Template Filling in a Biomedical Domain

Trends in Bioinformatics

Janos Brauer

February 3rd 2017

Imagine you are an oncology researcher

Your task: Find new clinically proven treatments in literature for a clinician

Read papers that are relevant to the topic

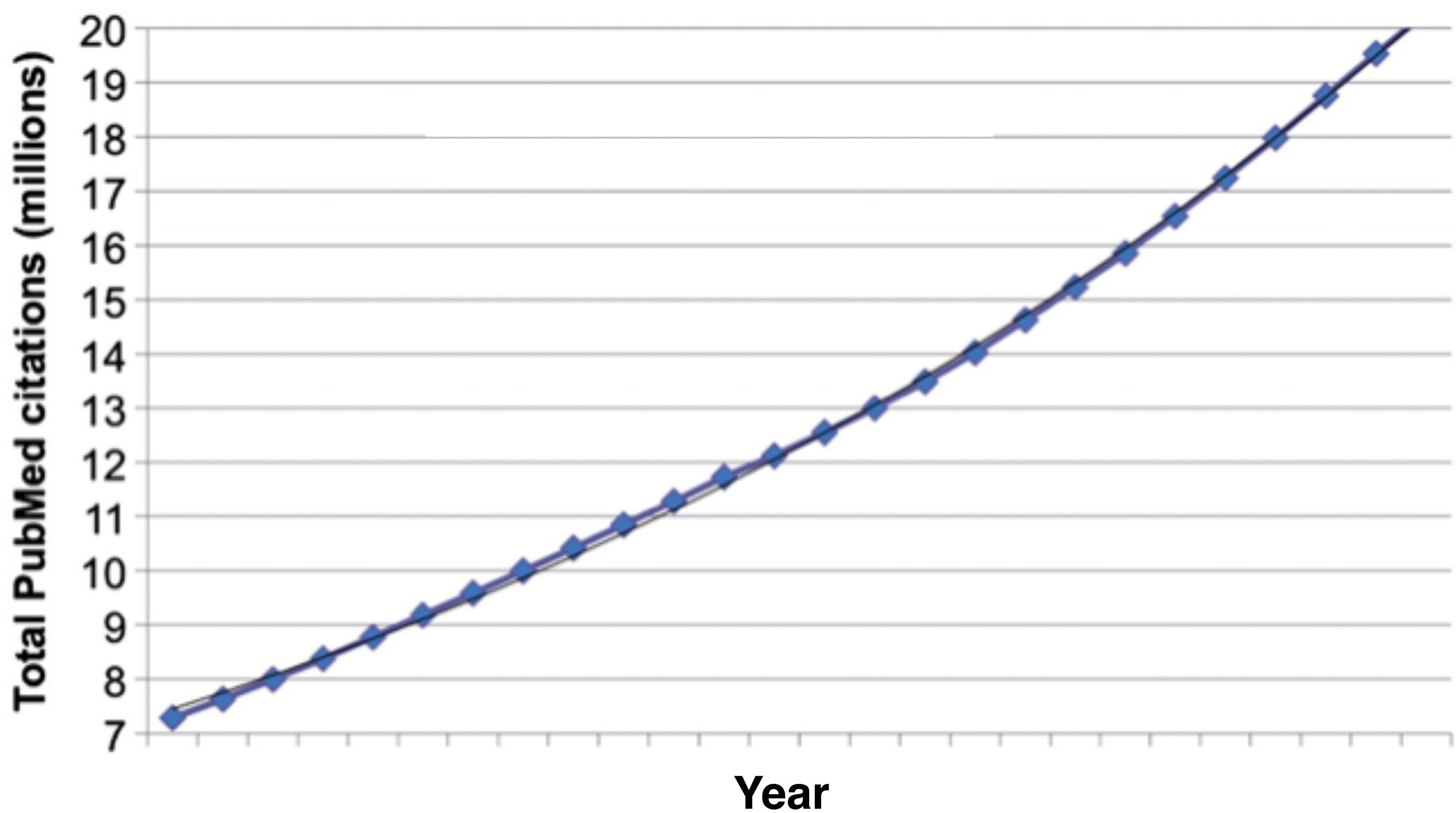
PubMed

Google Scholar

Cochrane

Embase®

Size of Biomedical Literature is growing



Problem

- Literature search for biomedical researcher is becoming more complex

- Reasons:

    - Size of Biomedical Literature is growing exponentially

    - Biomedical domain is becoming more multidisciplinary

→ **Manual literature search is not possible**

Agenda

- **Ansatz DKG**
- Template Filling
- My Approach
- Hands-On Session
- Future Work

DKG

- **D**eutsche **K**rebs**G**esellschaft e. V.
- Non-profit organization
- Section B: medical experts and non-academic professionals active in research, treatment and control of cancer
- Consult clinicians regarding cancer treatments

