Automatic Verbalisation of Biological Events

Bikash Gyawali, Claire Gardent, Christophe Cerisara

Automatic Verbalisation of Biological Events

Bikash Gyawali [1], Claire Gardent [2] and Christophe Cerisara [2]
[1]Université de Lorraine, Nancy, France
[2]CNRS and Université de Lorraine, Nancy, France

ABSTRACT

We present a method for automatically generating descriptions of biological events encoded in the KB BIO 101 Knowledge base. In this knowledge base, events are concepts (e.g., RELEASE) related by role relations (e.g., AGENT, PATIENT, PATH, INSTRUMENT) to the concepts denoting their arguments (e.g., GATED-CHANNEL, VASCULAR-TISSUE, IRON). We propose a probabilistic, unsupervised method which extracts possible verbalisation frames from large biology specific domain corpora and uses probabilities both to select an appropriate frame given an event description and to determine the mapping between syntactic and semantic arguments. That is, probabilities are used to determine which event argument fills which syntactic function (e.g., subject, object) in the produced verbalisation. We evaluate our approach on a corpus of 336 event descriptions, provide a qualitative and quantitative analysis of the results obtained and discuss possible directions for further work.

1 INTRODUCTION

An ontology specifies a conceptualisation of a given domain by listing the objects, concepts and other entities that are assumed to exist in that domain and the relationships that hold among them (Genesereth and Nilsson, 1987). To constrain the possible interpretations for the defined terms, class definitions are also often included. These can be provided either as logical axioms or as natural language text. In general though, both styles (logical axioms and natural language descriptions) are equally important. Axioms are required to support reasoning while natural language descriptions are needed to help development (by improving inter-annotator agreement) and to facilitate usage by non experts. Indeed, in bio-ontologies, the provision of a textual definition for each entity present in the ontology has become one of the OBO Foundry criteria (Smith *et al.*, 2007).

Authoring both logical and textual definitions and keeping them consistent is time consuming however. To address this shortcoming, we therefore propose to explore ways of automatically generating natural language text from OWL data. As a first step towards this goal, we focus on the verbalisation of single events using data from the KB BIO 101 Knowledge base.

KB BIO 101 (Chaudhri *et al.*, 2013) was developed by the Halo Project to represent a significant fraction of an introductory college-level biology textbook (Reece *et al.*, 2011) and was used as part of a prototype of an intelligent digital textbook called Inquire designed to help students to learn better. The knowledge base (KB BIO 101) underlying this digital textbook contains descriptions of biological events and of their interrelation. To facilitate the description to the user of these events, we propose a method for automatically producing a natural language verbalisation of the event descriptions contained in the KB BIO 101 Knowledge base.

The paper is structured as follows. In Section 2, we present the method used to verbalise KB events and their participants. In Section 3, we situate our approach with respect to previous work, evaluate our approach on a corpus of 336 event descriptions, provide a qualitative and quantitative analysis of the results obtained and discuss possible directions for further work. Section 4 concludes.

2 METHODOLOGY

As mentioned above, our goal is to automatically generate natural language verbalisations of the event descriptions contained in KB BIO 101. To ensure portability to other domains, we develop an unsupervised method in which the natural language information required to produce the verbalisations is extracted from automatically constructed domain specfic corpora. The development of our generation system involves the following main steps.

Corpus Building. We first gather a large domain specific corpus from the web i.e., digitised texts which bear on biology.

Lexicon Creation. For each event and entity in the input KB, we build a lexicon associating event and entities with synonyms and morphological variants.

Frame extraction. For each event in the input KB, syntactic frames are extracted from the corpus using the lexicon as a bridge between KBGEN+ event/entity names and their natural language lexicalisations. Frequency counts are gathered about the number of times a given frame occurs, the event and roles it represents and the syntactic dependencies binding argument(s) in the frame.

Probabilistic Frame Selection. Given an input event KB representation, the set of frames associated by Frame extractions with mentions of that event is retrieved and ranked by decreasing order of probability. The most probable frame given the input event KB representation is chosen for generation.

Probabilistic Argument Linking. Given an input event KB representation and a syntactic frame, all possible mappings between KB and syntactic arguments are considered and the most probable mapping given the input event KB representation is selected.

Slot Filling The frame slots are filled with verbalisations of the arguments and of the events thus producing a verbalisation of the input KB event and its arguments.

We start by giving a brief overview of the content and the structure of KB BIO 101(Section 2.1). We the describe the steps involved in building our generation system.

