Temporal Expression Recognition for Cell Cycle Phase Concepts in Biomedical Literature

Negacy D. Hailu, Natalya Panteleyeva and **K. Bretonnel Cohen**
Computational Bioscience Program, University of Colorado Denver
School of Medicine
negacy.hailu@ucdenver.edu, natalya.panteleyeva@ucdenver.edu,
kevin.cohen@gmail.com

Abstract

In this paper, we present a system for recognizing temporal expressions related to cell cycle phase (CCP) concepts in biomedical literature. We identified 11 classes of cell cycle related temporal expressions, for which we made extensions to TIMEX3, arranging them in an ontology derived from the Gene Ontology. We annotated 310 abstracts from PubMed. Annotation guidelines were developed, consistent with existing time-related annotation guidelines for TimeML. Two annotators participated in the annotation. We achieved an inter-annotator agreement of 0.79 for an exact span match and 0.82 for relaxed constraints. Our approach is a hybrid of machine learning to recognize temporal expressions and a rule-based approach to map them to the ontology. We trained a named entity recognizer using Conditional Random Fields (CRF) models. An off-the-shelf implementation of the linear chain CRF model was used. We obtained an F-score of 0.77 for temporal expression recognition. We achieved 0.79 macro-averagee F-score and 0.78 micro-averaged F-score for mapping to the ontology.

1 Introduction

Storing and processing temporal data in biomedical informatics is important, but challenging (Zhou and Hripcsak, 2007; Augusto, 2005). Biomedical data is often intrinsically associated with time. For example, data from electronic medical records are on a clinical timeline (Zhou and Hripcsak, 2007) which links all information on the progress of a patient's status. Temporal reasoning remains a challenge for medical information systems (Combi and Shahar, 1997). Conventionally, dictionaries define time as "The continuous passage of existence in which events pass from a state of potentiality in the future, through the present, to a state of finality in the past" (Editorial Staff, undated). This traditional linear concept of temporality does not adequately capture the cyclical nature of some important biological processes, such as the cell cycle and circadian rhythms. In this paper, we describe a system for the recognition of temporal expressions related to cell cycle phases in biomedical literature. The cell cycle is a phenomenon that a cell goes through during its growth and replication. Its stages are depicted in Figure 1. We treat each phase as a distinct time component and we aim at recognizing expressions that describe them in biomedical literature, then mapping them to an ontology of cell cycle phases and transitions. Specifically, we are interested in recognizing expressions that contain one or more of the concepts shown in Table 1, where the Gene Ontology is taken as definitional of concepts related to phases of the cell cycle.

Recognition of cell cycle phase concepts from text is a non-trivial problem. Some of the ways that they can be mentioned in text, such as *interphase, anaphase,* and *prophase* are relatively unambiguous and can be recognized and mapped to an ontology using regular expressions. However, as is often the case both in general language and in biomedical language, many of the ways in which they can be mentioned are highly ambiguous. For example, *M*, which stands for mitosis, is often a unit of measurement, as in ... *removal of histone HI with 0,6 M NaCl.* (PMID: 6183061) *M* could also be an abbreviation of an author's first name, as in ... *Suzuki S, Nakata M.* (PMID: 23844291) *S*, which refers to S-phase or synthesis phase, could also stand for an author's first name, as well as a protein name, as in ... *Protein S acts as a cofactor for tissue factor pathway inhibitor.* (PMID: 23841464). In addition, the word *synthesis* is in it-

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 10–18,
Baltimore, Maryland USA, June 26-27 2014.

self ambiguous, even in the context of other mentions of cell cycle phases. In the following examples, it refers to something other than a cell cycle phase:

- *. . . histone synthesis by lymphocytes in G0 and G1.* (PMID: 6849885)
- *. . . metaphase-anaphase transition, as a result of fertilization, activation or protein synthesis inhibition.* (PMID: 9552372)

We treated recognition of temporal expressions from literature as a named entity recognition (NER) problem. Many approaches to named entity recognition are based on machine learning techniques. Nadeau and Sekine report that although semi-supervised learning algorithms have been employed in NER challenges, most systems that perform well are built based on supervised learning techniques (Nadeau and Sekine, 2007). Based on this survey report, we used Conditional Random Fields (CRFs) for the recognition phase of our approach. The details of our methods are described in section 4.2.