Approach

Paper

Read Abstract

Researcher @ DKG

Fill Slots

Form

ID	123
Chapter	colon cancer
Type	RCT
Intervention	Surgery
Study Count	140

Form

ID	124
Chapter	Colon Cancer
Type	RCT
Intervention	Laser Therapy
Study Count	100

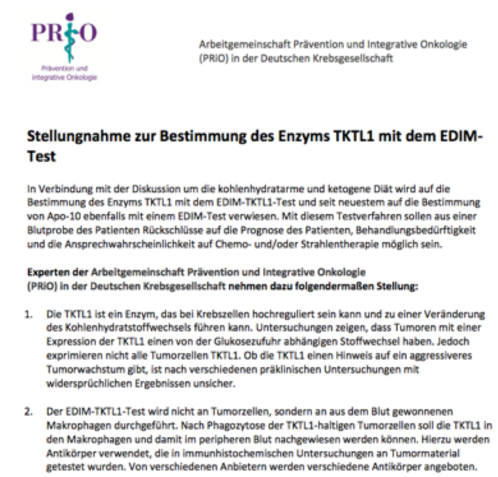
Read

Read

Researcher @ DKG

Create

Summary



Arbeitgemeinschaft Prävention und Integrative Onkologie (PRiO) in der Deutschen Krebsgesellschaft

Stellungnahme zur Bestimmung des Enzyms TKTL1 mit dem EDIM-Test

In Verbindung mit der Diskussion um die kohlenhydratarme und ketogene Diät wird auf die Bestimmung des Enzyms TKTL1 mit dem EDIM-TKTL1-Test und seit neuestem auf die Bestimmung von Apo-10 ebenfalls mit einem EDIM-Test verwiesen. Mit diesem Testverfahren sollen aus einer Blutprobe des Patienten Rückschlüsse auf die Prognose des Patienten, Behandlungsbedürftigkeit und die Ansprechwahrscheinlichkeit auf Chemo- und/oder Strahlentherapie möglich sein.

Experten der Arbeitgemeinschaft Prävention und Integrative Onkologie (PRiO) in der Deutschen Krebsgesellschaft nehmen dazu folgendermaßen Stellung:

1. Die TKTL1 ist ein Enzym, das bei Krebszellen hochreguliert sein kann und zu einer Veränderung des Kohlenhydratstoffwechsels führen kann. Untersuchungen zeigen, dass Tumoren mit einer Expression der TKTL1 einen von der Glukosezufuhr abhängigen Stoffwechsel haben. Jedoch exprimieren nicht alle Tumorzellen TKTL1. Ob die TKTL1 einen Hinweis auf ein aggressiveres Tumorwachstum gibt, ist nach verschiedenen präklinischen Untersuchungen mit widersprüchlichen Ergebnissen unsicher.
2. Der EDIM-TKTL1-Test wird nicht an Tumorzellen, sondern an aus dem Blut gewonnenen Makrophagen durchgeführt. Nach Phagozytose der TKTL1-haltigen Tumorzellen soll die TKTL1 in den Makrophagen und damit im peripheren Blut nachgewiesen werden können. Hierzu werden Antikörper verwendet, die in immunhistochemischen Untersuchungen an Tumormaterial getestet wurden. Von verschiedenen Anbietern werden verschiedene Antikörper angeboten.

Form

ID	[[23926_vongruenigen2007]]			
Kapitel	Vulvakarzinom	Vaginalkarzinom	In situ	
Zuordnung	4_op_vulva	4_op_vagina		
Typ	RCT			
Intervention(en)	Lasertherapie			
Vergleichsintervention(en)	Ultraschalltherapie			
Fallzahl	110			
Ergebnisse		12		
	Schmerzen stärker	VAS 25,1 vs. 20,7; p=0,032		
	Vulvakarzinom			
	Narbenbildung häufiger	P<0.01		
	Rezidivrate gleich			
Bemerkungen				

Positive Aspects

Abstract

OPEN ACCESS Freely available online — PLoS BIOLOGY

A Role for Parasites in Stabilising the Fig-Pollinator Mutualism

Derek W. Dunn[1,2,3], Simon T. Segar[1,2], Jo Ridley[3], Ruth Chan[1], Ross H. Crozier[4], Douglas W. Yu[3], James M. Cook[1,2,5*]

1 Division of Biology, Imperial College London, Ascot, United Kingdom, 2 School of Biological Sciences, University of Reading, Reading, United Kingdom, 3 School of Biological Sciences, University of East Anglia, Norwich, United Kingdom, 4 School of Marine and Tropical Biology, James Cook University, Townsville, Queensland, Australia, 5 Natural Environment Research Council (NERC) Centre for Population Biology, Imperial College London, Ascot, United Kingdom

Mutualisms are interspecific interactions in which both players benefit. Explaining their maintenance is problematic, because cheaters should outcompete cooperative conspecifics, leading to mutualism instability. Monoecious figs (*Ficus*) are pollinated by host-specific wasps (Agaonidae), whose larvae gall ovules in their "fruits" (syconia). Female pollinating wasps oviposit directly into *Ficus* ovules from inside the receptive syconium. Across *Ficus* species, there is a widely documented segregation of pollinator galls in inner ovules and seeds in outer ovules. This pattern suggests that wasps avoid, or are prevented from ovipositing into, outer ovules, and this results in mutualism stability. However, the mechanisms preventing wasps from exploiting outer ovules remain unknown. We report that in *Ficus rubiginosa*, offspring in outer ovules are vulnerable to attack by parasitic wasps that oviposit from outside the syconium. Parasitism risk decreases towards the centre of the syconium, where inner ovules provide enemy-free space for pollinator offspring. We suggest that the resulting gradient in offspring viability is likely to contribute to selection on pollinators to avoid outer ovules, and by forcing wasps to focus on a subset of ovules, reduces their galling rates. This previously unidentified mechanism may therefore contribute to mutualism persistence independent of additional factors that invoke plant defences against pollinator oviposition, or physiological constraints on pollinators that prevent oviposition in all available ovules.

