Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 31-38 recibido 03-05-2011 aceptado 22-05-2011

Biomedical event extraction using Kybots

Extracción de eventos biomédicos usando Kybots

Arantza Casillas (*) arantza.casillas@ehu.es
Arantza Díaz de Ilarraza (‡) a.diazdeilarraza@ehu.es
Koldo Gojenola (‡) koldo.gojenola@ehu.es
Maite Oronoz (‡) maite.oronoz@ehu.es
German Rigau (‡) german.rigau@ehu.es

IXA Taldea. University of the Basque Country. UPV/EHU
(*) Department of Electricity and Electronics
(‡) Department of Computer Languages and Systems

Resumen: Este artículo presenta la aplicación de un sistema general de Extracción de Información desarrollado para extraer conocimiento conceptual y factual de textos al dominio específico de la biomedicina. El sistema, previamente desarrollado durante el proyecto KYOTO, está basado en reglas, y es usado para la extracción de eventos biomédicos que implican proteínas y genes en textos anotados pertenecientes al BioNLP11 Shared Task.
Palabras clave: Extracción de información, Biomedicina, PLN

Abstract: This paper shows the applicability of a general Information Extraction technology developed for the extraction of conceptual and factual knowledge from texts, to the specific domain of biomedicine. The rule-based system previously developed for the KYOTO Project is used to extract biomedical events involving proteins or genes from texts annotated in the BioNLP11 Shared Task.
Keywords: Information Extraction, Biomedicine, NLP

1 Introduction

This paper presents the application of a general Information Extraction system to the domain of biomedical texts from the BioNLP Shared Task (Kim *et al.*, 2009) on biomedical event extraction. The applied system is part of a general framework for NLP developed in the KYOTO project[1] (Vossen *et al.*, 2008). The goal of KYOTO is the construction of a system for facilitating the exchange of information across cultures, domains and languages. This system allows, among other objectives, to use different types of information for the detection of knowledge and facts in text. Although the final KYOTO system is domain independent, it has been applied successfully to the domain of environment. The system allows, among many other functionalities, automatic fact mining on document collections. Albeit the KYOTO system allows the use of semantic information from WordNet and ontologies, we have no made use of its semantic capabilities for the present work.

The system will be applied in the context of the BioNLP Shared Task series, including the BioNLP 2009 and its follow-up BioNLP Shared Task 2011[2], which represents the application of fine-grained information extraction (IE) to bio-textmining. The task setup and data have served as the basis of numerous studies and published event extraction systems and datasets. The task defines biologically relevant extraction targets and a linguistically motivated approach to event representation. Manually annotated data where all annotations are bound to specific expressions in text are provided for training, development and evaluation of extraction methods, and tools for the evaluation of system outputs are made available. The main aim of the Genia Shared Task concerns the detection of molecular biology events in biomedical texts using NLP tools and methods. It requires the identification of events together with their gene or protein arguments. Nine event types are considered: localization, binding, gene expression, transcription, protein catabolism, phosphorylation, regulation,

[1] http://www.kyoto-project.eu

[2] http://sites.google.com/site/bionlpst/

ISSN 1135-5948

positive regulation and negative regulation.

The shared task consists in identifying events related to proteins, where it is mandatory to detect the event triggers, together with their associated event-type, and recognize their primary arguments. There are "simple" events, concerning an event together with its arguments (Theme, Site, ...) and also "complex" events, or events that have other events as secondary arguments. Our system did not participate in the optional tasks of recognizing negation and speculation.

As an example, table 1 presents an input text, together with its corresponding annotation files. Each target text consists of 3 files:

- The plain text input file. The files come from PubMed scientific documents (abstracts and full papers), which have been manually annotated with the targeted events and proteins (PMID-9032271.txt)
- The input text is accompanied by its corresponding PMID-9032271.a1 file, which contains stand-off annotations for proteins. In the table, there is an instance (term T5) of a protein ("I kappaB alpha"), with its character offset in the input text. Both the *.txt and *.a1 files are the input files for the task.
- The PMID-9032271.a2 file contains the targeted events and entities to be located by the participating systems. The *.a2 files are used for training and development of the systems, and are also the evaluation result to be obtained. In table 1, T20 represents a term that refers to an event trigger corresponding to a phosphorylation event, and E7 represents a simple event with T20 as a trigger and having the protein T5 filling the *Theme* role.