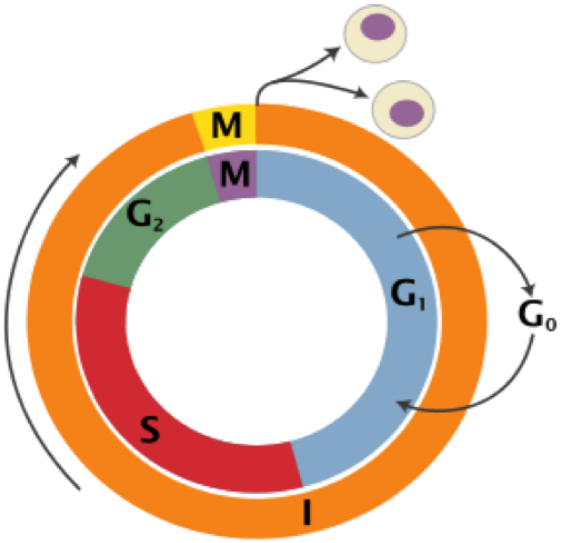

Figure 1: Schematic of the cell cycle. Outer ring: I = Interphase, M = Mitosis; inner ring: M = Mitosis, G1 = Gap 1, G2 = Gap 2, S = Synthesis; not in ring: G0 = Gap 0/Resting [Wikipedia].

2 Motivation

A vast collection of biomedical literature in PubMed/MEDLINE and other biomedical journal repositories is estimated to grow exponentially (Hunter and Cohen, 2006), as shown in Figure 2. Searching for papers specific to a researcher's interest in any domain is difficult. PubMed/MEDLINE allows search using keywords, but until recently did not rank results by document relevance. General-purpose search engines such as Google and Bing rank their results, but are not well-suited for search of specialized information related to genes and small molecules. Building a specialized search engine exclusively to search biomedical literature using genes and small molecules as keywords could be very useful, for instance, for cancer researchers.

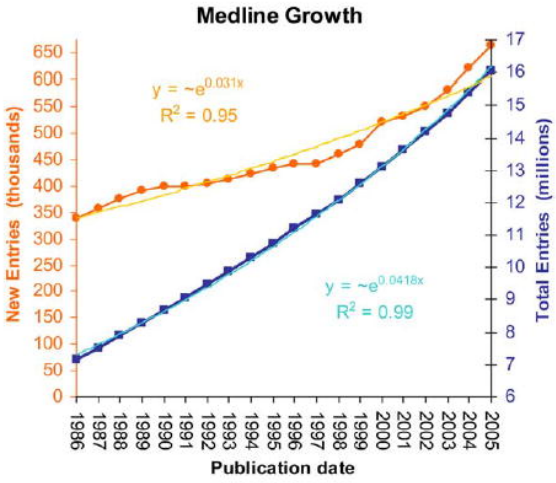

Figure 2: Publication growth rate at Medline (Hunter and Cohen, 2006)

Our long term goal is to build a specialized search engine specific to cancer research. The system will retrieve articles from PubMed/MEDLINE and rank them according to their relevance. The system will utilize gene, protein, and small molecule names as keywords in document search. We are also interested in identifying the phase(s) of the cell cycle during which the gene is expressed. After detecting the active phase(s) of a gene or gene product, the system will link relevant documents to this gene from PubMed/MEDLINE. In this paper we present our first step towards that goal, which is extraction of temporal expressions from biomedical literature. Temporal expressions will be used to identify active phases of genes or gene products.

3 Related Work

Automatic recognition of events and temporal expressions from text has attracted researchers from areas such as computer science and linguistics.

