

# Extracting Biological Processes with Global Constraints

## Author 1

XYZ Company  
111 Anywhere Street  
Mytown, NY 10000, USA  
author1@xyz.org

## Author 2

ABC University  
900 Main Street  
Ourcity, PQ, Canada A1A 1T2  
author2@abc.ca

## Abstract

Reasoning over processes is fundamental for language understanding applications such as Question Answering. In this paper we propose a method for extracting relations between events in a process. We annotate 150 paragraphs describing biological processes and show that by taking advantage of the global structure of a process we can substantially improve performance. In addition, we release our data set.

## 1 Introduction

Processes describe complicated phenomena that involve a series of events and multiple participants. Automatically extracting the process structure is necessary for text understanding applications, which require reasoning over process events. Consider, for example, the paragraph in Figure 1, which describes the biological process of ATP synthesis. A human reading this paragraph can create a mental model that allows her to answer questions such as:

1. *How do  $H^+$  ions contribute to the production of ATP?*
2. *What causes the rotor to spin?*
3. *In ATP synthesis, what happens if the rotor fails to spin?*

All these questions depend on extraction of the process structure and reasoning over the causal and temporal relations between the process events. Question answering systems that rely on bag-of-words representations will fail to correctly answer such questions.

Process extraction is related to two recent lines of works in Information Extraction – event extraction and timeline construction. The BioNLP 2009 and 2011 shared tasks (Kim et al., 2009; Kim et al., 2011) led to increasing interest in biomedical event extraction (Poon and Vanderwende, 2010; Miwa et al., 2010; Riedel and McCallum, 2011; McClosky et al., 2011; Björne et al., 2011), where given a single sentence annotated with protein mentions, events are identified and relations between events and proteins are extracted. In this shared task participants were asked to consider nine event types that are relevant for proteins (such as *Phosphorylation* and *Transcription*). Processes, on the other hand, contain many events and are centered around the various relations between the event mentions. Thus, process descriptions usually span multiple sentences, and must handle both an open-ended set of event types as well as a rich set of event-event relations.

Timeline construction involves identifying temporal relations between a collection of events (Chambers and Jurafsky, 2008; Yoshikawa et al., 2009; Denis and Muller, 2011; Do et al., 2012; McClosky and Manning, 2012), and is thus related to process extraction as both focus on event-event relations that span multiple sentences. However, fully capturing process structure requires handling a rich set of relations such as CAUSES and SUPEREVENT (see Section 3), which are often not addressed in timeline construction. Moreover, processes exhibit particular properties that do not hold generally in temporal ordering. For example, in processes all events are somehow related to one another, a property that can be exploited for improving extraction.