2.1 KB Bio 101

The foundational component of the KB is the Component Library (CLIB), an upper ontology which is linguistically motivated and designed to support the representation of knowledge for automated reasoning (Gunning *et al.*, 2010). CLIB adopts four simple top level distinctions: (1) entities (things that are); (2) events (things that happen); (3) relations (associations between things); and (4) roles (ways in which entities participate in events). Using this ontological inventory, KB BIO 101 encodes events, the entities that participate in

	Events		Entities		Roles	
	# Types	# Tokens	# Types	# Tokens	# Types	# Tokens
KBGEN+	126	336	271	929	14	929

Table 1. KBGEN+ Statistics

events and roles that the entities play in an event. Events and entities are concepts while roles are Event-to-Entity relations.

Figure 1 shows an example representation for a blocking event between a plasma membrane and hydrophobic compounds. *Block* is a subclass of the event class. *Plasma-Menbrane* and *Hydrophobic-Compound* are subclasses of the entity class. The *Plasma-Menbrane* and the *Hydrophobic-Compound* concepts stand respectively in an *instrument* and in an *object* role relation with the *Block* event.

```
SubClassOf(:Hydrophobic-Compound :Entity)
SubClassOf(:Plasma-Membrane :Entity)
SubClassOf(:Block
 ObjectIntersectionOf(:Event
  ObjectSomeValuesFrom(:instrument :Plasma-Membrane)
  ObjectSomeValuesFrom(:object :Hydrophobic-Compound)))
```

Fig. 1. Example Event Representation in KB BIO 101

KB BIO 101 is organized into a set of concept maps, where each concept map corresponds to a biological entity or process. It was encoded by biology teachers and contains around 5,000 concept maps. KB BIO 101 is available for download for academic purposes in various formats including OWL[1] .

To test and evaluate our approach, we focus on the subpart of KB BIO 101 isolated for the KBGEN surface realisation shared task by (Banik *et al.*, 2013). In this dataset, content units were semi-automatically selected from KB BIO 101 in such a way that:

- the set of relations in each content unit forms a connected graph
- each content unit can be verbalised by a single, possibly complex sentence which is grammatical and meaningful
- the set of content units contain as many different relations and concepts of different semantic types (events, entities, properties, etc) as possible.

That is, the KB content extracted for KBGEN isolate event descriptions which can be verbalised by a single, coherent sentence. To evaluate the ability of our generator to generate event description, we further process this dataset to produce all KB fragments which represent a single event. The statistics for the resulting dataset (dubbed KBGEN+) are shown in Table 2. More detailed statistics about the input data we test generation on is given in Table1.

Items	Count
Total nb of Triples set	336
Avg. nb of relations in a triple set	2.93
Total nb of distinct events	126
Total nb of distinct entities	271
Total nb of distinct relations	14

Table 2. Input Statistics

2.2 Corpus Collection

We begin by gathering sentences from several of the publicly available Biomedical domain corpus[2] This includes the BioCause (Mihil *et al.*, 2013), BioDef [3], BioInfer (Pyysalo *et al.*, 2007), Grec (Thompson *et al.*, 2009), Genia[4] and PubMedCentral (PMC)[5] corpus. We also include sentences provided by the KB BIO 101 challenge. This custom collection of sentences will be the corpus on which our unsupervised learning approach will build upon. Table 3 lists the count of sentences available in each corpus and in total.

	#Sentences
BioCause	3,187
BioDef	8,426
BioInfer	1,100
Genia	37,092,000
Grec	2,035
PMC	7,018,743
BioKB101	3,393
Total	44,128,884

Table 3. Corpus Size

2.3 Lexicon Creation

To enable the description in natural language of KB content, knowledge about how relations and concepts are realised in natural language is required. One way to capture such knowledge is by specifying a lexicon mapping concept and relation names to natural language words or phrases. Ideally this lexicon should map each concept/relation to the set of lexical and phrasal variants lexicalising that concept/relation and to their various forms (e.g., both singular and plural for a noun).

To identify corpus sentences which might contain verbalisation of KB events, we first build such a lexicon making use of existing resources namely, the lexicon provided by the KBGEN challenge and

[1] http://www.ai.sri.com/halo/halobook2010/exported-kb/biokb.html

[2] Ideally, since KB BIO 101 was developed based on a textbook, we would use this textbook as a corpus. Unfortunately, the textbook, previously licensed from Pearson, is no longer available.

[3] Obtained by parsing the ⟨Supplement⟩ section of html pages crawled from http://www.biology-online.org/dictionary/

[4] http://www.nactem.ac.uk/genia/

[5] ftp://ftp.ncbi.nlm.nih.gov/pub/pmc

synsets automatically extracted from the Mesh vocabulary[6] and the BioDef lexicon[7].

