Using Kybots for Extracting Events in Biomedical Texts

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 UPV/EHU
(*) Department of Electricity and Electronics
(‡) Department of Computer Languages and Systems

Abstract

In this paper we describe a rule-based system developed for the BioNLP 2011 GENIA event detection task. The system applies Kybots (Knowledge Yielding Robots) on annotated texts to extract bio-events involving proteins or genes. The main goal of this work is to verify the usefulness and portability of the Kybot technology to the domain of biomedicine.

1 Introduction

The aim of the BioNLP'11 Genia Shared Task (Kim *et al.*, 2011b) 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, positive regulation and negative regulation.

When identifying the events related to the given proteins, it is mandatory to detect also the event triggers, together with its 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 secundary arguments. Our system did not participate in the optional tasks of recognizing negation and speculation.

The training dataset contained 909 texts together with a development dataset of 259 texts. 347 texts were used for testing the system.

The main objective of the present work was to verify the applicability of a new Information Extraction (IE) technology developed in the KYOTO project[1] (Vossen *et al.*, 2008), 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.

Currently, our system follows a rule-based approach (i.e. (Kim *et al.*, 2009), (Kim *et al.*, 2011a), (Cohen *et al.*, 2011) or (Vlachos, 2009)), using a set of manually developed rules.

2 System Description

Our system proceeds in two phases. Firstly, text documents are tokenized and structured using an XML layered structure called *KYOTO Annotation Format* (KAF) (Bosma et al., 2009). Secondly, a set of *Kybots* (Knowledge Yielding Robots) are applied to detect the biological events of interest occurring in the KAF documents. Kybots form a collection of general morpho-syntactic and semantic patterns on sequences of KAF terms. These patterns are defined in a declarative format using Kybot profiles.

2.1 KAF

Firstly, basic linguistic processors apply segmentation and tokenization to the text. Additionally, the offset positions of the proteins given by the task organizers are also considered. The output of this basic processing is stored in KAF, where words, terms, syntactic and semantic information can be stored in separate layers with references across them.

Currently, our system only considers a minimal amount of linguistic information. We are only using

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

the word form and term layers. Figure 1 shows an example of a KAF document where proteins have been annotated using a special POS tag (PRT). Note that our approach did not use any external resource apart of the basic linguistic processing.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<KAF xml:lang="en">
<text>
...
<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<...
</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"/></span>
    <target id="w213"/>
    <target id="w214"/></span>
</term>...
</terms>
</KAF>
```

Figure 1: Example of a document in KAF format.

2.2 Kybots

Kybots (Knowledge Yielding Robots) are abstract patterns that detect actual 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. These patterns are defined in a declarative format using Kybot profiles, which describe general morpho-syntactic and semantic conditions on sequences of terms. Kybot profiles are compiled to XQueries to efficiently scan over KAF documents uploaded into an XML database. These patterns extract and rank the relevant information from each match.

Kybot profiles are described using XML syntax and each one consists of three main declarative parts:

- *Variables*: In this part, the entities and its properties are defined
- *Relations*: This part specifies the positional relations among the previously defined variables
- *Events*: describes the output to be produced for every matching

Variables (see the Kybot section *variables* in figure 2) describe the term variables used by the Kybot. 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 (see the Kybot section *relations* in figure 2) define the sequence of variables the Kybot is looking for. For example, in the Kybot in figure 2, the variable named `Phosphorylation` is the main pivot, the variable `Of` must follow `Phosphorylation` (`immediate` is `true` indicating that it must be the next term in the sequence), and a variable representing a `Protein` must follow `Of`. Proteins and genes are identified with the `PRT` tag.

Events (expressions marked as *events* in figure 2) describes 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. For example, the Kybot in figure 2 selects some features of the event represented with the variable called `Phosphorylation`: its term-identification (*@tid*), its lemma, part of speech and offset. 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 2 is shown in figure 3. Comparing the examples in table 1 and in figure 3 we observe that all the features needed for generating the files for describing the results are also produced by the Kybot.