The files are divided in three sets corresponding to training, development and test sets, as usual. The training dataset contained 909 texts together with a development dataset of 259 texts, and 347 texts were used for testing the system. The final test evaluation results can only be obtained once for each participating system, with the aim of avoiding iterative reuse of the same test set and overfitting.

One of the main objectives of the present work was to verify the applicability of the Information Extraction (IE) technology developed in the KYOTO project, to a new specific domain. The KYOTO system comprises a general and extensible multilingual architecture for the extraction of conceptual and factual knowledge from texts, which has already been applied to the environmental domain.

PMID-9032271.txt
... As well as culminating in the inducible **phosphorylation of I kappaB alpha** on serines 32 and 36, all the stimuli that are inactive on 1.3E2 cells exhibit a sensitivity to the antioxidant pyrrolidine dithiocarbamate (PDTC).
PMID-9032271.a1 file
T5 Protein **1214 1228 I kappaB alpha**
PMID-9032271.a2 file
T20 **Phosphorylation 1195 1210** phosphorylation E7 Phosphorylation:T20 **Theme:T5**

Table 1: An example of the GENIA shared task files.

2 Related Work

Numerous projects have pursued ways of extracting information from text documents. The approaches differ in various features and can hardly be classified strictly, although we could make a broad binary classification into the Knowledge Engineering (KE) approach and the machine learning approach. The interested reader is invited to revise (Sarawagi, 2008) for comprehensive surveys on Information Extraction approaches. In this work we will follow what can be called the *traditional* or KE approach to IE.

The KE approach uses an iterative process, whereas within each iteration the rules are modified as a result of the system's output on a training corpus, consequently demanding a lot of effort. The FASTUS (Hobbs *et al.*, 1997b) system can be considered a *classical* representative of many current KE-based IE systems, as the present one, which follows a rule-based approach (i.e. (Kim *et al.*, 2009), (Cohen *et al.*, 2011) or (Vlachos, 2009)), as opposed to systems based on machine learning.

Regarding general text mining, shared tasks such as those organized in the MUC, TREC and ACE events, have significantly contributed to the progress of their respective fields. This has also been the case in bio-text-mining. With the emergence of Named Entity Recognizers with performance capable

of supporting practical applications, the recent interest of the community shifted toward IE. The BioNLP09 Shared Task addressed bio-IE, concerning the detailed behavior of bio-molecules, characterized as bio-molecular events (bio-events). As the first shared task of its type, the BioNLP task aimed to define a bounded, well-defined bioevent extraction task, considering both the needs and the state of the art in bio-text-mining. Special consideration was given to providing evaluation at diverse levels and aspects, so that the results could drive continuous efforts in relevant directions. (Kim *et al.*, 2009) discusses the design and implementation of the BioNLP task, and reports the main results.

3 System Overview: KYOTO

For the Information Extraction task described in this paper we will make use of the KYOTO technology. Figure 1 gives a detailed overview of the KYOTO architecture. The system includes a "capture module", a "document base" and a "job dispatcher" to manage the documents to be analysed. A "term database" stores new terms that are learned from Kyoto Annotation Format (KAF) representations of documents; and there is a platform for creating pipelines of processing modules through input and output stream connections. The following modules are combined in a pipeline architecture to produce KAF, a term database and facts:

1. Tybots. Extract the terms and their relations using structural, distributional and pattern-based rules.

2. Syntactic processors including tokenization, lemmatized term representation, chunks and dependencies for Dutch, Spanish, Basque, English and Italian. It also includes a multiword tagger that groups sequences of terms based in generic wordnets and domain wordnets.

3. Semantic processors including i) a sense tagger that uses a graph (UKB) and makes word-sense-disambiguation, ii) a Named-entity (NE) tagger for detecting time points and places and applies disambiguation and finally, iii) an ontological tagger (Ontotagger) that reads synsets and inserts the full set of ontological implications.

4. Kybots (Knowledge Yielding Robots). They read KAFs and a specified set of profiles to extract events and facts from KAF, where the profiles can specify patterns at any level of KAF.