The KBGEN lexicon is composed of entries that provide inflected forms and nominalizations for the event variables and singular and plural noun forms for the entity variables, such as :

Secretion , secretes , secrete , secreted , secretion
Earthworm , earthworm , earthworms

On the other hand, the lexical entries obtained from Mesh and BioDef are usually synsets, such as :

Block , prevent , stop
Neoplasms , Neoplasm , Tumors , Neoplasia , Cancer

In Table 4, we list the count of lexical entries available in each source and those that co-occur in our input. Table 5 shows the proportion of KBGEN+ event and entity classes for which a lexical entry was found as well as the max, min and average number of lexical items associated with event and entities.

	KBGen	Mesh	BioDef
#Lexical Entries	469	26795	14934
#Intersecting Entries	397	65	99

Table 4. Lexical Entries

	KBGen	Mesh	BioDef	ALL	Min/MAx/Avg
Event	100%	10.31%	25.39%	100%	1/21/5.68
Entity	100%	19.18%	24.72%	100%	3/18/3.88
All	100%	16.37%	24.93%	100%	1/21/4.40

Table 5. Proportion of Event and Entity Names for which a Lexical Entry was found. Min, max and average number of lexical items associated with event and entities

2.4 Frame Extraction

Events in KBGEN+ take an arbitrary number of participants ranging from 1 to 8. Knowing the lexicalisation of an event name is therefore not sufficient. For each event lexicalisation, information about syntactic subcategorisation and syntactic/semantic linking is also required. Consider for instance, the following event representation:

```
SubClassOf(:PC/EBP beta :Entity)
SubClassOf(:TNF-activation :Entity)
SubClassOf(:Myeloid-Cells :Entity)
SubClassOf(:Block
 ObjectIntersectionOf(:Event
  ObjectSomeValuesFrom(:instrument :C/EBP beta)
  ObjectSomeValuesFrom(:object :TNF-activation)))
  ObjectSomeValuesFrom(:base :Myeloid-Cells)))
```

Knowing that a possible lexicalisation of a *Block* event is the finite verb form *blocked* is not sufficient to produce an appropriate verbalisation of the KB event e.g.,

(1) C/EBP beta blocked TNF activation in myeloid cells.

In addition, one must know that this verb (i) takes a subject, an object and an optional prepositional argument introduced by a locative preposition (subcategorisation information) and (ii) that the INSTRUMENT role is realised by the subject slot, the OBJECT role by the DOBJ slot and the BASE role by the PREP-LOC slot (syntax/semantics linking information). That is, we need to know, for each KB event e and its associated roles (i.e., event-to-entity relations), first, what are the syntactic arguments of each possible lexicalisations of e and second, for each possible verbalisation, which role maps to which syntactic function.

To address this issue, we extract syntactic frames from our constructed corpus and use the collected data to learn the mapping between KB and syntactic arguments.

Frame extraction proceeds as follows. For each event name in the KBGEN+event set, we look for sentences in the corpus that mention this event name or one of its several verbalisations available in the merged lexicon (ALL in Table 5).

From all such sentences, event frames are then extracted where an event frame is a syntactic frame obtained from the dependency parse tree of the sentence by selecting the local subtree originating at the node labelled with the event name (or one of its variants). For instance, given the sentence and the dependency tree shown in Figure 2, the extracted frame will be:

nsubj:NP,VB,dobj:NP

indicating that the verb form *block* requires a subject and an object noun phrase (NP). That is, a syntactic frame describes the arguments required by the lexicalisations of an event, the syntactic function they realise and their syntactic category (e.g., NP).

We use the Stanford Dependency Parser[8] to produce the dependency trees and take as input for frame extraction, the collapsed typed dependency variant. NP variants (NN, NNS, NNP, NNPS, PRP, PRP) are generalized as NP and the VB variants (VB ,VBD ,VBG ,VBN ,VBP ,VBZ) are generalized as VB. When extracting the frames, we only consider a subset of the dependency relations[9] produced by the Stanford parser to avoid including in the frame adjuncts such as temporal or spatial phrases which are optional rather than required arguments.

Note that for each event, many event frames can arise from a single sentence (if the sentence has multiple mentions of the event) and several sentences can be extracted for the same event. Also note that the same event frame can be observed for different events in the event set although their event representation may differ. Table 6 lists the count of events on our event set for which at least a sentence was found in the individual and total corpus.

A total of 2383 distinct event frames were observed whereby 96.06% of the KBGEN+events were assigned at least one frame and each event was assigned an average of 116.48 distinct frames. The high variety of frames assigned to each event results from both lexical and syntactic variations. Each event can be lexicalised by

[6] http://www.nlm.nih.gov/mesh/filelist.html

[7] Obtained by parsing the entries in ⟨Synonyms⟩ section of html pages crawled from http://www.biology-online.org/dictionary/

[8] http://nlp.stanford.edu/software/lex-parser.shtml

[9] The dependency relations considered for frame construction are: *advcl, agent, appos, csubj, csubjpass, dobj, expl, iobj, mwe, nn, npadvmod, nsubj, nsubjpass, number, pobj, possessive, tmod, vmod* and the variants of *prep_* and *prepc_*.