Citation: Dunn DW, Segar ST, Ridley J, Chan R, Crozier RH, et al. (2008) A role for parasites in stabilising the fig-pollinator mutualism. PLoS Biol 6(3): e59. doi:10.1371/journal.pbio.0060059

Introduction

In a biosphere driven by selection at the level of the individual gene [1], explaining the existence of cooperation, such as mutualism, is a major scientific challenge. Mutualisms are interspecific ecological interactions characterised by reciprocal benefits to both partners [2] that usually involve costly investments by each. What factors thus prevent one partner from imposing unsustainable costs onto the other to enable mutualism stability [3–7]? In some mutualisms, the larger, more sessile partner, manipulates the other by directing benefits to cooperative individuals and costs to cheaters [4–7]. However, a general consensus on mutualism persistence has only recently been formulated, and this

from overexploiting figs remain unknown, despite intensive study over four decades.

Within receptive syconia, the lengths of floral styles are highly variable [13,14], and ovipositing pollinators (foundresses) favour flowers with shorter styles for their offspring [15–18]. Style and pedicel lengths of flowers are negatively correlated. Short-styled ovules develop into seeds or galls (when a wasp is present) near the syconium inner cavity, while most long-styled ovules develop into seeds near the outer wall [19,20] (Figure 1). These patterns have been shown to reflect the oviposition preferences of foundresses, and are unlikely to be the result of greater elongation of pedicels containing eggs during syconial maturation, because in receptive syconia, pollinators' eggs are mainly present in short-styled inner ovules [16]. These widespread observations have been

- Relevant information is extracted
- Not the whole paper is read

Form

ID	
Chapter	
Type	
Intervention	
Study Count	

- Unstructured Text is structured
- Table can be inserted into Database

DEMO

Negative Aspects

Long Templates

ID	[[23926_vongruenigen2007]]			
Kapitel	Vulvakarzinom	Vaginalkarzinom	In situ	
Zuordnung	4_op_vulva	4_op_vagina		
Typ	RCT			
Intervention(en)	Lasertherapie			
Vergleichsintervention(en)	Ultraschalltherapie			
Fallzahl	110			
Ergebnisse		12		
	Schmerzen stärker	VAS 25,1 vs. 20,7; p=0,032		
	Vulvakarzinom			
	Narbenbildung häufiger	P<0.01		
	Rezidivrate gleich			
Bemerkungen				

Language Problems

- Problems are also generated by human factors:
  - different languages
    - e.g. English text has to be translated into German
  - synonyms
    - e.g. pegylated liposomal doxorubicin vs. PLD
  - conflicts
    - e.g. two distinct nouns are possible slot fillers

Time & Error Rate

- Time-consuming: up to 30 minutes for one paper
- Very tedious and error-prone approach

Let’s automate it

Agenda

- Ansatz DKG
- **Template Filling**
- My Approach
- Hands-On Session
- Future Work

Recap: Writing an abstract

- Six steps to write a good abstract:

  1. **Motivation**: Why do we care about the problem and results?

  2. **Problem Statement**: What problem are you trying to solve?

  …

- Every abstract has these patterns

Definition

- Abstracts of papers contain common and stereotypical structure called **script**
- A script is represented by a **template**
- A template consists of a fixed set of **slots**
- A Slot is filled with slot-fillers belonging to particular classes according to a **slot-filling rule**
- A slot-filling rule serves as an extraction guideline

Template Filling

Filling slots in predefined templates

Template & Slots

Study Count: 110
Treatment: Laser Therapy
Type: RCT

Flat template

vs.

Study Count: 110
Treatment: Laser Therapy
Type: RCT
Results: [ResultTemplate]

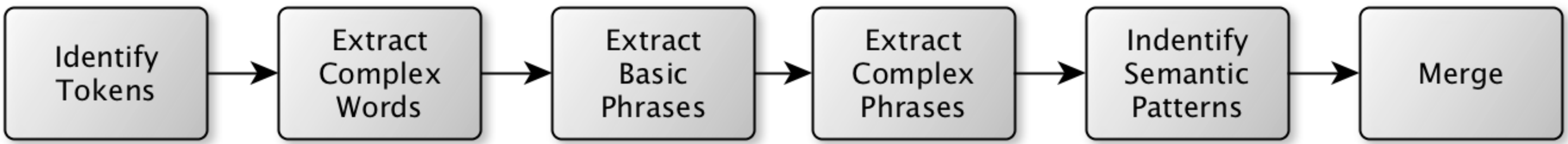
Object-oriented template

- Design of template can either be manual or automatic

- Populated Templates can also be provided

Slot-filling Rules: Finite-state Cascade

Pipeline of Transducers



- Each transducer is a finite-state automata which extracts a specific type of information
- Transducers can contain hand-written regular expressions or grammar rules
- Used to fill object-oriented templates

Agenda

- Ansatz DKG
- Template Filling
- **My Approach**
- Hands-On Session
- Future Work