Concept	ID	Activities in each phase	Synonyms
Interphase	GO:0051325	The cell readies itself for meiosis or mitosis and the replication of its DNA occurs.	karyostasis
G0 phase	GO:0044838	Cells enter in response to cues from the cell's environment.	quiescence
G1 phase	GO:0051318	Gap phase	-
S phase	GO:0051320	DNA synthesis takes place.	S-phase, synthesis
G2 phase	GO:0051319	Gap phase	-
Mitosis	GO:0007067	The nucleus of a eukaryotic cell divides	-
Prophase	GO:0051324	Chromosomes condense and the two daughter centrioles and their asters migrate toward the poles of the cell.	-
Metaphase	GO:0051323	Chromosomes become aligned on the equatorial plate of the cell.	-
Anaphase	GO:0051322	The chromosomes separate and migrate towards the poles of the spindle.	-
Telophase	GO:0051326	The chromosomes arrive at the poles of the cell and the division of the cytoplasm starts.	-

Table 1: Cell cycle phase concepts. Definitions from the Gene Ontology.

The results have contributed to the development of diverse natural language processing applications, such as information extraction, information retrieval, question-answering systems, text summarization, etc. TimeML: Robust Specification of Event and Temporal Expressions in Text (Pustejovsky et al., 2003) is a specification language for annotation of events and temporal expressions in human language. TimeML addresses specification issues like time stamping, order of events, reasoning about events, and time expressions.

TempEval is one of the shared challenges included in SemEval (Agirre et al., 2009) as of 2007. It aims at advancing research on processing temporal information. Primarily it focuses on three tasks: event extraction and classification, temporal expression extraction and normalization, and temporal relation extraction (UzZaman et al., 2013). However, this ongoing work on temporal evaluation is based on language data collected from the news. In the clinical domain, (Styler IV et al., Undated; Palmer and Pustejovsky, 2012; Albright et al., 2013) describe the THYME annotation project. The scope and language of temporality related to the cell cycle is different from that of both TempEval and the clinical domain, and supports (and demands) different types of reasoning, specifically related to cyclical time.

Cyclical phenomena are ubiquitous in cancer development and progression. The connection between the cell cycle and cancer is well known (Vermeulen et al., 2003; Kastan and Bartek, 2004; Malumbres and Barbacid, 2009), and the fact that the cell cycle is the main target for cancer regulation, deregulation, and therapy is well established (Vermeulen et al., 2003; Kastan and Bartek, 2004; Malumbres and Barbacid, 2009). Circadian rhythms, rounds of chemotherapy, remissions, and re-occurrences all have a cyclic nature.Circadian rhythms have been investigated in the study of cancer treatment (Sahar and Sassone-Corsi, 2009; Ortiz-Tudela et al., 2013; Lengyel et al., 2009; Kelleher et al., 2014).

From the perspective of cancer research, identifying cell cycle concepts in the literature is crucial to being able to retrieve and explore information related to cyclical biological processes like the cell life cycle. From the natural language processing perspective, the novelty of this work consists in modeling cyclical time. To our knowledge, temporal event recognition grounded in a cyclical model of time has not been previously proposed.

4 Methodology

4.1 Materials

We built a corpus of 360 abstracts, consisting of 70,570 words. The concepts are presented in Table 1. We balanced our corpus by collecting articles from the PubMed/MEDLINE database using the concepts individually as keywords. We used

the PubMed/MEDLINE[1] and BioMedLib search engines[2], two keyword-based search engines built on top of MEDLINE, for this purpose. The following keywords were used to collect the abstracts from PubMed and BioMedLib:

- interphase, G0, G0 phase, G1, G1 phase, synthesis, S phase, G2, G2 phase
- Mitosis, M phase, prophase, metaphase, anaphase, telophase
- checkpoint

The annotation guidelines addressed the following issues:

- The goal of the project: the goal of the annotation project was to develop a highly annotated corpus specific to CCP concepts, which will be used for automatic recognition and classification.
- Specification of each tag: this is shown in Figure 3.
- Tool used to annotate the project: We used Knowtator (Ogren, 2006), a text annotation tool built on top of the Protégé knowledge representation system.