```
<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 3: Output obtained after the application of the Kybot in figure 2.

3 GENIA Event Extraction Task and Results

We developed a set of basic auxiliary programs to extract event patterns from the training corpus. These programs obtain the struc-

```
<?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 2: Example of a Kybot for the pattern `Event of Protein`.

.a1 file
T5 Protein **1214 1228 I kappaB alpha**
.a2 file
T20 **Phosphorylation 1195 1210** phosphorylation E7 Phosphorylation:T20 **Theme:T5**

Table 1: Results in the format required in the GENIA shared task.

ture 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 in the training data into account, we manually developed the set of Kybots used to extract the events from the development and test corpora. For example, in this way we wrote the Kybot in figure 2 that fulfils the conditions of the pattern of interest.

The two phases mentioned in section 2, corresponding to the generation of the KAF documents and the application of Kybots, have different input files depending on the type of event we want to detect: *simple* or *complex* events. When extracting *simple* events (see figure 4), we used the input text and the files containing protein annotations (".a1" files in the task) to generate the KAF documents. These KAF documents and Kybots for simple events are provided to the mining module. In the case of *complex* events (events that have other

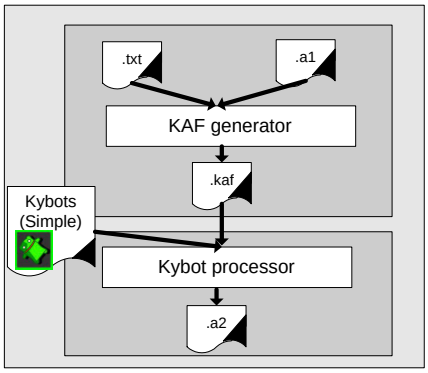

Figure 4: Application of Kybots. Simple events.

events as arguments), the identifiers of the detected simple events are added to the KAF document in the first phase. A new set of Kybots describing complex events and KAF (now with annotations of the simple events) are used to obtain the final result (see figure 5).

For the evaluation, we also developed some programs for adapting the output of the Kybots (see figure 3) to the required format (see table 1).

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 regulations) were detected using a set of 24 Kybots.

The evaluation of the task was based on the output

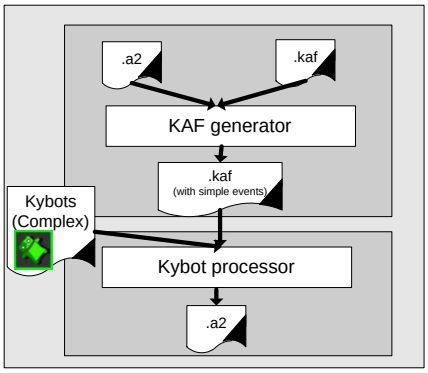

Figure 5: Application of Kybots. Complex events.

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

Table 2: Number of Kybots generated for each event.

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 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%).

4 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. Although the results are far from satisfactory, we must take into account the limited effort we dedicated to adapting the system and designing the kybots, which can be roughly estimated in two

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.

person/months.

After the final evaluation, our system obtained the thirteenth position out of 15 participating systems in the main task (processing PubMed abstracts and full paper articles), 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 limited time we dedicated to adapting the system and designing the kybots. 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 paper articles, 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 (5 full texts correspond to approximately 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 will probably facilitate our work and will benefit the system results. In the near future we will study the application of 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.

Acknowledgments

This research was supported by the the KYOTO project (STREP European Community ICT-2007-211423) and the Basque Government (IT344-10).

References

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

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.

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

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, and Junichi Tsujii. 2011a. Overview of BioNLP Shared Task 2011. *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*. Association for Computational Linguistics. Portland, Oregon, 2011.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011b. Overview of the Genia Event task in BioNLP Shared Task 2011. *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*. Association for Computational Linguistics. Portland, Oregon, 2011.

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.