Input


OPEN ACCESS Freely available online — PLoS BIOLOGY

A Role for Parasites in Stabilising the Fig-Pollinator Mutualism

Derek W. Dunn[1,2,3], Simon T. Segar[1,2], Jo Ridley[3], Ruth Chan[1], Ross H. Crozier[4], Douglas W. Yu[3], James M. Cook[1,2,5*]

1 Division of Biology, Imperial College London, Ascot, United Kingdom, 2 School of Biological Sciences, University of Reading, Reading, United Kingdom, 3 School of Biological Sciences, University of East Anglia, Norwich, United Kingdom, 4 School of Marine and Tropical Biology, James Cook University, Townsville, Queensland, Australia, 5 Natural Environment Research Council (NERC) Centre for Population Biology, Imperial College London, Ascot, United Kingdom

Mutualisms are interspecific interactions in which both players benefit. Explaining their maintenance is problematic, because cheaters should outcompete cooperative conspecifics, leading to mutualism instability. Monoecious figs (*Ficus*) are pollinated by host-specific wasps (Agaonidae), whose larvae gall ovules in their "fruits" (syconia). Female pollinating wasps oviposit directly into *Ficus* ovules from inside the receptive syconium. Across *Ficus* species, there is a widely documented segregation of pollinator galls in inner ovules and seeds in outer ovules. This pattern suggests that wasps avoid, or are prevented from ovipositing into, outer ovules, and this results in mutualism stability. However, the mechanisms preventing wasps from exploiting outer ovules remain unknown. We report that in *Ficus rubiginosa*, offspring in outer ovules are vulnerable to attack by parasitic wasps that oviposit from outside the syconium. Parasitism risk decreases towards the centre of the syconium, where inner ovules provide enemy-free space for pollinator offspring. We suggest that the resulting gradient in offspring viability is likely to contribute to selection on pollinators to avoid outer ovules, and by forcing wasps to focus on a subset of ovules, reduces their galling rates. This previously unidentified mechanism may therefore contribute to mutualism persistence independent of additional factors that invoke plant defences against pollinator oviposition, or physiological constraints on pollinators that prevent oviposition in all available ovules.

Citation: Dunn DW, Segar ST, Ridley J, Chan R, Crozier RH, et al. (2008) A role for parasites in stabilising the fig-pollinator mutualism. PLoS Biol 6(3): e59. doi:10.1371/journal.pbio.0060059

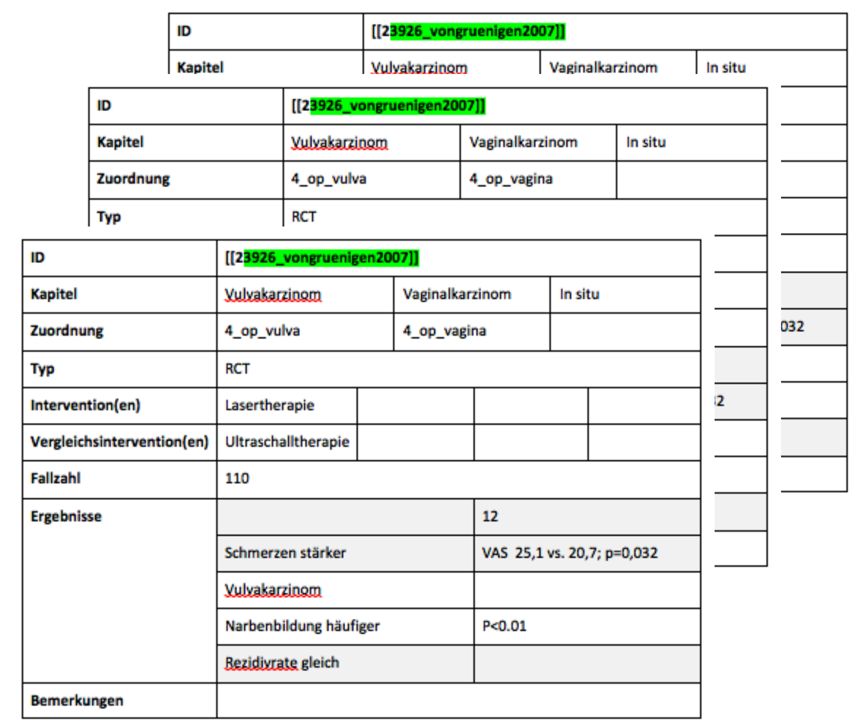
Introduction

In a biosphere driven by selection at the level of the individual gene [1], explaining the existence of cooperation, such as mutualism, is a major scientific challenge. Mutualisms are interspecific ecological interactions characterised by reciprocal benefits to both partners [2] that usually involve costly investments by each. What factors thus prevent one partner from imposing unsustainable costs onto the other to enable mutualism stability [3–7]? In some mutualisms, the larger, more sessile partner, manipulates the other by directing benefits to cooperative individuals and costs to cheaters [4–7]. However, a general consensus on mutualism persistence has only recently been formulated, and this

from overexploiting figs remain unknown, despite intensive study over four decades.