Modeling the phenomenon was the first step in understanding the annotation process (Pustejovsky and Stubbs, 2012). We modeled our corpus as a triple, Model = <T, R, I>, as shown below:

- Model = <T, R, I> where T = terms, R = relation between the terms, and I = interaction
- T = {Named Entity, time expression, not time expression}
- R = {Named_Entity ::= TIMEXCCP | not TIMEXCCP}
- I = {TIMEXCCP = list of concepts from Table 1 or checkpoints. Examples of checkpoints are G1/G2 phase, S/G2 phase, etc.

TimeML is a specification for annotating human language in text (Pustejovsky et al., 2003). TIMEX3 is defined in TimeML as a tag for capturing dates, times, durations, and sets of dates and times. In our work we extended TIMEX3. We employ a single tag set called TIMEXCCP, where the naming is intended to be consistent with existing time-related tag sets. Figure 3 shows the attributes and functions of the tag TIMEXCCP, as well as examples of usage.

Attribute	Function	Example
Value	Interphase, G0, G1, S, G2, M, prophase, metaphase, anaphase, telophase, checkpoint.	The <TIMEXCCP value ="G1 checkpoint"> G1 checkpoint </TIMEXCCP> control mechanism ensures that everything is ready ...
Modifier	Handles temporal modifiers such as early, mid, late ...	Apoptosis induction and <TIMEXCCP modifier = "early" value = "G2/M"> early G2/M </TIMEXCCP>arrest of ...
Set	A boolean value if the time expression is a set or not.	
Comments	Comments by the annotators.	

Figure 3: Attributes and functions of the TIMEXCCP tag.

Two annotators with training in the domain performed the annotation. Inter-annotator agreement was calculated as F-measure, following (Hripcsak and Rothschild, 2005). Inter-annotator agreement was 0.79 for an exact-span match and 0.82 for relaxed matching. The constraints, which are values of the attributes, were not considered while computing IAA for the latter case.

The annotation effort developed through several iterations, applying the annotation development cycle introduced by Pustejovsky and Stubbs (Pustejovsky and Stubbs, 2012). This methodology is depicted in Figure 4. It is called the MATTER cycle, which stands for Model, Annotation, Train, Test, Evaluation, Revise. The advantage of this methodology is that it allows us to discover hidden specifications and refine them during the MATTER cycle.

4.2 Methods

We are particularly interested in recognizing and classifying temporal expressions in the literature. For example, in the following sentence, taken from Wikipedia, the recognition task is to recognize the blue boxes as shown below and classify them. The mapping task is to categorize the recognized temporal expressions into the concepts shown in Table 1.

```
"Microhomology-mediated end
joining (MMEJ) uses a Ku
protein and DNA-PK independent
repair mechanism, and
```

[1]http://www.ncbi.nlm.nih.gov/pubmed/
[2]http://bmlsearch.com/

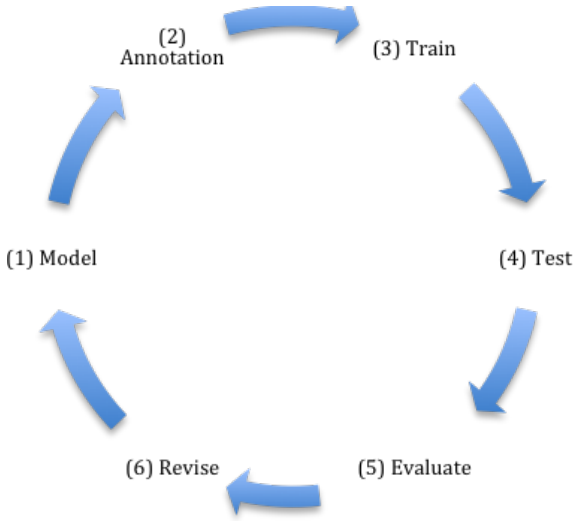

Figure 4: The MATTER cycle (Pustejovsky and Stubbs, 2012)

```
repair occurs during the
S phase of the cell cycle,
as oposed to the G0/G1 and
early S phases in NHEJ and
late S to G2 phases in HR."
```

... the S phase of the cell cycle, as opposed to the G0/G1 and early S phases in NHEJ and late S to G2 phase in HR.