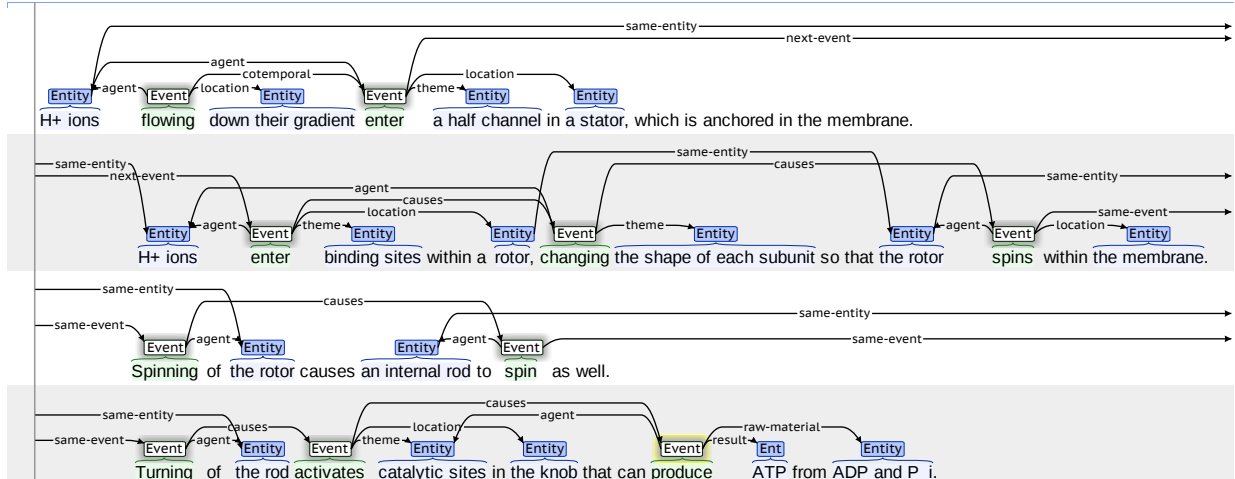


Figure 1: An annotation of the ATP synthesis process

In this paper, we present the task of process extraction and describe methods for extracting relations between process events. Our method works over multiple sentences and extracts a rich set of event-event relations, where the set of possible event types is open ended. Process structure is characterized by global properties that can be utilized during process extraction. For example, most processes exhibit a “chain-like” structure corresponding to process progression over time, and all process events are connected to one another, as previously noted. We will show that by incorporating global properties into our model and performing joint inference over the extracted relations we can significantly improve process quality. Our empirical experiments are performed over a novel data set of 150 process descriptions from the textbook “Biology” (Cambell and Reece, 2005) that were annotated by trained biologists. We note however that our method does not utilize any domain-specific knowledge and thus can be easily applied to domains other than Biology.

To conclude, this paper presents the following three contributions:

1. We define the task of process extraction and characterize the structural properties of processes.
2. We show that by modeling structural properties we can significantly improve the quality of extracted processes comparing to baselines.
3. We publicly release a novel data set of 150 fully

annotated biological processes.

## 2 Related Work

BioNLP work

Timeline construction work.

Scripts work - Chambers, Poon 2013.

Work that uses global constraints with ILP or dual decomposition or whatever.

## 3 Process Definition and Data Set

A process is defined by a series of events, where each event involves some participants, and also by relations between the events (temporal, causal, etc.).

**Notation and definition of a process.** [think whether all of the notation is needed]

A process  $\mathcal{P} = (V, E)$  is a graph with labeled edges (maybe the sentence  $\mathbf{x}$  is also part of it?), where the nodes  $V = \{1, \dots, |V|\}$  are event mentions and edges represent event-event relations. Given a paragraph  $\mathbf{x} = \{x_1, \dots, x_{|x|}\}$ , we define  $x_{i:j}$  to be a span of words  $\{x_i, x_{i+1}, \dots, x_j\}$ . An event mention  $v$  is defined by an event trigger  $t_v$ , which is some span of words  $x_{i:j}$  and by a set of arguments  $A_v$ , where each argument  $a_v \in A_v$  is again a span of words and a semantic role label  $a_l$  taken from a set  $\mathcal{L}$ . an event-event relation is a labeled edge  $(u, v, r)$  where  $r \in \mathcal{R}$  is a closed set of relations<sup>1</sup>.

<sup>1</sup>Our process annotation also contains coreference relations between arguments but we omit this since it is not very relevant.

For example, the first sentence of Figure 1 contains two event mentions, where  $t_1 = \text{flowing}$  and  $t_2 = \dots$ . An example for the notation. [example that will make the notation complete and clear]

In general, one can think of process extraction as comprising of two steps - the first is standard semantic role labeling which comprises trigger identification, argument identification and role labeling, and the second is extracting a rich set of event-event relations. We focus on the second task - basically we get as input the set of event mention triggers  $\mathcal{T}$  and extract the  $E$ . Performing the two tasks jointly is challenging direction for future work.

Next we will describe the full set of semantic role labels  $\mathcal{L}$ , and more importantly the set of event-event relations  $\mathcal{R}$ .

The set of semantic role labels  $\mathcal{L}$  contains standard labels such as *agent*, *theme*, *origin*, *destination* and *location*. In addition we have two semantic role labels that are relevant for the biological text domain - *result* and *raw-material*, which correspond to arguments that are the result of the event and materials used during the event. The last sentence in Figure 1 gives an example for these two labels.