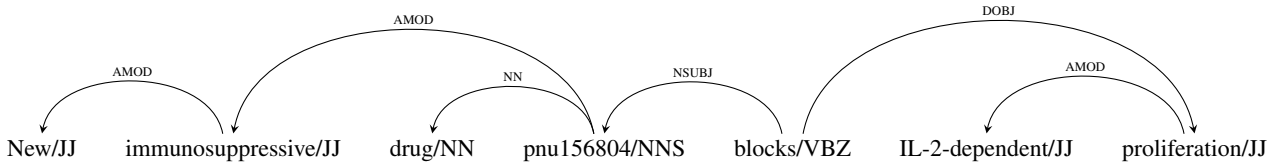


Fig. 2. Example Dependency Parse Tree

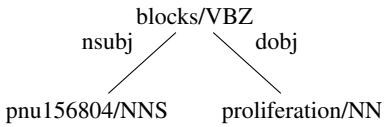
	#Events Found	#Stces/Event
BioCause	30	5
BioDef	71	9.9
BioInfer	31	4
Genia	51	26.33
Grec	32	5.15
PMC	93	2163.43
BioKB101	58	22.15
TOTAL	122	1078.59

Table 6. Sentences containing Event Verbalisations

several natural language words or phrases and each natural language expressions may occur in several syntactic environments.

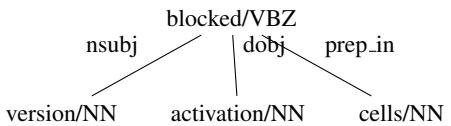
Here are some examples showing for a given event name and a corpus sentence, the tree extracted from the dependency parse tree and the corresponding frame.

Event BLOCK
Sentence *New immunosuppressive drug PNU156804 blocks IL-2-dependent proliferation and nf-kappa b and ap-1 activation.*
Tree



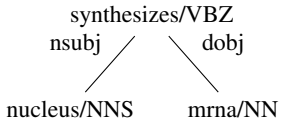
Frame nsubj:NP,VB,dobj:NP

Event BLOCK
Sentence *Finally, a dominant-negative version of C/EBP beta blocked TNF alpha promoter activation in myeloid cells.*
Tree



Frame nsubj:NP,VB,dobj:NP,prep_in:NP

Event SYNTHESIS-OF-FAT
Sentence *Nucleus synthesizes mrna using instructions provided by the DNA.*
Tree



Frame nsubj:NP,VB,dobj:NP

2.5 Probabilistic Learning

Using the frequency counts produced by the frame extraction process, we estimate the probability $P(f|e)$ of frame f given an event e as follows:

$$P(f|e) = \frac{counts(f,e)}{counts(e)}$$

where *counts(f,e)* is the number of time the combination of frame f and event e was observed and *counts(e)* is the number of time event e was observed.

We further observe that an entity can be bound via different relations to different/same events in different triplesets of the input. Thus we build a mapping of each entity in the entity set to the set of all relations it is bound with across all the events in the event set and refer it as entity-relation map. On average, an entity was found to be bound via 1.35 distinct relations.

The entity-relation map will be useful in computing the likelihood of roles to be ascribed to event frames. We assume that if an event dependency tree bears a mention of any entity (or one of its several verbalisations available from the lexicon) from the entity set as its immediate dependent node, the corresponding event frame of this event dependency tree can be ascribed to all of the roles associated for that entity in the entity-relation map. By checking this criteria on all of the event dependency trees obtained for all events in the event set, we compute the frequency of all of the roles available for each event frames and this will give us the probability $P(f \mid r)$ which we estimate as follows :

$$P(f|r) = \frac{counts(f,r)}{counts(r)}$$

Add-one smoothing is applied to account for any role that does not hold for the given frame.

One final probabilistic parameter that we compute from the event frames is the likelihood of the syntactic dependency relation (Dep)

being assigned to the role relation. Here, we assume that the dependency relation of the event dependency tree binding an entity (or one of its several verbalisations available from the lexicon) serves as a representative of all the roles associated to the entity in the entity-relation map. By checking this criteria on all of the event dependency trees obtained for all events in the event set, we compute the frequency of all of the dependency relations available for each role and this will give us the probability $P(d \mid r)$. We estimate $P(d \mid r)$ using frequency counts as follows :

$$P(d|r) = \frac{counts(d, r)}{counts(r)}$$

2.6 Surface Realisation

The set of probabilities so learnt are then used for generating sentences to verbalize input triplesets as follows. Given an input tripleset to verbalize, we first identify the event name and the roles present in the input. All possible frames for the current event can be retrieved but we only retain those frames (for further processing) that have an arity matching the number of roles in the input. Then, for each frame thus obtained, we compute its probability in conjunction with all the roles present in the input, i.e.