In this example, there are four temporal expressions: *S*, *G0/G1*, *early S*, and *late S to G2*. The expression "S" is of the type S-phase or synthesis phase according to the conceptual ontology in Table 1. The expression "G0/G1" can be classified as G0 and G1. Similarly, the expression "late S to G2" can be of type S and G2.

Our approach is a hybrid of machine learning and rule-based techniques. The machine learning technique, which we refer to as the first layer, is applied for temporal expression recognition. In this layer, CRFs are trained to learn to recognize the expressions from the list of features which is shown below.

1. Word-level features:
   - Is the word in uppercase?
   - Is the first character of the word in uppercase?
   - Words themselves are also treated as features.
   - Length of the word.
2. Punctuation-related features:
   - Does the word contain at least one of the most common punctuation marks?
3. Digit-related features:
   - Is the word a digit?
   - Does the word contain a digit?
4. Does the word contain either of the following: *phase, arrest, entry*? These words typically come before or after the cell cycle concepts. For example, *early mitosis, G0 phase*.
5. Part-of-Speech tagging: Window size of 2 before and after the word.
6. Presence of concept modifiers before the word. Modifiers include: *early, mid, late, early-mid*.

Conditional Random Fields (CRFs) are one of the probabilistic graphical model sequence tagging techniques. They are understood as a sequential version of Maximum Entropy Models (Klinger and Tomanek, 2007). One advantage of CRFs over other probabilistic models like Hidden Markov Models and Maximum Entropy Models for complex systems is their support for features interacting with one another. The linear chain CRF representation is shown in Figure 5.

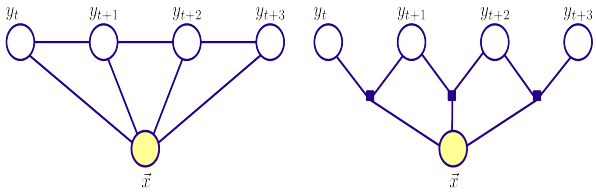

Figure 5: A linear chain Conditional Random Field representation (Klinger and Tomanek, 2007).

In this representation, $\vec{x}$ is a vector of observations, also known as features in machine learning, and the y_t's are states or labels. In this linear chain model, a given state is dependent on its previous, current, and next states. It is also influenced by the observations for that state. This argument can be formulated as Equation 1. Accordingly, state prediction will be an optimization of Equation 1. $\psi_c(\vec{x}, \vec{y})$ are the factor matrices of

the maximal cliques read from the factor graph in Figure 5 (Klinger and Tomanek, 2007).

$$P(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \prod_{c \epsilon C} \psi_c(\vec{x}, \vec{y}) \quad (1)$$

We used the *IOB* format, which is the most common method of representation for sequence tagging. In this format, *I* stands for the inside, *O* is the outside, and *B* is the beginning of a temporal expression. Table 2 shows an example of *IOB* labeling for the phrase ... *late S to G2 phase in HR*.

token	tag
...	...
late	B_TIMEXCCP
S	I_TIMEXCCP
to	I_TIMEXCCP
G2	I_TIMEXCCP
phase	O
in	O
HR	O
.	O

Table 2: IOB format representation of a segment of a sentence.

The rule-based system is keyword-based. The rules match simple cell cycle phase concepts. For example, the phrase *early S phase* is classified as synthesis, since there is *S* in it. The expression *G0/G1 phase* is classified as a *G0/G1* checkpoint.

