bioinformatics Gent

Analyzing text in search of bio-molecular events: a high-precision machine learning framework

Sofie Van Landeghem, Yvan Saeys,

Bernard De Baets, Yves Van de Peer

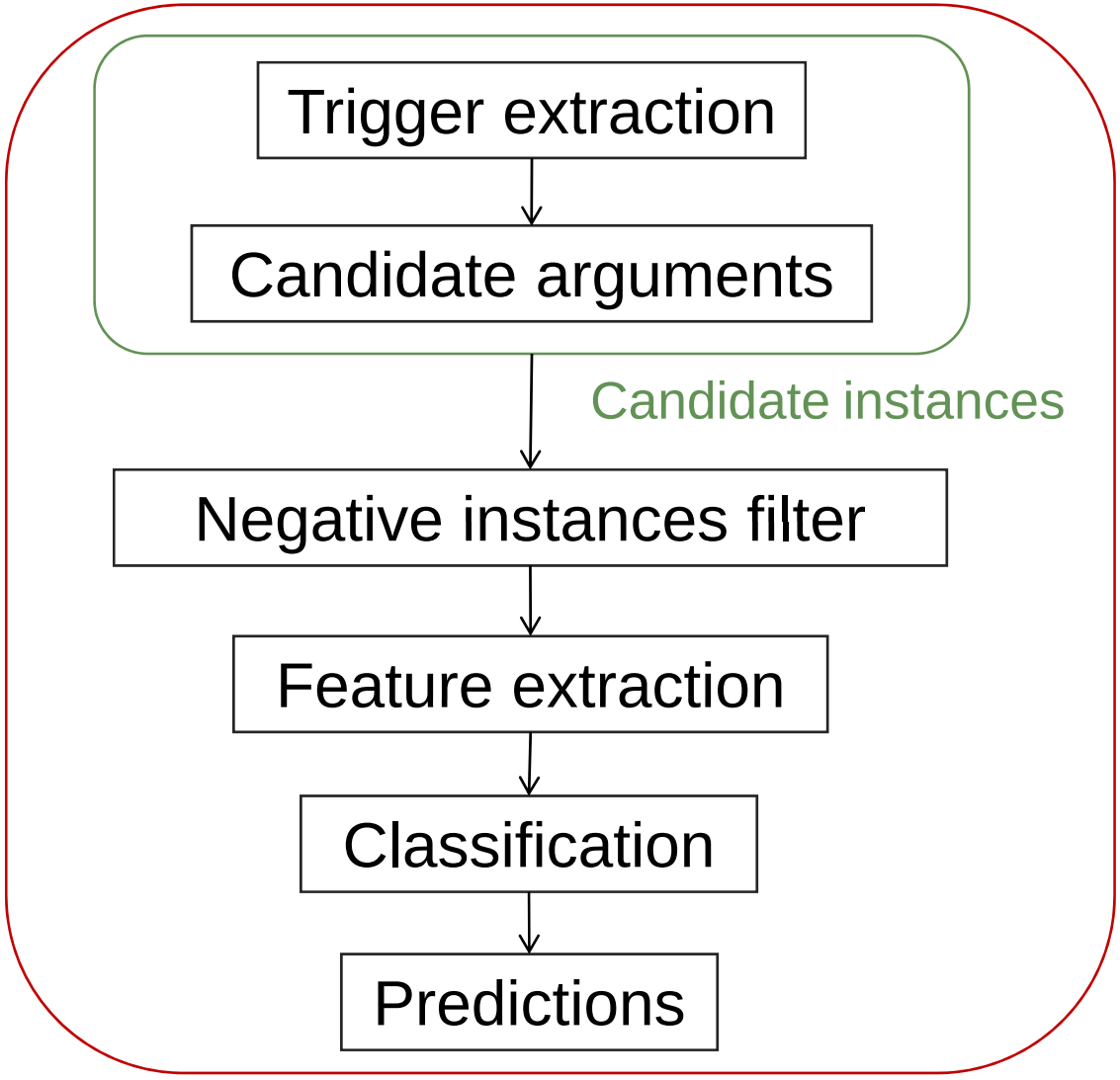
Friday June 5th, 2009

BioNLP 2009, Boulder, Colorado

UNIVERSITEIT GENT

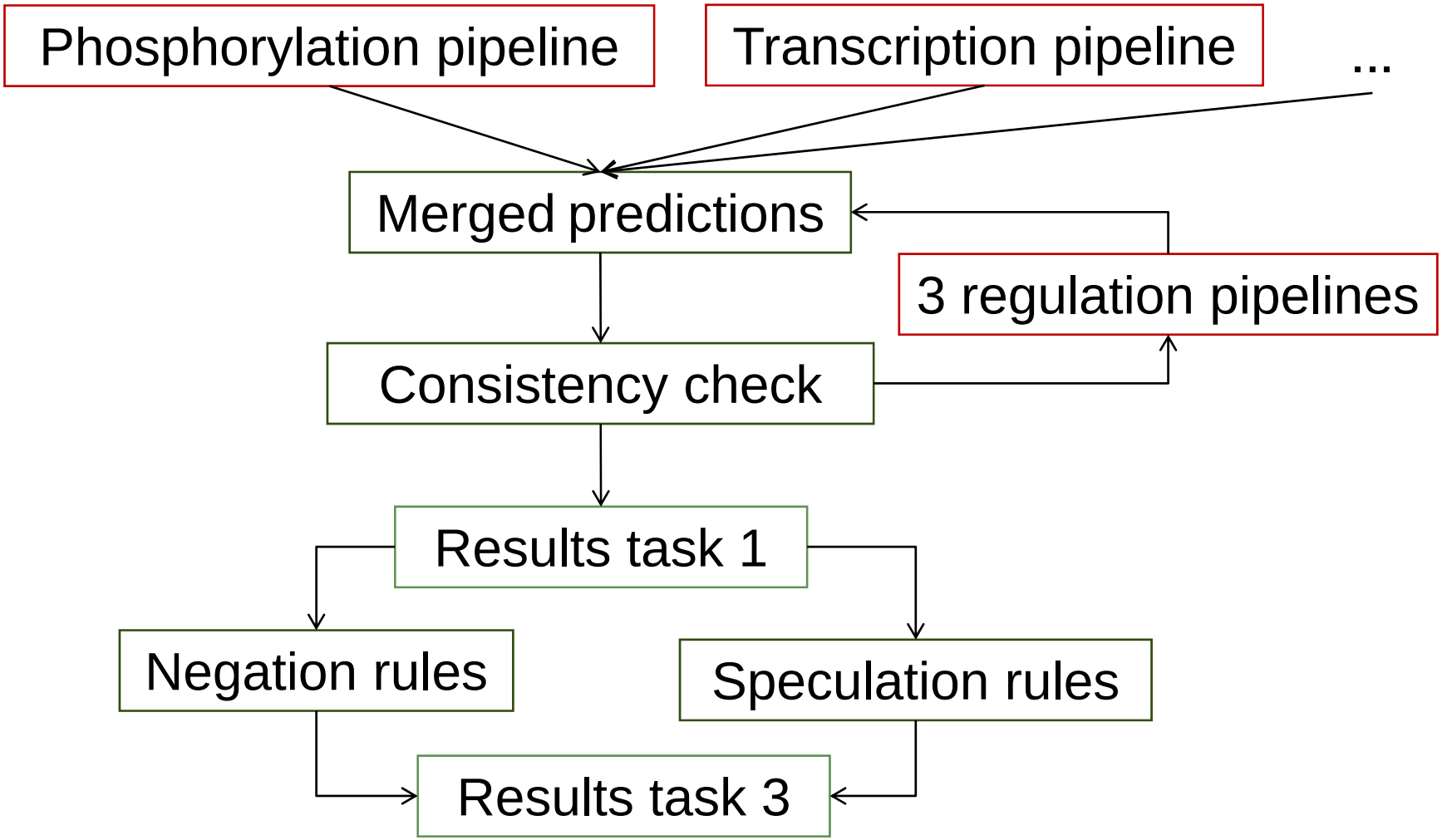
VIB

General ML pipeline



Pipeline for one specific event type

Framework overview