The output of this linguistic analysis is stored in an XML annotation format (Agirre et al., 2009) that is the same for all the languages (KAF).

3.1 KAF

KAF adopts a stand-off strategy for annotating the source text and comprises several linguistic annotation levels corresponding to a text:

- Tokenized and segmentized word forms as they appear in the text;
- Lemmatized and typed terms that correspond to one or more word forms;
- Constituents (chunks) that span a series of terms and exhibit head-modifier relations;
- Syntactic dependencies between the constituents;
- Semantic roles for the constituents;
- Facts pointing back to constituents mapped to entities related in time and place.

If a process adds information which cannot be held by existing layers, a new annotation layer is added. Any previous layer will remain intact and can still be used by other processes. Layers may be linked by means of references from one layer to items in another (lower level) layer.

3.2 Kybots

Kybots (Knowledge Yielding Robots) are abstract patterns that detect factual concept instances and relations in KAF. The extraction of factual knowledge by the mining module is done by processing these abstract patterns on the KAF documents and obtaining events as a result. These patterns are defined in a declarative format using Kybot profiles, which describe general morpho-syntactic and semantic conditions on sequences of terms.

The Kybot server reads a profile and compiles it into a program that can be applied to any document collection. Kybot profiles are described using XML syntax and each one consist of three main parts:

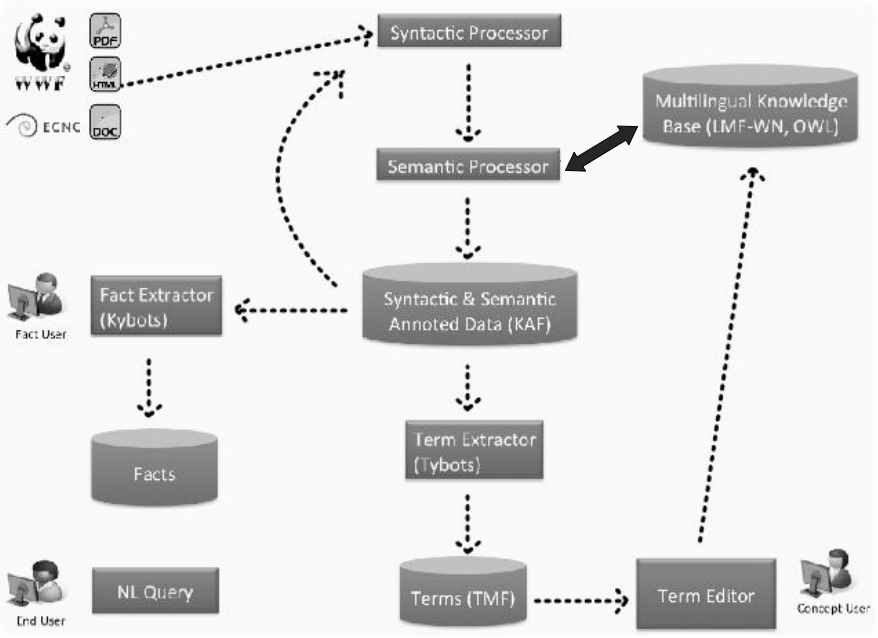

Figure 1: KYOTO System Architecture

- *Variables*: In this part, the entities and its properties are defined. They have been designed with the aim of being flexible enough to deal with many different information associated with the KAF terms including semantic and ontological statements.
- *Relations*: This part specifies the positional relations among the previously defined variables. Define the sequence of variables the Kybot is looking for.
- *Events*: describes the output to be produced for every matching, that is, the output template of the Kybot. For every matched pattern, the kybot produces a new event filling the template structure with the selected pieces of information.

4 Application to the BioNLP shared task

Our system proceeds in two phases. Firstly, text documents are tokenized and converted to KAF (Bosma et al., 2009). Additionaly, the offset positions of the proteins given by the task organizers are also represented in KAF. Secondly, a set of Kybots are applied to detect the biological events of interest occurring in the KAF documents.