5 Experimental setup

We split our dataset of more than 70K tokens into 80% training and 20% test sets. We used 5-fold cross validation to balance the distribution of the dataset. The number of positive instances for the 5 runs is shown in Figure 6. The expressions S and synthesis are displayed separately, despite their identical meaning, to allow for more granular evaluation of performance. The same rationale applies to displaying M and mitosis separately.

The ratio of the individual concepts that we have in the 5 runs is balanced, as shown in Figure 6. However, the training dataset is skewed, since there are almost 98% negative labels, with the remaining small portion as positive labels. Among the approximately 10K test tokens, 180 of them are labeled as positive TIMEXCCP, but the others are negative, i.e. they have the label *O*. A positive TIMEXCCP in this case could be B_TIMEXCCP or I_TIMEXCCP—beginning or inside of a temporal expression.,

6 Results

Since the task consisted of two separate steps—temporal expression recognition, and mapping or normalization—in this section, we report our findings independently. Our evaluation metrics are in terms of precision *P*, recall *R*, and *F-measure*. The system achieved precision P = 0.83, recall R = 0.72 and F = 0.77 for recognizing TIMEXCCP in biomedical literature.

The temporal expression mapper, which is a rule-based system, achieved a macro-averaged *P*= 0.90, *R* = 0.70, and *F* = 0.79 and a micro-averaged *P* = 0.86, *R* = 0.71, and *F* = 0.78. The system performance for the individual concepts is shown in Figure 7.

7 Discussion

Some of the false positive predictions were due to human annotation errors.

There were some conditions where the annotators disagreed. For example, ... *early G1 to G2 phase*. This examples addresses two questions that should be explicitly mentioned in the annotation guidelines:

- Does the modifier "early" modify only G1, or both G1 and G2?
- Should there be an attribute for the range of time from G1 to G2 in the annotation guidelines?

Our system achieved good performance on both time expression recognition and mapping of highly ambiguous concepts. In spite of the challenges presented by ambiguity, we obtained 0.85, 0.81, and 0.80 F-measures for recognizing and mapping the concepts *synthesis*, *M*, and *S*, respectively. The most informative features that contribute to this score are the discriminating words before and after a target token. These words are: *phase, arrest*, and *entry*. They are often present before or after CCP concepts. Also, presence of modifiers is a good indication of CCP concepts. For example, in the phase *early S phase*, the modifier *early* is one of the most informative features. However, recognition of complex phrases as in *late S to G2 phase* remained a challenge.

The challenges of complex temporal expressions can be seen from a different perspective.

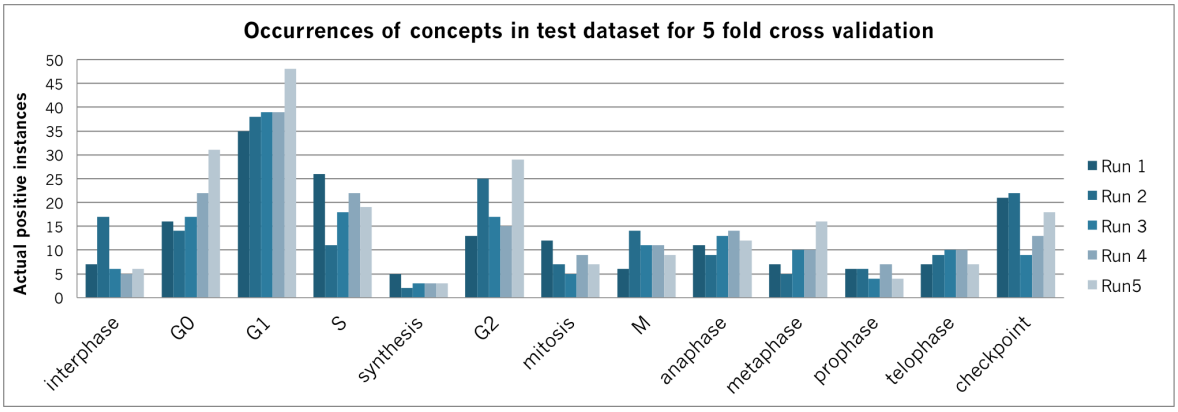

Figure 6: Distribution of concepts in 5 runs.

