binary form the candidates in each category being referred to or not of that category. To do this, first we define the characteristics to look at each candidate. The areas of the features are the following:

- **Accounting for the occurrence of words in sentences:** Using the scikit-learn tool, it identified all the words in the sentences of all examples, excluding words that are just sentence connectors and counted the number of occurrences in the sentences of the candidates;
- **Total words in the phrase:** counted the length of the sentence related to the number of words;
- **Distance between the value and specific words:** Knowing the words that may exist in the sentences of the candidates, we identified which words were present and counted the number of words that made the distance;

- **Range checking with parameterized numbers:** Knowing the figures for the average of the various categories, as well as the median, we checked that the value in question was close to the average value and the median value;

- **Distance from parameterized numbers:** We counted the numerical distance value in analysis compared to the average and median values.

After obtaining the analysis of the features of the various candidates we had to make up the group of training and test examples. As we did not have many positive examples of each category we decided to follow the approach Leave One Out in the way that it performs many iterations as there are candidates in test. That is, all candidates are seen as a test sample and the remaining are in the group of training examples. Then, we used various classifiers to verify those that obtained better results, namely in Naïve Bayes [2], Support Vector Machine (SVM) [10], SVM Linear [10], Logistic Regression [3], Random Forest [10], Decision Trees [10] and K Nearest Neighbors [10]. To make the evaluation of the results we used the valuation metrics that were used to analyze the systems of information extraction, namely Accuracy, Precision, Recall and F1-Score. Finally, we obtained a database of the various categories and the values that our system identified as possible values for the birds.

3.4 Feature Selection

When we defined the features, we did not know what would be its impact on the classification thus, to optimize the use of the various features and analyze which ones best suited and enabled better results, we performed the Feature Selection module following the Sequential Feature algorithm Selection. According to this algorithm the system moves through the various features and measures a defined evaluation metric and, depending on these results it goes creating a series of those features that get higher values. In our case, as we didn't have many positive examples for the various categories, sometimes the results were biased to one of the two classes, resulting in a very high accuracy values and we were not able to discover the target. Thus, we defined as an evaluation metric the F1-score. This way, the system ran all the features and calculated the value of F1-score for the sequence and finally the obtained the sequence which allowed the best value for the metric used.

4 Results

Analyzing all the data and comparing, we can see that the classification procedure without the selection features classifiers have worked best in Decision Tree and Random Forest classifiers, getting a maximum accuracy of 0.95652 and 0.97674 respectively, the first in the category Body Mass and the second in the category Body Temperature. The minimum value obtained was 0.83333 and 0.81666, respectively, in the category Egg Temperature.

As the classifier Decision Tree achieved good results for most categories, we decided to use it for Feature Selection. The results for category Body Mass were equal of the results without Feature Selection. However, in categories Body Temperature, Egg Temperature, Fledging, Incubation and Total Body Water the results were must high than the results with all features. Although of Body Mass values were not changed, we can concluded that several features weren't relevant and with less features it was possible to obtain the same result.

5 Conclusions

To meet the biology information needs and taking as an example the taxonomic birds group that most existing databases only aggregate information on its physiognomy (mass, length), being rare to find data of their physiology (average body temperature or water content), this paper comes up with the main objective of giving answer to the question: "Is it possible to build a system that can extract data from certain bird species from scientific articles? ".

In response to the question asked, we created a system that receives scientific papers and a document describing the features to extract and then proceeds to the extraction of relevant information to fill a Knowledge Base. The system follows a hybrid approach complementing the rules approach and the Machine Learning approach. To the analysis of the texts we used Natural Language Processing techniques, regular expressions and various classifiers. From these are three modules: the extraction of candidates, classification and feature selection.

We conducted experiments with various classifiers and data before and after implementing the feature selection algorithm. The classifier that achieved the best results was Decision Tree after the selection of features. The values subsequently obtained from the Feature Selection were not as superior as expected because of the existence of few positive examples of the various categories.

6 Work limitations and recommendations for future work

This study answered the central question previously mentioned and contributed to the extraction of multiple correct examples, however, there are some points that can be deepened.

The first point relates to the tool used for optical character identification. One of the major limitations of this study referred to the interpretation of the tool over the text of .pdf documents. Therefore, we suggest the use of a more powerful tool for differentiating text in multiple columns, images and graphics and that allows to extract the information in the most correct way, without incomprehensible codes.