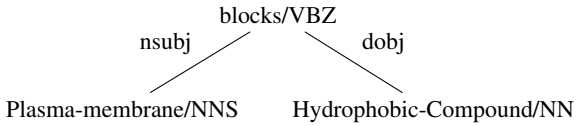
$$P(f|e) \times P(f|r_1) \times \ldots \times P(f|r_n) \quad (1)$$

We select the top 5 highest scoring frames given by Equation 1 as our candidate frames for generation. Each of those frames serves as a template for our generation task. In particular, we replace the dependents of the root node with the entities of the roles of the input so that $\prod_i P(d_i|r_i)$ is maximized. An example below will illustrate (For readability reasons, the probabilities are in logarithmic of base 10) :

Input:

```
SubClassOf(:Plasma-membrane :Entity)
SubClassOf(:Hydrophobic-Compound :Entity)
SubClassOf(:Block
 ObjectIntersectionOf(:Event
  ObjectSomeValuesFrom(:instrument :Plasma-membrane)
  ObjectSomeValuesFrom(:object :Hydrophobic-Compound)))
```

Selected Frame : nsubj:NP,VB,dobj:NP
Known P(d|r) :
P(nsubj|instrument)=-1.25, P(dobj|instrument)=-1.15
P(nsubj|object)=-0.99, P(dobj|object)=-0.76
Assignment Possibilities :
instrument $\rightarrow$ nsubj & object $\rightarrow$ dobj = -1.25+-0.76 = -2.01
instrument $\rightarrow$ dobj & object $\rightarrow$ nsubj = -1.15+-0.99 = -2.14
Chosen Assignment :



Generated Sentence : Plasma membrane blocks Hydrophobic compounds.

3 DISCUSSION AND EVALUATION

3.1 Related Work

There has been much research in recent years on developing natural language generation systems which support verbalising knowledge bases.

Many of the existing KB Verbalising tools rely on generating so-called Controlled Natural Languages (CNL) i.e., a language engineered to be read and written almost like a natural language but whose syntax and lexicon is restricted to prevent ambiguity. Some CNLs are completely formal and can be automatically mapped to logic. Examples of such languages include ACE (Fuchs *et al.*, 2008), CELT (Pease and Li, 2010), CLCE (Pool, 2006), CLP (Clark *et al.*, 2005), Formalized-English (Martin, 2002) and PENG (Kaljurand and Fuchs, 2007).

Thus, the OWL verbaliser integrated in the Protégé tool is a CNL based generation tool, (Kaljurand and Fuchs, 2007) which provides a verbalisation of every axiom present in the ontology under consideration and (Wilcock, 2003) describes an ontology verbaliser using XML-based generation.

More complex NLG system have also been developed to generate text (rather than simple sentences) from knowledge bases. Thus, the MIAKT project (Bontcheva and Wilks., 2004) and the ONTOGENERATION project (Aguado *et al.*, 1998) use symbolic NLG techniques to produce textual descriptions from some semantic information contained in a knowledge base. Both systems require some manual input (lexicons and domain schemas). More sophisticated NLG systems such as TAILOR (Paris, 1988), MIGRAINE (Mittal *et al.*, 1994), and STOP (Reiter *et al.*, 2003) offer tailored output based on user/patient models. While offering more flexibility and expressiveness, these systems are difficult to adapt by non-NLG experts because they require the user to understand the architecture of the NLG systems (Bontcheva and Wilks., 2004). Similarly, the NaturalOWL system (Galanis *et al.*, 2009) has been proposed to generate fluent descriptions of museum exhibits from an OWL ontology. This approach however relies on extensive manual annotation of the input data. Finally, recent work by the SWAT project[10] has focused on producing descriptions of ontologies that are both coherent and efficient (Williams and Power, 2010).

More recently, statistical, data-driven approaches have focused on learning a generation system from parallel corpora of data and text. In particular, (Angeli *et al.*, 2010; Chen and Mooney, 2008; Wong and Mooney, 2007; Konstas and Lapata, 2012b,a) trained and developed data-to-text generators on datasets from various domains including the air travel domain (Dahl *et al.*, 1994), weather forecasts (Liang *et al.*, 2009; Belz, 2008) and sportscasting (Chen and Mooney, 2008). Here, the dominant approach consists in learning a direct mapping between meaning representations and natural language.

Our approach differs from previous work in two main ways.