Describe the set of relations  $\mathcal{R}$ : (a) NextEvent - directed relation (b) Causes - directed relation (c) Enables - directed relation (d) SuperEvent - directed relation (e) Coterporal - undirected relation (f) SameEvent - event coreference. (g) None. Differences from previous work: (1) we have not only the temporal relation but also "cause" and "enable" that are important in our domain. (2) We have SuperEvent that most work on temporal ordering did not deal with (we should write who did deal with it) (3) We add coreference as just one of the event-event relations. This allows us to add constraints between cored and other relations in our global formulation. Write something about that we believe this is a succinct and good set of relation for process representation.

**Properties of processes** Naturally there are various properties to coherent processes (1) in the semantic role labeling part - different event triggers don't overlap, different argument of the same event don't overlap (this was used by Toutanova and Haghighi, 2006) [we have to think what to mention depending on whether we show results of the full pipeline] (2) in the event-event relations - all events

are somehow linked to one another. Also in general the events are "chain-like" that is most events are related to one or two other events but not more than that (see Table that shows distribution of degrees of event mentions). In Section ... we will show that by using these global properties of processes and more we can improve performance. (3) Some triads are not possible etc. (4) more? [In general I think it is interesting and relevant to discuss here the global properties of processes even if we eventually don't use all of them in our final model - this is simply since the local model does a good enough job but still identifying these properties is of interest in this paper].

**Data set.** We annotated 150 processes using the described scheme. Table ... describes the statistics of this data set. Stats in the data set: average number of tokens, average number of events mentions, average number of relations, more? We briefly describe the annotation procedure in Section ?? . Inter-annotator agreement is...

Next we will describe our method that given an annotation of event triggers  $\mathcal{T}$  extracts all the event-event relations. We first describe a local pairwise model that classifies each pair of events independently of others and then show our full model that uses global properties of structure.

## 4 Joint Model for Process Extraction

Our task is given a paragraph  $x$  and a set of event mention triggers  $\mathcal{T}$ , to extract all event-event relations  $E$ . Similar to () our model consists of two parts. In section 4.1, we use a local pairwise classifier that considers each pair of triggers, and then in Section 4.2 we perform joint inference over the set of relations using global constraints.

### 4.1 Local pairwise classifier

Our local classifier is a function  $f : \mathcal{T} \times \mathcal{T} \rightarrow \mathcal{R}$ . As a baseline we combine features from previous work (Chambers and Jurafsky, 2008; Do et al., 2012). However, since our set of relations is different we also add some new features more relevant for our task. However, we do not use biological dictionaries as has been done in BioNLP.

Description of baseline features according to categories. Either short or longer depending on impor-

tance. Mention that features that were in previous work but we did not use are simply those that didn't help on the dev set. Have a table with list of features and from what paper we took them.

Description of novel features. We have some that are for SuperEvent (first and nominalization), for coreference (use of determiners), use of dependency paths between the triggers, and the clustering. Maybe about the clustering we can talk a bit more - one of the problems in our scenario is that we have very little training data so it is important for us to use the information we have so we clustered time and cause words to share statistics. I think about this we can talk. I think we should have a short experiment with ablations on the new features to see how much they hurt the local pairwise classifier performance.

classifier - say what classifier we used.

## 4.2 Global Constraints

[I think we should have a short experiment with soft constraints on the degree of nodes, I think this will add some substance even if our intuition is that it might not work I don't think it is a lot of work]

Motivation paragraph - naturally there are cases where local decision can lead to global structures that don't make sense. Give examples - one for something that is a hard constraint and something for soft constraints. Maybe we should have a figure with examples for bad local predictions - connectivity and triads.

Define the notations for formulating the objective function and formulate the objective function (variables are indicators  $e_{ijr}$ ). Our formulation will probably have in the objective the local model scores and the soft constraints. Our hard constraints are those to make the formulation make sense and the hard constraints. Say that we use log probabilities from the pairwise classifier as our weights in the model.

Then we describe the modeling constraints.

Connectivity - short explanation. and then show the formulation which is a slight variation on (Martins et al., 2009). Refer to the motivating figure.

chain-like - we did not implement this because we didn't think it would help but I think it is easy to formulate and experiment with this. Explain the motivation. Refer to the table that shows that the local classifier is doing already pretty well and say that

later we show whether this helps or not in the process of choosing soft constraints.