Currently, our system only considers a minimal amount of linguistic information. We are only using the word form and term layers. Figure 2 shows an example of a KAF document where proteins have a special tag (`PRT`). Note that our approach did not use any external resource apart of the basic linguistic processing.

Figure 3 shows an example Kybot for detecting phosphorylation events. In this Kybot we have defined three variables named: `Phosphorylation`, `Of` and `Protein`. Kybots also define relations between variables. For example, in the Kybot in figure 3, the variable named `Phosphorylation` is the main pivot, the variable `Of` must follow `Phosphorylation` at a distance of 1 (`immediate` is `true`), and a variable representing a `Protein` must follow `Of` at a distance of 1.

The last part of the Kybot in figure 3 defines the output of the event selecting some of its features represented with the variable called `Phosphorylation`: its term-identification (*@tid*), its lemma (*@lemma*), part of speech (*@pos*) and offset (*@start* and *@end*). The expression also describes that the variable `Protein` plays the role of being the "Theme" of the event. The output obtained when aplying the Kybot in figure 3 is shown in figure 4. Comparing the examples

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<KAF xml:lang="en">
<text>
...
<wf wid="w207" sent="10">in</wf>
<wf wid="w208" sent="10">the</wf>
<wf wid="w209" sent="10">inducible</wf>
<wf wid="w210" sent="10">phosphorylation</wf>
<wf wid="w211" sent="10">of</wf>
<wf wid="w212" sent="10">I</wf>
<wf wid="w213" sent="10">kappaB</wf>
<wf wid="w214" sent="10">alpha</wf>
...
</text>
<term tid="t210" type="open" lemma="phosphorylation" start="1195" end="1210" pos="W">
<span><target id="w210"/></span>
</term>
<term tid="t211" type="open" lemma="of" start="1211" end="1213" pos="W">
<span><target id="w211"/></span>
</term>
<term tid="T5" type="open" lemma="I kappaB alpha" start="1214" end="1228" pos="PRT">
<span>
    <target id="w212"/>
    <target id="w213"/>
    <target id="w214"/>
</span>
</term>...
</terms>
</KAF>
```

Figure 2: Example of a document in KAF format.

in table 1 and in figure 4 we observe that all the features needed for generating the files for describing the results are also produced by the Kybot.

We developed a set of basic auxiliary programs to extract event patterns from the training corpus. These programs obtain the structure of the events, their associated trigger words and their frequency. For example, in the training corpus, a pattern of the type `Event of Protein` appears 35 times, where the *Event* is further described as `phosporylation, phosphorylated...`. Taking the most frequently occurring patterns into account, we manually developed the set of Kybots used to extract the events from the development and test corpora. For example, the Kybot in figure 3 fulfils the conditions of the pattern of interest.

Kybots are applied in two different ways depending on the type of target event we want to detect: *simple* or *complex* events. When extracting *simple* events (see figure 5), we used the input text and the files containing protein annotations (".a1" files in the task) to generate the KAF documents.

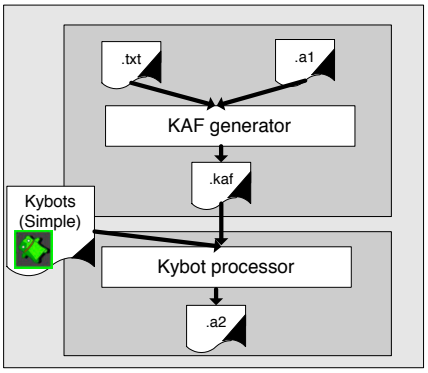

Figure 5: Application of Kybots. Simple events.

These KAF documents and Kybots for simple events are provided to the mining module. As we have said before, complex events are those having other events as arguments. For example, in figure 6 an event of the type "positive regulation" (identified with the trigger word *"culminating"*) has the simple event *"phosphorylation of I kappaB alpha"* as its theme (see table 1 and figure 6). The identifiers of the detected simple events are added to the KAF document in the first phase, with the aim of simplifying the detection of