First, it is unsupervised. As mentioned above, most of the previous work on generating from knowledge bases either makes use of hand-crafted grammars (and sometimes lexicons) or of a parallel data/text corpus. In both cases, considerable time and expertise must be spent on developing the required linguistic resources (aligned data-text corpus, grammar, lexicon)

[10] `http://crc.open.ac.uk/Projects/SWAT`

thereby restricting domain independence. Porting the system to a new domain is therefore costly. In contrast, our approach is fully unsupervised; extracting and learning the relevant linguistic and probabilistic information from available text only corpora. Consequently, it can be used for any knowledge base for which there exists large textual corpora.

Second, we focus on the verbalisation of n-ary relations and on the task of appropriately mapping KB roles to syntactic functions. Little attention has been paid to this issue so far. In symbolic approaches (e.g., the CNL or KB based approaches mentioned above), this mapping is determined by the lexicon and must be manually specified. In data-driven approaches on the other hand the mapping is learned from the alignment between text and data. In both cases, predicting the appropriate mapping depends on having the appropriate linguistic resources (manually specified lexicon , parallel data-text corpus) and is restricted to cases which either are specified in the lexicon or have been seen in the training data. Instead, we view the syntax/semantic mapping task as a bipartite graph alignment problem (each syntactic function must be aligned with exactly one semantic role and vice versa) and learn a probabilistic model designed to select the most probable mapping. In this way, we provide a domain independent, fully automatic, means of verbalising n-ary relations.

3.2 Evaluation

We evaluate our approach on the 336 event representations included in the KBGEN$^+$dataset. For each event representation, we generate the 5 best natural language verbalisations using the method described in the preceding section. We then evaluate the results both qualitatively and quantitatively.

We first consider coverage i.e., the proportion of input in the test set for which a verbalisation is produced. Because we limit the choice of selected frames to the ones that bear the right arity and are VB rooted frame, we fail in selecting a frame for some of the input. There are 45 input cases for which none of the selected frames had a matching arity and 9 input cases where a VB rooted event frame was not found. Thus we generated an output for 82.5% of the input dataset. We are currently investigating whether relaxing these constraints would improve coverage and how it would impact the quality of the generated verbalisations.

Taking a random sample of 100 inputs from the KBGEN$^+$dataset, we examine the quality of the output, in particular the syntax/semantic mapping induced by our probabilistic model and the lexicalisation of events and arguments. For each randomly sampled input, we consider the 5 best output and manually annotated the input as follows:

1. Correct: both the syntax/semantic linking of the arguments and the lexicalisation of the event and of its arguments is correct. Some examples are shown in Table 7.
2. Incorrect Linking: the lexicalisation of the event and of its arguments is correct but the syntax/semantic linking of the arguments is not. Some examples are shown in Table 8.
3. Incorrect Frame or Lexicalisation: the lexicalisation or the frame chosen for the event is incorrect. Some examples are shown in Table 9.

29% of the ouput were found to be correct , 17% to have incorrect linking and 54% to lack a correct frame. Manual examination of the results indicates that often, the correct results are found but are not in the 5 best list. We are currently exploring various directions for improving these first results.

One first possibility is to improve ranking by acquiring a more fine grained probabilistic model which in addition to the probabilities presented in Section 2.5, also takes into account e.g., the probability of a dependency given not only a role ($P(d|r)$) but also an event or an event class ($P(d|r,e)$. These two probability distributions could be combined using linear interpolation ($\lambda_1 P_1(d|r) + \lambda_2 P_2(d|r,e)$) for instance. A further interesting avenue for further research would be to use a backoff

Example 1

```
SubClassOf(:Radioactive-Isotope :Entity)
SubClassOf(:Cancer :Entity)
SubClassOf(:Radioactive-Treatment
 ObjectIntersectionOf(:Event
  ObjectSomeValuesFrom(:instrument :Radioactive-Isotope)
  ObjectSomeValuesFrom(:object :Cancer)))
```

Generated Sentence Cancer is treated with radioactive isotope.

Example 2

```
SubClassOf(:Sucrose-Hydrogen-ion-Cotransporter :Entity)
SubClassOf(:Plant-Cell :Entity)
SubClassOf(:Sucrose :Entity)
SubClassOf(:Cotransport-of-sucrose-and-hydrogen-ion
 ObjectIntersectionOf(:Event
  ObjectSomeValuesFrom(:agent :Sucrose-Hydrogen-
                              ion-Cotransporter)
  ObjectSomeValuesFrom(:base :Plant-Cell)
  ObjectSomeValuesFrom(:object :Sucrose)))
```

Generated Sentence Sucrose hydrogen ion cotransporter transports sucrose in plant cells.

Table 7. Correct Examples

Example 1

```
SubClassOf(:Earthworm :Entity)
SubClassOf(:Mucus :Entity)
SubClassOf(:Alimentary-Canal :Entity)
SubClassOf(:Secretion
 ObjectIntersectionOf(:Event
  ObjectSomeValuesFrom(:object :Mucus)
  ObjectSomeValuesFrom(:base :Earthworm)
  ObjectSomeValuesFrom(:site :Alimentary-Canal)))
```

Generated Sentence Alimentary canal secretes earthworm in mucus.