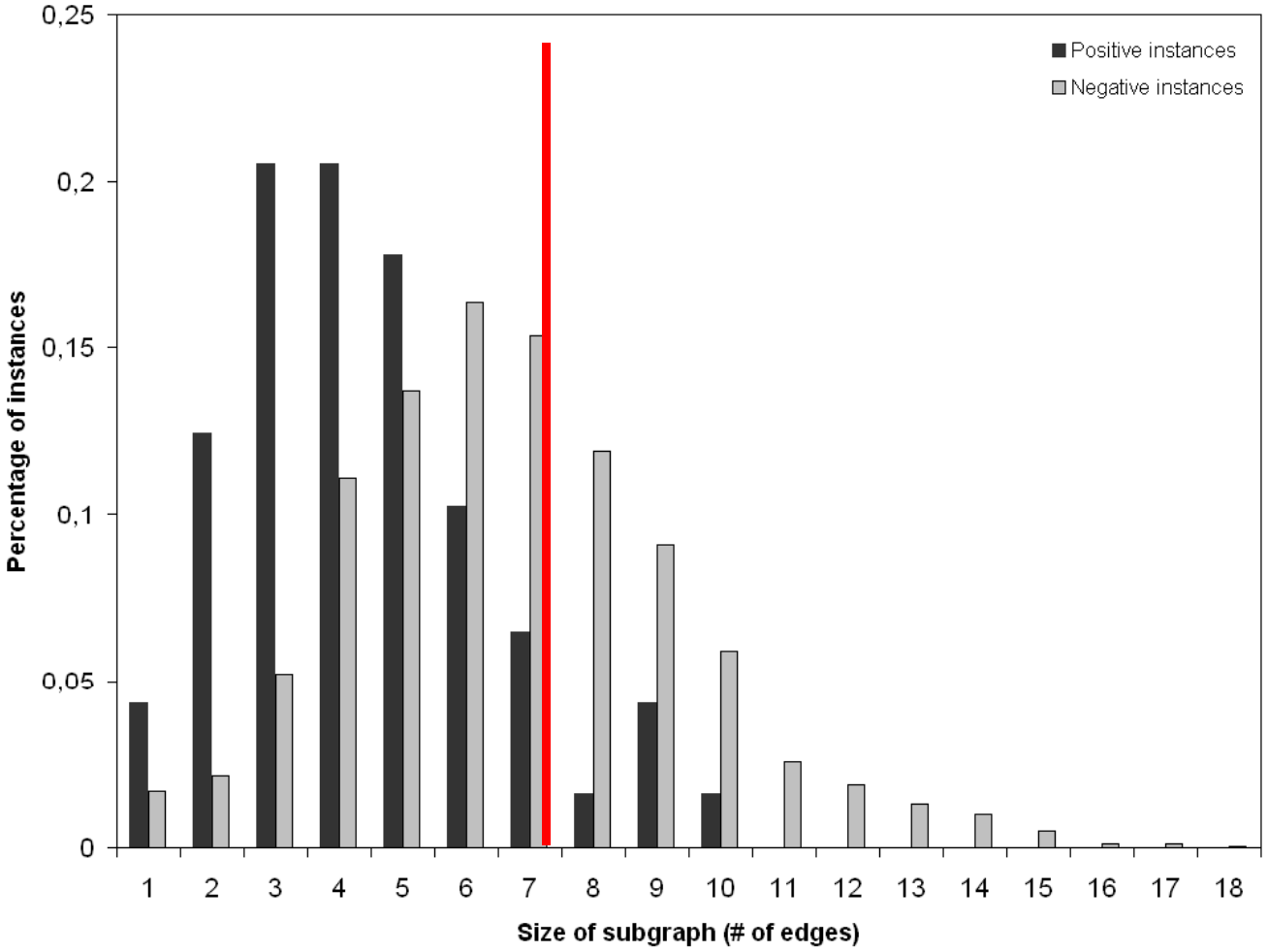
Trigger dictionaries

- E.g. “phosphorylated”, “overexpression”
- Porter stemming
- Separate dictionary for each type of event
- Compiled automatically from training data
- Manually filtered to remove general words
  - Such as “are”, “via” or “through”
- Binding : distinction between
  - “Single” (e.g. “homodimer”, “binding site”)
  - “Multiple” (e.g. “heterodimer”, “complex”)

Instance creation

1. Finding triggers in the text
2. Initially : all (combinations of) proteins / events that appear in the same sentence as the trigger
   - Lots of noise -> high-dimensional datasets
3. Implementation of Negative-instances (NI) filter
   - Checks the length of the sub-sentence spanned by the candidate event & applies a cut-off value
   - Checks the size of the subgraph of the dependency graph, corresponding to the candidate event & applies a cut-off