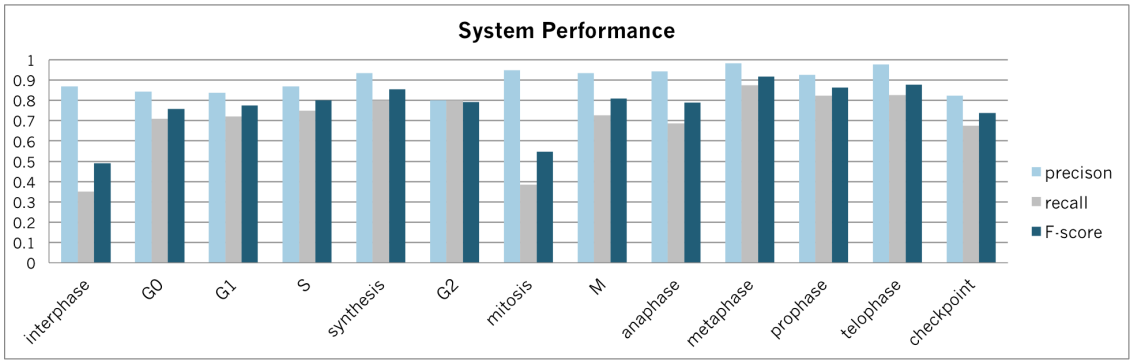

Figure 7: Rule-based classification performance. Average score for 5 runs.

Mostly the system recognizes the individual concepts within a complex phrase, but not the modifiers nor the words like prepositions within the complex phrase. In the example given previously, the system recognizes *S* and *G2* but not the modifier *late*, nor the preposition *to*. These challenges could be tackled by having features that address the modifiers as well as words within two concepts.

We used a naive tokenizer that splits the text into words based on white space. In the future, we would like to test the system with other more sophisticated tokenizers. We kept punctuation marks in temporal expressions, for example, the forward slash in *G0/G1 phase*. Presence of punctuation marks, such as hyphen (-), forward slash (/), comma (,) and single quote (’), within a token is one of our features in training the machine learning algorithm to recognize temporal expressions.

8 Conclusions & Future work

Cell cycle phase concepts are time expressions, and can be annotated in a fashion similar to TimeML. In this work, we annotated a corpus with cell cycle phase information. This corpus can be used to train machine learning algorithms to predict cell cycle phase concepts. The concepts were annotated using the TIMEXCCP tag, an extension of TIMEX3, which has the following attributes: value, modifier, set, and comments. The details are in Figure 3.

We have developed a temporal expression recognizer and classifier based on a hybrid of machine learning and rule-based techniques. We propose a two-tiered architecture to solve temporal expression recognition and mapping for CCP concepts. The first tier recognizes temporal expressions using CRFs. In the second tier, a rule-based system classifies the concepts.

Some of the main future directions for this works are testing the system with the addition of more annotated data. We will focus on how we can capture complex time expressions. This might take us to redefining the annotation guidelines that we have right now.

Acknowledgments

The authors thank Richard Osborne and Scott Cramer for helpful discussion of the significance of this work from a cancer research perspective.

References

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski. 2009. Computational semantic analysis of language: Semeval-2007 and beyond. *Language Resources and Evaluation*, 43(2):97–104.

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930.

Roberta Alfieri, Ivan Merelli, Ettore Mosca, and Luciano Milanesi. 2007. The Cell Cycle DB: a systems biology approach to cell cycle analysis. *Nucleic Acids Research*.

Juan Carlos Augusto. 2005. Temporal reasoning for decision support in medicine. *Artificial Intelligence in Medicine*, 33(1):1–24.

Matteo Brucato, Leon Derczynski, Hector Llorens, Kalina Bontcheva, and Christian S. Jensen. 2013. Recognising and interpreting named temporal expressions. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, *RANLP*, pages 113–121. RANLP 2011 Organising Committee/ACL.