Example 2

```
SubClassOf(:Food-Vacuole :Entity)
SubClassOf(:Solid-Substance :Entity)
SubClassOf(:Confine
 ObjectIntersectionOf(:Event
  ObjectSomeValuesFrom(:base :Food-Vacuole)
  ObjectSomeValuesFrom(:object :Solid-Substance)))
```

Generated Sentence Food vacuole is confined to solid substance.

Table 8. Incorrect Linking Examples

Example 1

```
SubClassOf(:Water-Molecule :Entity)
SubClassOf(:Cellulose :Entity)
SubClassOf(:Cellulase :Entity)
SubClassOf(:Monomer :Entity)
SubClassOf(:Cellulose-digestion
 ObjectIntersectionOf(:Event
  ObjectSomeValuesFrom(:object :Cellulose)
  ObjectSomeValuesFrom(:raw-material :Water-Molecule)
  ObjectSomeValuesFrom(:result :Monomer)
  ObjectSomeValuesFrom(:agent :Cellulase)))
```

Generated Sentence In cellulose, water molecules were digested with cellulase before monomer.

Example 2

```
SubClassOf(:Protein :Entity)
SubClassOf(:Kinetochore-Microtubule :Entity)
SubClassOf(:Divide
 ObjectIntersectionOf(:Event
  ObjectSomeValuesFrom(:site :Kinetochore-Microtubule)
  ObjectSomeValuesFrom(:object :Protein)))
```

Generated Sentence Kinetochore microtubules share protein.

Table 9. Incorrect Frame Examples

approach with several levels of specificity of probabilities following the approach presented in (Swier and Stevenson, 2004) for unsupervised semantic role labelling.

An alternative track consists in investigating vector based approaches and to measure the similarity of a syntactic frame and an event representations by aligning the arguments in the syntactic frame with the arguments in the event representations and computing the similarity of the aligned arguments. Following (Cheung and Penn, 2014), this problem could be solved as a maximum-weight bipartite graph matching problem and similarity could compound both similarity between words and similarity between the slot fillers of syntactic and semantic roles.

Finally, we plan to investigate whether a more intensive use of the lexical (synonymy, hyperonymy) and KB relations (SubClass) could help either approach. These relations could be used in different phases of our approach to generalise over specific facts thereby increasing the frequency counts and reducing the data sparsity.

4 CONCLUSION

We have presented an approach for verbalising biological event representations which differs from previous work in that (i) it is unsupervised and (ii) it focuses on n-ary relations and on the issue of how to automatically map natural language and KB arguments. A first evaluation gives encouraging results but also shows that the current approach has limitations. We are currently exploring two main directions for improvements. On the one hand, we are investigating whether a more sophisticated probabilistic model could help improve results. On the other hand, we are looking at an alternative, vector-based approach.

ACKNOWLEDGEMENTS

This work has been partially funded by the ANR Project WebNLG (Natural Language Generation for the Semantic Web). We would like to thank Vinay Chaudhri for fruitful discussion concerning the KB Bio 101 knowledge base and the Synalp group for valuable feedback and comments during the Thursday Lunch Seminars.

REFERENCES

Aguado, G., Bañón, A., Bateman, J., Bernardos, S., Fernández, M., Gómez-Pérez, A., Nieto, E., Olalla, A., Plaza, R., and Sánchez, A. (1998). ONTOGENERATION: Reusing Domain and Linguistic Ontologies for Spanish Text Generation. In *Workshop on Applications of Ontologies and Problem Solving Methods, ECAI*, volume 98.

Angeli, G., Liang, P., and Klein, D. (2010). A Simple Domain-independent Probabilistic Approach to Generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–512. Association for Computational Linguistics.

Banik, E., Gardent, C., and Kow, E. (2013). The KBGen Challenge. In *the 14th European Workshop on Natural Language Generation (ENLG)*, pages 94–97.

Belz, A. (2008). Automatic Generation of Weather Forecast Texts using Comprehensive Probabilistic Generation-Space Models. *Natural Language Engineering*, **14**(4), 431–455.

Bontcheva, K. and Wilks., Y. (2004). Automatic Report Generation from Ontologies: the MIAKT Approach. In *Ninth International Conference on Applications of Natural Language to Information Systems (NLDB'2004)*. Lecture Notes in Computer Science 3136, Springer, Manchester, UK.

Chaudhri, V. K., Wessel, M. A., and Heymans, S. (2013). KB_Bio_101: A Challenge for OWL Reasoners. In *ORE*, pages 114–120. Citeseer.