Negative instances filter



Distribution of Multiple binding instances (training data)

Dimensionality reduction

		# instances	positive instances
Manually filtered dictionaries	Binding	34.612	2%
Distinction between single and multiple binding	Single Multiple	4708 3861	11% 5%
Application of the NI filter	Single Multiple	4070 2365	13% 8%

Distribution of Binding instances for different design choices

(training data)

Feature generation

- Stanford dependency parsing : smallest subgraph
- Vertex walks extracted from the dependency subgraph
  - Vertex – edge – vertex
  - Lexical variant : trigger/protein blinded : e.g. "*trigger nsubj protx*"
  - Syntactic variant : e.g. "*nn nsubj nn*"
- Bag-of-words : nodes of dependency graph
  - Excludes uninformative words such as prepositions
- Stemmed trigrams e.g. "*by induc transcript*"
- Lexical and syntactic information of the triggers
- Length of the sub-sentence & size of the subgraph
- Regulation : whether arguments are proteins or events

Classification

- High-dimensional and highly unbalanced datasets
- Support vector machine (SVM)
- LibSVM implementation as provided by WEKA
- Kernel type : radial basis function (default)
- Internal 5-fold CV loop to tune parameters

Consistency check (1)

Overlapping triggers of different event types

- Predictions for different event types are processed in parallel, independently of each other, and merged afterwards
- One word in the text might lead to two distinct triggers of different type
- E.g. “expression” can lead to both a Transcription and a Gene expression event
- But in real life, this never happens at the same time!

➢ Keep only the prediction with the highest SVM score

➢ Minimal overlap between dictionaries can avoid this inconsistency to occur in the first place

Consistency check (2)

Different events from the same type, based on the same trigger

- One trigger is involved in many events from the same type
- E.g. “It induces expression of STAT5-regulated genes is CTLL-2, i.e. beta-casein, and oncostatin M (OSM)”
- 2 Gene expression events based on the trigger “expression”, one with beta-casein and one with OSM.
- This happens often, and the predictions will have similar SVM scores

➢ However, for some types, usually only one true event is linked to each trigger (Protein catabolism & Phosphorylation) → keep only top-ranked result

Performance - Task 1

Event type	Recall	Precision	F-score
Localization	43.68	78.35	56.09
Binding	38.04	38.60	38.32
Gene expression	59.42	81.56	68.75
Transcription	39.42	60.67	47.79
Protein catabolism	64.29	60.00	62.07
Phosphorylation	56.30	89.41	69.09
Total	50.75	67.24	57.85
Regulation	10.65	22.79	14.52
Positive regulation	17.19	32.19	22.41
Negative regulation	22.96	35.22	27.80
Total	17.36	31.61	22.41
Task 1	33.41	51.55	**40.54**

Negation

- Three cases
  1. Negation construct in close vicinity of the trigger
     - "CsA was found not to inhibit lck gene expression."
  2. Trigger is inherently negative
     - "This was associated with a reduction in endothelial MCP-1 secretion and GRO-alpha immobilization."
  3. The "but not" pattern
     - "Overexpression of Vav, but not SLP-76, augments CD28-induced IL-2 promoter activity."
- Custom made rule based system : locating certain words, triggers and patterns that indicate negation

Speculation

- Two cases
  1. Uncertainty (stating the research)
     - "We examined the ability of type I and type II IFNs to regulate activation of STAT6 by IL-4 in primary human monocytes."
  2. Hypothesis (interpreting the research)
     - "(…) suggesting that these nuclear proteins may determine the IP-10 mRNA inducibility by IFNgamma."
- Custom made rule based system : locating certain words that indicate speculation

Performance - Task 3

Event type	Recall	Precision	F-score
Negation	10.57	45.10	17.13
Speculation	8.65	15.79	11.18
Total	9.66	24.85	13.91

- The results of this subtask heavily depend on the results of subtask 1
- When we only consider events found in subtask 1, recall of the rule-based system is actually higher: above 50%.

Thanks to ...

- My supervisors
  - Yvan Saeys
  - Yves Van de Peer
  - Bernard De Baets

- The whole Bioinformatics team @ University of Ghent, Belgium

- Many thanks to the BioNLP'09 organizers, for offering to the community a very valuable and well organized task about event extraction!