Within receptive syconia, the lengths of floral styles are highly variable [13,14], and ovipositing pollinators (foundresses) favour flowers with shorter styles for their offspring [15–18]. Style and pedicel lengths of flowers are negatively correlated. Short-styled ovules develop into seeds or galls (when a wasp is present) near the syconium inner cavity, while most long-styled ovules develop into seeds near the outer wall [19,20] (Figure 1). These patterns have been shown to reflect the oviposition preferences of foundresses, and are unlikely to be the result of greater elongation of pedicels containing eggs during syconial maturation, because in receptive syconia, pollinators' eggs are mainly present in short-styled inner ovules [16]. These widespread observations have been

Abstracts



ID	[[23926_vongruenigen2007]]			
Kapitel	Vulvakarzinom		Vaginalkarzinom	In situ
Zuordnung	4_op_vulva		4_op_vagina	
Typ	RCT			
Intervention(en)	Lasertherapie			
Vergleichsintervention(en)	Ultraschalltherapie			
Fallzahl	110			
Ergebnisse			12	
	Schmerzen stärker		VAS 25,1 vs. 20,7; p=0,032	
	Vulvakarzinom			
	Narbenbildung häufiger		P<0.01	
	Rezidivrate gleich			
Bemerkungen				

Templates

Input

ID	[[23926_vongruenigen2007]]			
Kapitel	Vulvakarzinom	Vaginalkarzinom	In situ	
Zuordnung	4_op_vulva	4_op_vagina		
Typ	RCT			
Intervention(en)	Lasertherapie			
Vergleichsintervention(en)	Ultraschalltherapie			
Fallzahl	110			
Ergebnisse		12		
	Schmerzen stärker	VAS 25,1 vs. 20,7; p=0,032		
	Vulvakarzinom			
	Narbenbildung häufiger	P<0.01		
	Rezidivrate gleich			
Bemerkungen				

Template Slots (1)

ID	Internal ID + First Author's Last Name + Year
Chapter	Diseases, Cancer-Types
Mapping	Internal
Type	Study Type
Intervention	Risk Factors, Prevention, Medications, Therapies
Comparison Intervention	For comparison
Study Count	Start and Completion Numbers
Results	Results

Template Slots (2)

Ergebnisse		12
	Schmerzen stärker	VAS 25,1 vs. 20,7; p=0,032
	Vulvakarzinom	
	Narbenbildung häufiger	P<0.01
	Rezidivrate gleich	

- Unstructured format
- Slot contains another template

Template Slots (3)

- After Deletion of Mapping & Results columns

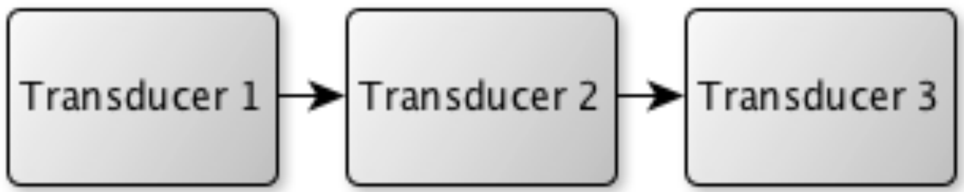
ID	Internal ID + First Author's Last Name + Year
Chapter	Diseases, Cancer-Types
Type	Study Type
Intervention	Risk Factors, Prevention, Medications, Therapies
Comparison Intervention	For comparison
Study Count	Start and Completion Numbers

Fill Rules (1)

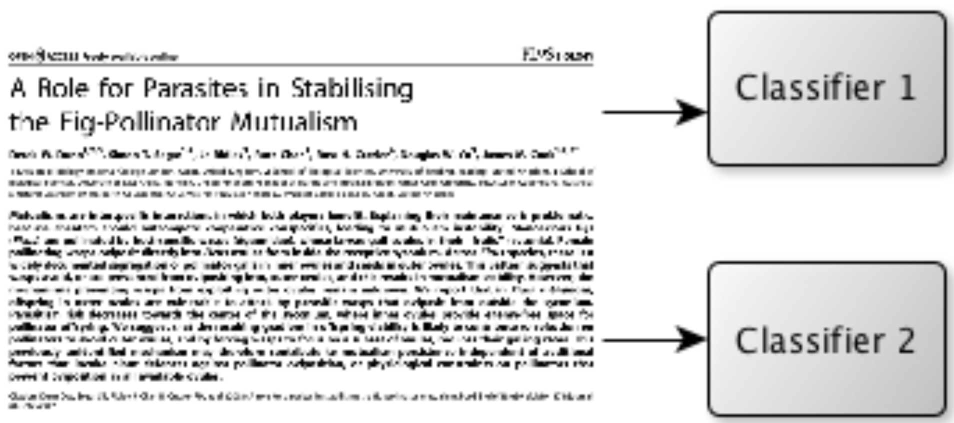
- Searched for existing tools in the biomedical domain
- Found systems in papers:
  - are not available
  - do not offer public access
  - are highly specialized for one template or subdomain

→ **Customized Approach**

Fill Rules (2)