C. Combi and Y. Shahar. 1997. Temporal reasoning and temporal data maintenance in medicine: Issues and challenges. *Comput Biol Med*, 27 (5).

Carlo Combi, Elpida Keravnou-Papailiou, and Yuval Shahar. 2010. *Temporal Information Systems in Medicine*. Springer Publishing Company, Incorporated, 1st edition.

Collins Editorial Staff. undated. *Collins Concise English Dictionary*.

Nicholas Paul Gauthier, Lars Juhl Jensen, Rasmus Wernersson, Sören Brunak, and Thomas S. Jensen. 2009. Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Research*, 9.

Erik Hatcher, Otis Gospodnetic, and Mike McCandless. 2nd revised edition. edition.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Lawrence Hunter and K. Bretonnel Cohen. 2006. Biomedical Language Processing: Perspective What's Beyond PubMed? *Molecular Cell*, 21:589–594.

Michael Kahn. December 1991. Modeling time in medical decision-support programs. *Med Decision Making*, 11(4):249–264.

Michael B. Kastan and Jiri Bartek. 2004. Cell-cycle checkpoints and cancer. *Nature*, 432:316–323.

Fergal C. Kelleher, Aparna Rao, and Anne Maguire. 2014. Circadian molecular clocks and cancer. *Cancer Letters*, 342:9–18.

Roman Klinger and Katrin Tomanek. 2007. Classical Probabilistic Models and Conditional Random Fields. Technical Report TR07-2-013, Department of Computer Science, Dortmund University of Technology, December.

R. Leaman and Gonzalez G. 2008. BANNER: An executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, 13:652–663.

Zsuzsanna Lengyel, Zita Battyáni, György Szekeres, Valér Csernus, and András D. Nagy. 2009. Circadian clocks and tumor biology: what is to learn from human skin biopsies? *Nature Reviews Cancer*, 9:153–166.

Marcos Malumbres and Mariano Barbacid. 2009. Cell Cycle, CDKs and cancer: a changing paradigm. *Nature Reviews Cancer*, 9:153–166.

Inderjeet Mani, Ben Wellner, Marc Verhagen, and James Pustejovsky. 2007. Three approaches to learning TLINKs in TimeML. Technical report, Computer Science Department, Brandeis University. Waltham, USA.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January. Publisher: John Benjamins Publishing Company.

Philip V. Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.

Elisabet Ortiz-Tudela, Ida Iurisci, Jacques Beau, Abdoulaye Karaboue, Thierry Moreau, Maria Angeles Rol, Juan Antonio Madrid, Francis Lévi, and Pasquale F. Innominato. 2013. The circadian rest-activity rhythm, a potential safety pharmacology endpoint of cancer chemotherapy. *International Journal of Cancer*.

Martha Palmer and James Pustejovsky. 2012. 2012 i2b2 temporal relations challenge annotation guidelines.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'REILLY.

James Pustejovsky, Josè Castano, Robert Ingria, Roser Saur, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5)*.

Saurabh Sahar and Paolo Sassone-Corsi. 2009. Metabolism and cancer: the circadian clock connection. *Nature*, 9:886–896.

Yuval Shahar and Carlo Combi. 1999. Editors' foreword: Intelligent temporal information systems in medicine. *J. Intell. Inf. Syst.*, 13(1-2):5–8.

Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, and Guergana Savova. Undated. Temporal annotation in the clinical domain.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Katrien Vermeulen, Dirk R. Van Bockstaele, and Zwi N. Berneman. 2003. The Cell Cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Proliferation*, 36:131–149.

Michael L. Whitield, Gavin Sherlock, Alok J. Saldanha, John I. Murray, Catherine A. Ball, Karen E. Alexander, John C. Matese, Charles M. Perou, Myra M. Hurt, Patrick O. Brown, and David Botstein. 2002. Identication of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, 13.

Li Zhou and George Hripcsak. 2007. Temporal reasoning with medical data - a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40(2):183–202.