```
<?xml version="1.0" encoding="utf-8"?>
<!-- Sentence:  phosphorylation of Protein
      Event1:  phosphorylation
      Role:  Theme Protein -->
<Kybot id="bionlp">
<variables>
    <var name="Phosphorylation" type="term" lemma="phosphorylat*''>
    <var name="Of" type="term" lemma="of"/>
    <var name="Protein" type="term" pos="PRT"/>
</variables>
<relations>
    <root span="Phosphorylation"/>
    <rel span="Of" pivot="Phosphorylation" direction="following" immediate="true"/>
    <rel span="Protein" pivot="Of" direction="following" immediate="true"/>
</relations>
<events>
    <event eid="" target="$Phosphorylation/@tid" kybot="phosphorylation_of_P"
    type="Phosphorylation" lemma="$Phosphorylation/@lemma"
    pos="$Phosphorylation/@pos" start="$Phosphorylation/@start" end="$Phosphorylation/@end"/>
    <role target="$Protein/@tid" rtype="Theme" lemma="$Protein/@lemma" start="$Protein/@start"
    end="$Protein/@end"/>
</events>
</Kybot>
```

Figure 3: Example of a Kybot for the pattern `Event of Protein`.

```
<doc shortname="PMID-9032271.kaf">
  <event eid="e1" target="t210" kybot="phosphorylation_of_P" type="Phosphorylation" lemma="phosphorylation"
  start="1195" end="1210" />
  <role target="T5" rtype="Theme" lemma="I kappaB alpha" start="1214" end="1228" />
</doc>
```

Figure 4: Output obtained after the application of the Kybot in figure 3.

complex-events. A new set of Kybots describing complex events is used to obtain the final result (see figure 7).

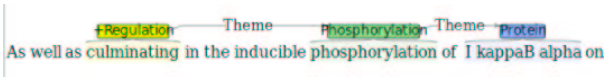

Figure 6: Example of a complex event.

5 *Evaluation*

The files obtained from the Kybots and the ".a2" files giving the results in the BioNLP Shared Task (see table 1) gather the same information, but with different formats. We developed some programs for adapting our output to the required format.

We used the development corpus to improve the Kybot performance. We developed 65 Kybots for detecting simple events. Table 2 shows the number of Kybots for each event type. Complex events relative to regulation (also including negative and positive

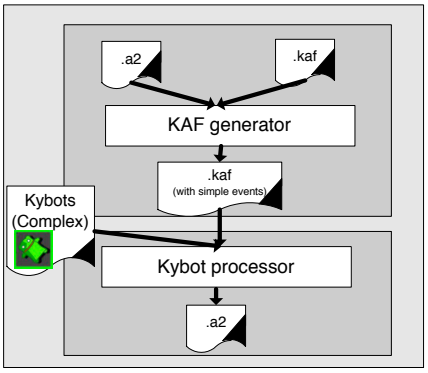

Figure 7: Application of Kybots. Complex events.

regulations) were detected using a set of 24 Kybots.

The evaluation of the task was based on the output of the system when applied to the test dataset of 347 previously unseen texts. Table 3 shows in the `Gold` column the number of instances for each event-type in the test

Event Class	Simple Kybots	Complex Kybots
Transcription	10	
Protein Catabolism	5	
Binding	5	
Regulation		3
Negative Regulation	5	4
Positive Regulation	3	17
Localization	7	
Phosphorylation	18	
Gene Expresion	12	
Total	65	24

Table 2: Number of Kybots generated for each event.

corpus. `R`, `P` and `F-score` columns stand for the recall, precision and f-score the system obtained for each type of event. As a consequence of the characteristics of our system, precision is primed over recall. For example, the system obtains 95% and 97% precision on "Phosphorylation" an "Localization" events, respectively, although its recall is considerably lower (41% and 19%). Sometimes some linguistic structures describing events overlap among different event types, creating ambiguity. For example, the "Gene Expression" event-type appears very frequently in the corpus (1002 occurrences) so describing a pattern `expression of protein` seems to be very fruitful. However, the same structure describes, for example, events of type "Transcription" and "Localization". As we do not use any linguistic information for disambiguation, our precision is not very good in, for instance, "Gene Expression" (42.22%).

Event Class	Gold	R	P	F-score
Localization	191	19.90	97.44	33.04
Binding	491	5.30	50.00	9.58
Gene Expression	1002	54.19	42.22	47.47
Transcription	174	13.22	62.16	21.80
Protein catabolism	15	26.67	44.44	33.33
Phosphorylation	185	41.62	95.06	57.89
Non-reg total	2058	34.55	47.27	39.92
Regulation	385	7.53	9.63	8.45
Positive regulation	1443	6.38	62.16	11.57
Negative regulation	571	3.15	26.87	5.64
Regulatory total	2399	5.79	26.94	9.54
All total	4457	19.07	42.08	26.25