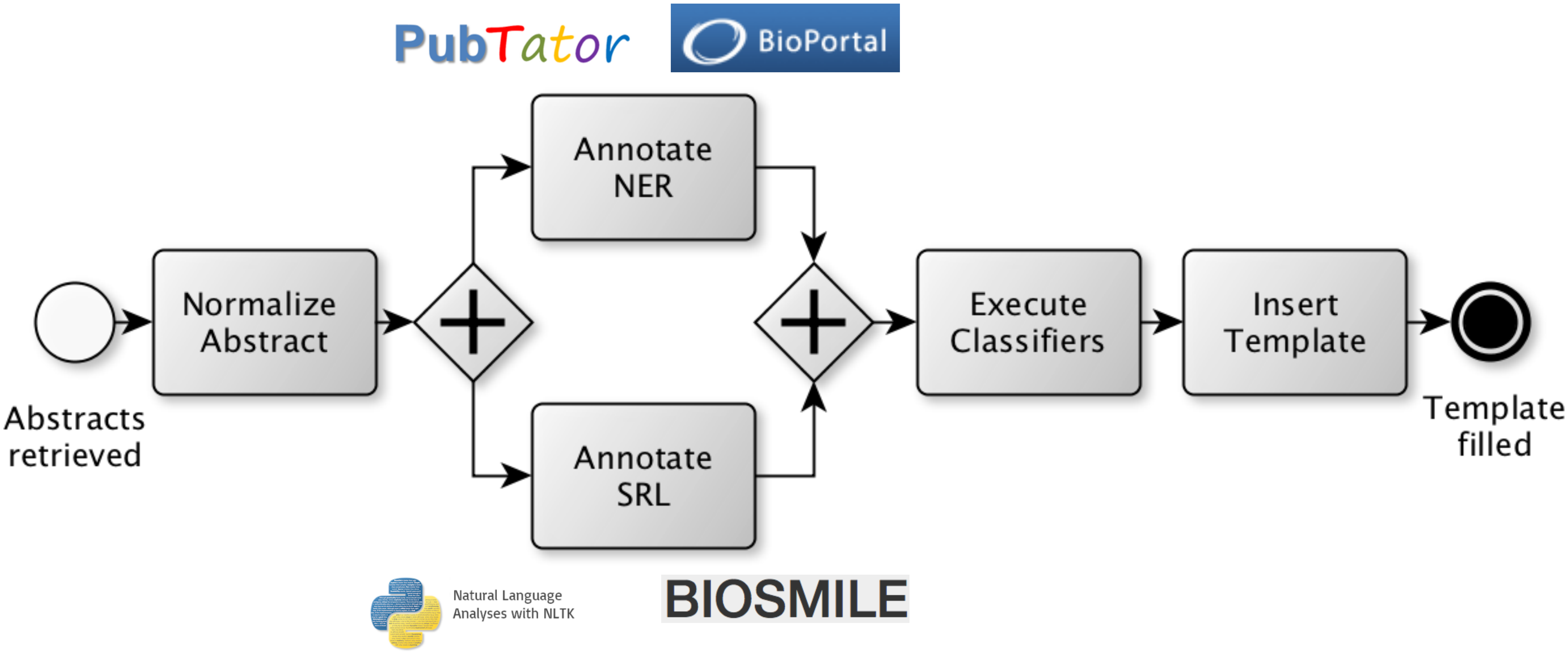
- Finite-State Cascade Approach



- One Classifier per Slot

- Reuse existing NLP Tools

Workflow


PubTator
BioPortal
Annotate NER
Normalize Abstract
Execute Classifiers
Insert Template
Abstracts retrieved
Template filled
Annotate SRL
Natural Language Analyses with NLTK
BIOSMILE

Slot Filling: **Study Count**

Candidate Sentences

“110 patients were randomly assigned.”

“Hundred and forty women […] received 6 courses of PLD 40 mg/m2 and carboplatin (AUC 6) every 28 days”

Normalization (1)

Problem: Numbers are in different format

*"**One hundred and ten** patients were randomly assigned."*

*"**140** women […] received 6 courses of PLD 40 mg/m2 and carboplatin (AUC 6) every 28 days"*

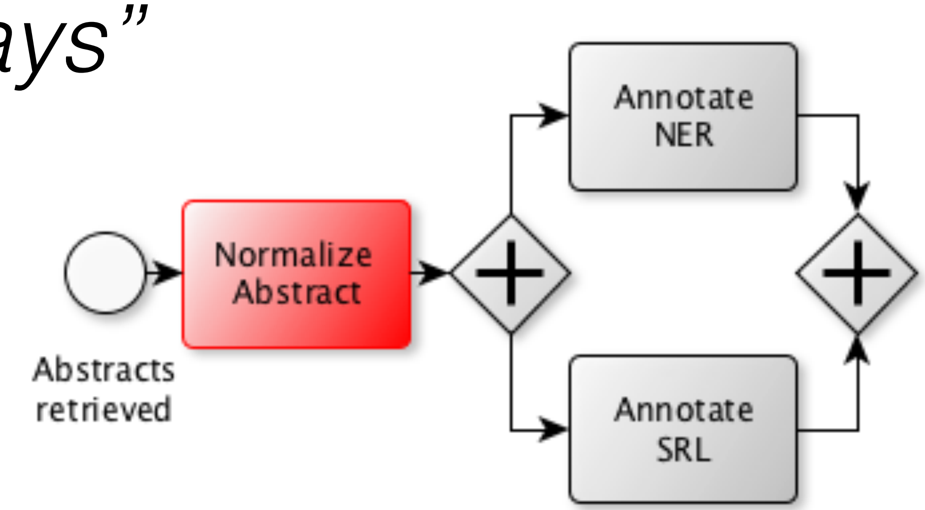
Normalization (2)

1. Convert word numbers to numeric numbers
   - Use regular expressions for substitutions

"110 patients were randomly assigned."