In the classifier stage, we had just few examples for each category, which limited and skewed results of several classifiers. Therefore, we suggest getting more positive examples, maybe obtained using the method of Distant Supervision, as used in the resolutions of the participants in the competition for knowledge bases population. This way, we would trust more on the obtained results and then be able to use the feature selection taking into account the values of the metric Accuracy.
In future work, it would be interesting to create an interface for user to use the classification.

References

[1] Ananiadou, S., Kell, D.B., Tsujii, J.i.: Text mining and its potencial application in systems biology. Trends in Biotechnology 24(12), p571–579 (2006)

[2] Bentor, Y., Viswanathan, V., Mooney, R.: University of texas at austin kbp 2014 slot filling system: Bayesian logic programs for textual inference. In: Proceedings of the TAC-KBP 2014 Workshop (2014)

[3] Berry, M.W., Kogan, J.: Text Mining : applications and theory. Chichester, U.K. Wiley, Boston, MA, USA, 1 edn. (2010)

[4] Campbell, N.: Biology: Concepts & Connections. Pearson/Benjamin Cummings (2006)

[5] Craven, M., Kumlien, J.: Constructing biological knowledge bases by extracting information from text sources. In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology (1999)

[6] Department of Statistics: Chapter 12 Logistic Regression. September (2016). http://www.stat.cmu.edu/ cshalizi/uADA/12/lectures/ch12.pdf

[7] Goyvaerts, J., Levithan, S.: Regular Expressions Cookbook - Detailed Solutions in Eight Programming Languages, Second Edition. O'Reilly (2012)

[8] Henerey,R.: Classification. Chapter 2. pp.6–16. Machine Learning, Neural and Statistical Classification. Ellis Horwood. USA (1994)

[9] Hirschman, L., Morgan, A.A., Yeh, A.S.: Rutabaga by ant other name: extracting biological names. pp. 247–259. No. 35, Academic Press (2002)

[10] Hong, Y., Wang, X., Chen, Y., Wang, J., Zhang, T., Zheng, J., Yu, D., Li, Q., Zhang, B., Wang, H., Pan, X., Ji, H.: Rpi blender tac-kbp2014 knowledge base population system. In: Proceedings of the TAC-KBP 2014 Workshop (2014)

[11] Humphreys, K., Demetriou, G., Gaizauskas, R.: Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In: Proceedings of the Pacific Symposium on Biocomputing (PSB-2000). pp. 505–516 (2000)

[12] Indurkhya, N., Damerau, F.J.: Handbook of Natural Language Processing. Chapman & Hall/CRC, 2nd edn. (2010)

[13] Kaufmann, M., Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., Wilks, Y.: University of sheffield: Description of the lasie system as used for muc-6 (1995)

[14] Krallingera, M. , Erhardtb, R. and Valencia, A.: Text-mining approaches in molecular biology and biomedicine. Drug Discovery Today. 10, p. 439–445 (2005)

[15] Linné, C.v., Salvius, L.: Caroli linnaei...systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. 1, 881 (1758)

[16] Nguyen, T.H., He, Y., Pershina, M., Li, X., Grishman, R.: New york university 2014 knowledge base population systems. In: Proceedings of the TAC-KBP 2014 Workshop (2014)

[17] Pazienza, M.T. (ed.): Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, Lecture Notes in Computer Science, vol. 1299. Springer (1997)

[18] Richardson, M., Domingos, P.: Markov logic networks. Mach. Learn. 62, pp.107–136 (2006)

[19] Surdeanu, M., Heng, J.: Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In: Proceedings of the TAC-KBP 2014 Workshop (2014)

[20] Vibhav, G., Dechter, R.: SampleSearch: A scheme that searches for consistent samples. In Proceedings of AISTATS (2007)

[21] Witten, I. H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc. San Francisco, CA, USA (2005)

[22] Zhai, C.,Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Information Retrieval. ACM Trans. Inf. Syst.New York, NY, USA (2004)

[23] Tilman, D.: Causes, consequences and ethics of biodiversity. Nature 405, 208–211 (2000)

[24] Zhou, X., Zhang, X., Hu, X.: Maxmatcher: Biological concept extraction using approximate dictionary lookup. In: PRICAI. vol. 4099, pp. 1145–1149. Springer (2006)