Table 3: Performance analysis on the test dataset.

After the final evaluation, our system obtained the thirteenth position out of 15 participating systems in the main task (processing PubMed abstracts and full documents), obtaining 19.07%, 42.08% and 26.25 recall, precision an f-score, respectively, far from the best competing system (49.41%, 64.75% and 56.04%). Although they are far from satisfactory, we must take into account the short time we dedicated to adapting the system and designing the kybots, which can be roughly estimated in two person/months. Apart from that, due to time restrictions, our system did not make use of the ample set of resources available, such as named entities, coreference resolution or syntactic parsing of the sentences. On the other hand, the system, based on manually developed rules, obtains reasonable accuracy in the task of processing full documents, obtaining 45% precision and 21% recall, compared to 59% and 47% for the best system, which means that the rule-based approach performs more robustly when dealing with long texts (each full text corresponds to 150 abstracts). As we have said before, our main objective was to evaluate the capabilities of the KYOTO technology without adding any additional information. The use of more linguistic information probably will facilitate our work and will benefit the system results.

6 Conclusions and Future work

This work presents the first results of the application of the KYOTO text mining system for extracting events when ported to the biomedical domain. The KYOTO technology and data formats have shown to be flexible enough to be easily adapted to a new task and domain. Furthermore, high precision kybot profiles have been developed for this new domain.

In a near future we plan to apply machine learning techniques for the automatic generation of Kybots from the training data. We also plan to include additional linguistic and semantic processing in the event extraction process to exploit the current semantic and ontological capabilities of the KYOTO technology.

Acknowledgements

This work has been supported by the Spanish Ministerio de Ciencia e Innovación (KNOW2 TIN2009-14715-C04-01), the European Comission (KYOTO ICT-2007-211423) and the Basque Government (IT344-10).

Bibliografía

Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini and Carlo Aliprandi. *KAF: a generic semantic annotation format* Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL 2009 Pisa, Italy, September 17-19, 2009

Eneko Agirre, Xabier Artola, Arantza Díaz de Ilarraza, German Rigau, Aitor Soroa, Wauter Bosma. *KAF: Kyoto Annotation Format* Tech. rep., University of the Basque Country, dept. Computer Science and Artificial Intelligence, Donostia-San Sebastian. 2009.

Kevin Bretonnel Cohen, Karin Verspoor, Helen L. Johnson, Chris Roeder, Philip V. Ogren, Willian A. Baumgartner, Elizabeth White, Hannah Tipney, and Lawerence Hunter. High-precision biological event extraction: Effects of system and data. *Computational Intelligence*, to appear, 2011.

Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel and Mabry Tyson. FASTUS: A cascaded finite-state transducer for extracting informat ion from natural-language text. *Finite-State Language Processing*, Cambridge, MA ,1997. MIT Press.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano and Jun'ichi Tsujii. Overview of BioNLP'09 Shared Task on Event Extraction. *Proceedings of the BioNLP 2009 Workshop*. Association for Computational Linguistics. Boulder, Colorado, pp. 89–96., 2011

Sunita Sarawagi. Information extraction. Fnt Databases, 1(3), 2008.

Andreas Vlachos. Two Strong Baselines for the BioNLP 2009 Event Extraction Task. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics Uppsala, Sweden, pp. 1–9., 2010

Piek Vossen, Eneko Agirre, Nicoletta Calzolari, Christiane Fellbaum, Shu-kai Hsieh, Chu-Ren Huang, Hitoshi Isahara, Kyoko Kanzaki, Andrea Marchetti, Monica Monachini, Federico Neri, Remo Raffaelli, German Rigau, Maurizio Tescon, Joop VanGent. KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures. *Proceedings of LREC 2008*. Marrakech, Morocco, 2008.