"140 women [...] received 6 courses of PLD 40 mg/m2 and carboplatin (AUC 6) every 28 days"

Study Count := Number



Named Entity Recognition (NER) (1)

Problem: Numbers are not efficient to detect study count

*“**110** patients were randomly assigned.”*

*“**140** women […] received **6** courses of PLD **40** mg/m2 and carboplatin (AUC **6**) every **28** days”*

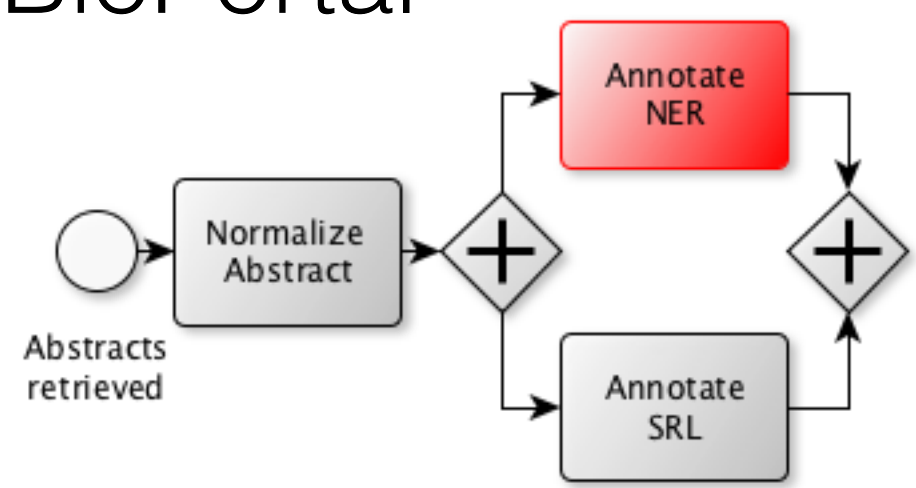
→ **Look at type of entity using NER**

Named Entity Recognition (NER) (2)

PubTator

BioPortal

2. Identify Named Entities
   - Identify and classify nouns
   - Resolve synonyms, normalizes entities
   - e.g. done by looking for presence in a named entity list
   - Usage of tools PubTator and BioPortal



NER: PubTator (1)

- Among other functionality pre-annotates Bio-Entities in PubMed articles
- Nightly updates of annotations using state-of-the-art text-mining tools
- Provides species entity type among others
- Accessible via web or REST API

NER: PubTator (2)

- Uses SR4GN tool for identifying species
- Comparably good accuracy of 85.42%
- Uses dictionary look up
  - 4 dictionaries: NCBI Taxonomy, etc.
  - Handles synonyms using regular expressions
- Maps them to NCBI Taxonomy IDs

NER: PubTator (3)

“110 ***patients*** *were randomly assigned.”*

“140 ***women*** *[…] received 6 courses of PLD 40 mg/m2 and carboplatin (AUC 6) every 28 days”*

Entity type	Entity mention	Concept ID	Nomenclature
Species	patients women	9606	NCBI Taxonomy

Study Count := Number + Entity of NCBI Concept 9606

NER: PubTator (4)

Problem: What if trial group are men?

It still works, as 9606 is the superclass Homo Sapiens

Problem: Does PubTator match every needed entity type?

No, e.g. treatments are not matched

→ **Use a more general NER tool**

NER: BioPortal (1)

- Annotates Bio-Entities in texts using many different ontologies
- Matches words in the text to terms in ontologies by doing an exact string comparison
- Possible to provide UMLS Semantic Type
- Accessible via web or REST API

NER: BioPortal (2)

- Select MESH ontology

Class	Ontology	Type	Matched Class
Patients	MESH	direct	Patients
Persons	MESH	ancestor	Patients
Women	MESH	direct	Women
Persons	MESH	ancestor	Women

Study Count := Number + Entity of MESH Type Persons

Semantic Role Labelling (SRL) (1)

Problem: Role of the entities is unknown

*"**110 patients** were randomly assigned. **96 patients** completed 1 year follow-up."*

"140 women […] received 6 courses of PLD 40 mg/m2 and carboplatin (AUC 6) every 28 days."

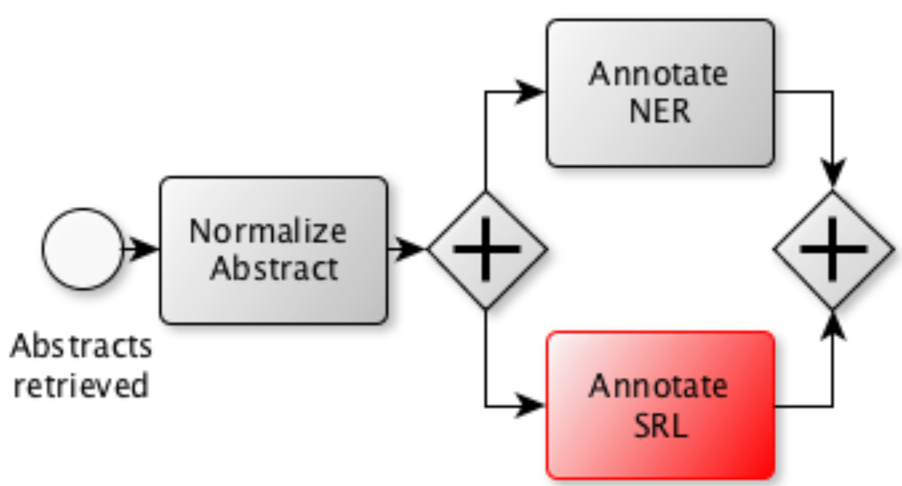
→ **Use SRL to identify role of the entity**