Triads - some triangles are not meaningful. To better understand what things are predicted by the model but are not in the gold and vice versa we counted and compared. Table... shows the top K of these. These guided us to engineer constraints that will improve the local classifier

Now we go over each one of the triad constraints we experimented with (even if some got 0 weight at the end) We explain the motivation and show the hard constraint formulation mentioning that turning them into soft is easy. We have to not make this boring so try to explain with examples.

Say something about the number of variables and constraints. Say that we use Gurobi ILP solver. Say that in principle one can use dual decomposition methods but in practice for this work we found ILP was fast enough.

Tuning of the soft constraints parameters - should we talk about this here or in the experimental section - probably in the experimental setting part

## 5 Experimental Evaluation

### 5.1 Experimental setup

**Annotation** Talk a bit about annotation of the data. Talk about the split to train and dev. Explain that the dev was used for feature selection in the local classifier and for tuning the parameters for global constraints. Explain that these parameters were chosen with coordinate ascent. Explain what are the values we tried and what are the values that were chosen - if we want we can have a table for the way performance increased on the dev set and what were the values that were chosen. might not be crucial.

Talk about the baselines (A) always next (B) Simple local (C) full local (D) local with chain structure (E) global model

Talk about evaluation measures. (a) full (b) collapsed. Maybe talk about the double-counting problem and we do nothing about it. We have to decide if to use only micro or also macro.

### 5.2 Results

Have a table with all results and discuss. We can see if interesting to have train/dev/test results. Maybe we can also have a confusion matrix for the final

model to see that there are difficulties distinguishing cause-next-cotemp which are harder to do with global constraints and require more work on local features or more data.

Maybe we can have a table with ablations for the new features we added to see which helps? not sure necessary.

what other tables and figures can we have?

### 5.3 Analysis and Discussion

What other interesting stats we can put? I think it would be good to have some interesting example for something that got corrected and also something that we did not correct. It's always nice to have some manual error analysis for intuition.

### 5.4 Full pipeline

If we have this we can briefly explain about our first step system and show some results. This is good to say we do everything and bad if this really sucks.

## 6 Conclusion

In this paper we presented the task of process extraction and a method for extracting processes. We focused on extracting relations between event triggers. We also release publicly a data set for the scientific community. We have shown that by taking advantage of the global structure of a process we can improve performance.

Future work - adding more constraints - Mengqiu's idea. This may result in inference problems (it does) and so we can try think of smarter inference. There is the problem of very little data and we can think about using data from other domains and do adaptations. We want to do the full pipeline jointly.

## References

- Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2011. Extracting contextualized complex biological events with rich graph-based feature sets. *Computational Intelligence*, 27(4):541–557.
- Neil Cambell and Jane Reece. 2005. *Biology*. Benjamin Cummings.
- Nathanael Chambers and Daniel Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *EMNLP*.
- Pascal Denis and Philippe Muller. 2011. Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition. In *IJCAI*.
- Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *EMNLP-CoNLL*, pages 677–687.
- Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Junichi Tsujii. 2009. Overview of bionlp 09 shared task on event extraction. In *Proceedings of BioNLP*.
- Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, and Junichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of BioNLP*.
- André L. Martins, Noah A. Smith, and Eric P. Xing. 2009. Concise integer linear programming formulations for dependency parsing. In *ACL/IJCNLP*, pages 342–350.
- David McClosky and Christopher D. Manning. 2012. Learning constraints for consistent timeline extraction. In *EMNLP-CoNLL*, pages 873–882.
- David McClosky, Mihai Surdeanu, and Christopher D. Manning. 2011. Event extraction as dependency parsing. In *ACL*, pages 1626–1635.
- Makoto Miwa, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event extraction with complex event classification using rich features. *J. Bioinformatics and Computational Biology*, 8(1).
- Hoifung Poon and Lucy Vanderwende. 2010. Joint inference for knowledge extraction from biomedical literature. In *HLT-NAACL*, pages 813–821.
- Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *Proceedings of the Conference on Empirical methods in natural language processing (EMNLP '11)*.
- Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly identifying temporal relations with markov logic. In *ACL/IJCNLP*.