Chen, D. L. and Mooney, R. J. (2008). Learning to Sportscast: A Test of Grounded Language Acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135. ACM.

Cheung, J. C. K. and Penn, G. (2014). Unsupervised Sentence Enhancement for Automatic Summarization. pages 775–786.

Clark, P., Harrison, P., Jenkins, T., Thompson, J., Wojcik, R., *et al.* (2005). Acquiring and Using World Knowledge using a Restricted Subset of English. FLAIRS.

Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Pao, C., Rudnicky, A., and Shriberg, E. (1994). Expanding the Scope of the ATIS Task: The ATIS-3 Corpus. In *Proceedings of the workshop on Human Language Technology*, pages 43–48. Association for Computational Linguistics.

Fuchs, N., Kaljurand, K., and Kuhn, T. (2008). Attempto Controlled English for Knowledge Representation. *Reasoning Web*, pages 104–124.

Galanis, D., Karakatsiotis, G., Lampouras, G., and Androutsopoulos, I. (2009). An Open-Source Natural Language Generator for OWL Ontologies and its use in Protégé and Second Life. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 17–20. Association for Computational Linguistics.

Genesereth, M. R. and Nilsson, N. J. (1987). Logical Foundations of Artificial. *Intelligence. Morgan Kaufmann*.

Gunning, D., Chaudhri, V. K., Clark, P. E., Barker, K., Chaw, S.-Y., Greaves, M., Grosof, B., Leung, A., McDonald, D. D., Mishra, S., *et al.* (2010). Project Halo UpdateProgress Toward Digital Aristotle. *AI Magazine*, **31**(3), 33–58.

Kaljurand, K. and Fuchs, N. (2007). Verbalizing OWL in Attempto Controlled English. *Proceedings of OWLED07*.

Konstas, I. and Lapata, M. (2012a). Concept-to-Text Generation via Discriminative Reranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 369–378. Association for Computational Linguistics.

Konstas, I. and Lapata, M. (2012b). Unsupervised Concept-to-Text Generation with Hypergraphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–761. Association for Computational Linguistics.

Liang, P., Jordan, M. I., and Klein, D. (2009). Learning Semantic Correspondences with Less Supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 91–99. Association for Computational Linguistics.

Martin, P. (2002). Knowledge Representation in CGLF, CGIF, KIF, Frame-CG and Formalized-English. *Conceptual Structures: Integration and Interfaces*, pages 77–91.

Mihil, C., Ohta, T., Pyysalo, S., and Ananiadou, S. (2013). Biocause: Annotating and Analysing Causality in the Biomedical Domain. *BMC BioInformatics*.

Mittal, V., Carenini, G., and Moore, J. (1994). Generating Patient Specific Explanations in Migraine. In *Proceedings of the eighteenth annual symposium on computer*

applications in medical care. McGraw-Hill Inc.

Paris, C. (1988). Tailoring Object Descriptions to a User's Level of Expertise. *Computational Linguistics*, **14**(3), 64–78.

Pease, A. and Li, J. (2010). Controlled English to Logic Translation. *Theory and Applications of Ontology: Computer Applications*, pages 245–258.

Pool, J. (2006). Can Controlled Languages Scale to the Web? In *International Workshop on Controlled Language Applications 5*.

Pyysalo, S., Ginter, F., Heimonen, J., Bjrne, J., Boberg, J., Jrvinen, J., and Salakoski, T. (2007). Bioinfer: A Corpus for Information Extraction in the Biomedical Domain. *BMC Bioinformatics*.

Reece, J., Urry, L. A., Meyers, N., Cain, M. L., Wasserman, S. A., Minorsky, P. V., Jackson, R. B., and Cooke, B. N. (2011). *Campbell Biology*. Pearson Higher Education AU.

Reiter, E., Robertson, R., and Osman, L. (2003). Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence*, **144**(1), 41–58.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., *et al.* (2007). The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nature biotechnology*, **25**(11), 1251–1255.

Swier, R. S. and Stevenson, S. (2004). Unsupervised Semantic Role Labelling. In *Proceedings of EMNLP*, volume 95, page 102.

Thompson, P., Iqbal, S. A., McNaught, J., and Ananiadou, S. (2009). Construction of an Annotated Corpus to Support Biomedical Information Extraction. *BMC Bioinformatics*.

Wilcock, G. (2003). Talking OWLs: Towards an Ontology Verbalizer. *Human Language Technology for the Semantic Web and Web Services, ISWC*, **3**, 109–112.

Williams, S. and Power, R. (2010). Grouping Axioms for More Coherent Ontology Descriptions. In *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010)*, pages 197–202, Dublin.

Wong, Y. W. and Mooney, R. J. (2007). Generation by Inverting a Semantic Parser that Uses Statistical Machine Translation. In *HLT-NAACL*, pages 172–179.