Semantic Role Labelling (SRL) (2)

BIOSMILE

Natural Language Analyses with NLTK

3. Identify Semantic Roles of Entities
    - Identify relationships between entity
    - Classify role of entity in relationship
    - Usage of tools BioSmile



SRL: BioSmile (1)

- BioProp contains list of biomedical predicate-argument relations

Roleset **eliminate.01**
Roles:
Arg0:entity removing
Arg1:thing being removed

Arg0 **Arg1**
Doctor *eliminates* *tumor*

- BioSmile uses BioProp to annotate roles to biomedical entities in texts
- F-Score of 87%

SRL: BioSmile (2)

“110 patients were randomly assigned.”

“140 women […] received 6 courses of PLD 40 mg/m2 and carboplatin (AUC 6) every 28 days”

- BioSmile does not detect any roles, due to small size of BioProp

→ **Use more general proposition bank**

SRL: PropBank + NLTK (1)

- PropBank contains basic semantic propositions gathered from news articles
- NLP Library NLTK contains PropBank corpus with examples
- Train supervised model to identify roles

SRL: PropBank + NLTK (2)

Verb Receive

Roles:

Arg0-PAG: *receiver* (vnrole: 13.5.2-1-Agent)
Arg1-PPT: *thing gotten* (vnrole: 13.5.2-1-Theme)
Arg2-DIR: *received from* (vnrole: 13.5.2-1-Source)

Verb Assign

Roles:

Arg0-PAG: *agent, assigner* (vnrole: 13.3-Agent)
Arg1-PPT: *thing assigned* (vnrole: 13.3-Theme)
Arg2-GOL: *assigned to* (vnrole: 13.3-Goal)

thing assigned

*"110 **patients** were randomly **assigned**."*

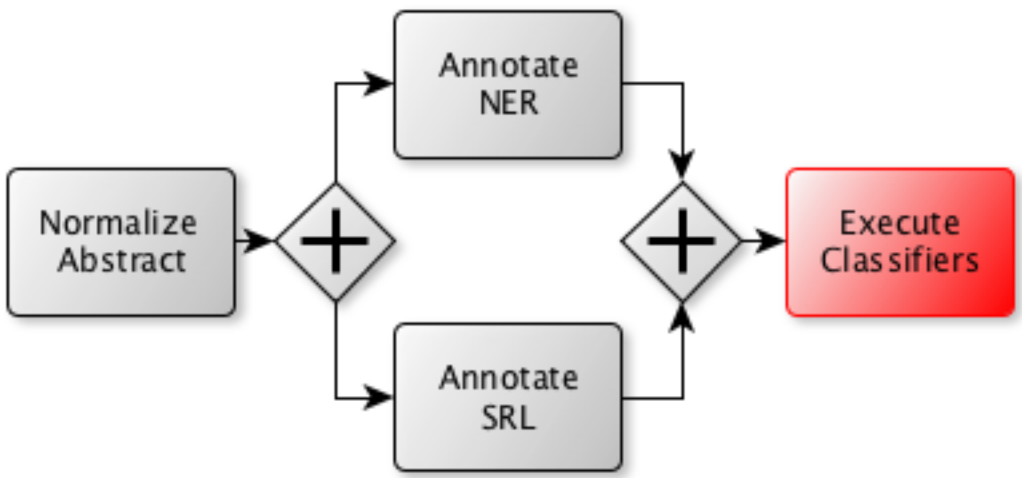
receiver

*"140 **women** **received** 6 courses of PLD 40 mg/m2 and carboplatin (AUC 6) every 28 days"*

Study Count := Number +
Entity of Type Persons +
Role thing assigned | receiver

Classifier

- Executes rules to extract candidates for slots
- Resolves conflicts using a learned model
- Outputs slot values



Shortcomings

PubTator

- Latest papers are not annotated
- Contains only five entity classes

BioPortal

- Performance depends on ontology

BIOSMILE

- BioProp contains few predicates

Natural Language Analyses with NLTK

- No biomedical semantic
- Model has to be created to learn annotation

Hands-On Session

What are your results?

Preliminary Results

- **ID:** good accuracy, as easy to extract
- **Chapter:** good accuracy, use title as indicator
- **Type:** good accuracy, look for key words
- **Intervention, Comparison:** ok accuracy, lots of candidates
- **Study Count:** ok accuracy, many different candidates

Future Work

- Look deeper into possible automation of filling rules
- Iterate over slot rules
- Train NLTK SRL model
- More data

Research Future Work

- More publicly available tools
- Integrate Multiple Documents
- Self-learned templates vs manually-crafted templates

Questions?