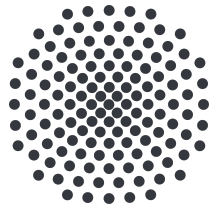
Towards Confidence Estimation for Typed Protein-Protein Relation Extraction

Camilo Thorne Roman Klinger

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
{name.surname}@ims.uni-stuttgart.de

bioRANLP 17, Varna, Bulgaria, 08.09.2017



Motivation

IL-6	positive regulation	STAT-3
⋮	⋮	⋮
TIPE2	negative regulation	Snail2
⋮	⋮	⋮
growth factor	positive regulation	WSC domain
⋮	⋮	⋮

Institut für Maschinelle Sprachverarbeitung

Motivation

IL-6	positive regulation	STAT-3	true
⋮	⋮	⋮	
TIPE2	negative regulation	Snail2	
⋮	⋮	⋮	
growth factor	positive regulation	WSC domain	
⋮	⋮	⋮	

Institut für Maschinelle Sprachverarbeitung

Motivation

IL-6	positive regulation	STAT-3	true
⋮	⋮	⋮	
TIPE2	negative regulation	Snail2	
⋮	⋮	⋮	
growth factor	positive regulation	WSC domain	false
⋮	⋮	⋮	

Institut für Maschinelle Sprachverarbeitung

Motivation

IL-6	positive regulation	STAT-3	true
⋮	⋮	⋮	
TIPE2	negative regulation	Snail2	interesting!
⋮	⋮	⋮	
growth factor	positive regulation	WSC domain	false
⋮	⋮	⋮	

▷ The middle fact is not in the STRING database, but PubMed/MEDLINE abstract 28186089 mentions it:

> *"(...) TIPE2 (...) downregulated (...) Snail2 (...)"*

Institut für Maschinelle Sprachverarbeitung

Motivation

c_1	IL-6	positive regulation	STAT-3	true
		⋮	⋮	
c_i	TIPE2	negative regulation	Snail2	interesting!
		⋮	⋮	
c_{i+k}	growth factor	positive regulation	WSC domain	false
		⋮	⋮	

▷ The middle fact is not in the STRING database, but PubMed/MEDLINE abstract 28186089 mentions it:

> *"(...) TIPE2 (...) downregulated (...) Snail2 (...)"*

Institut für Maschinelle Sprachverarbeitung

Motivation

c_1	IL-6	positive regulation	STAT-3	true
		⋮	⋮	
c_i	TIPE2	negative regulation	Snail2	interesting!
		⋮	⋮	
c_{i+k}	growth factor	positive regulation	WSC domain	false
		⋮	⋮	

▷ The middle fact is not in the STRING database, but PubMed/MEDLINE abstract 28186089 mentions it:

> *"(...) TIPE2 (...) downregulated (...) Snail2 (...)"*

▷ Requirements: $c_i \leq c_{i+1}$, $c_i \in [0, 1]$, for $i \geq 0$

Institut für Maschinelle Sprachverarbeitung

Strategies

We hypothesize that relation discovery confidence scores rely on three main kinds of sources:

S1: The (aggregated) confidence scores of the individual modules of RE pipelines

S2: The internal graph structure of the discovered relations

S3: Evidence gathered from external knowledge sources, such as textual evidence or knowledge retrieved or inferred from structured knowledge sources (biomedical ontologies and databases)

Institut für Maschinelle Sprachverarbeitung

Strategies

We hypothesize that relation discovery confidence scores rely on three main kinds of sources:

S1: The (aggregated) confidence scores of the individual modules of RE pipelines

S2: The internal graph structure of the discovered relations

S3: Evidence gathered from external knowledge sources, such as textual evidence or knowledge retrieved or inferred from structured knowledge sources (biomedical ontologies and databases)

▷ In this work, we focus on strategy S1

Institut für Maschinelle Sprachverarbeitung

Datasets

MEDLINE English-only, protein- and gene-centered subset of from May 1992 to May 2017

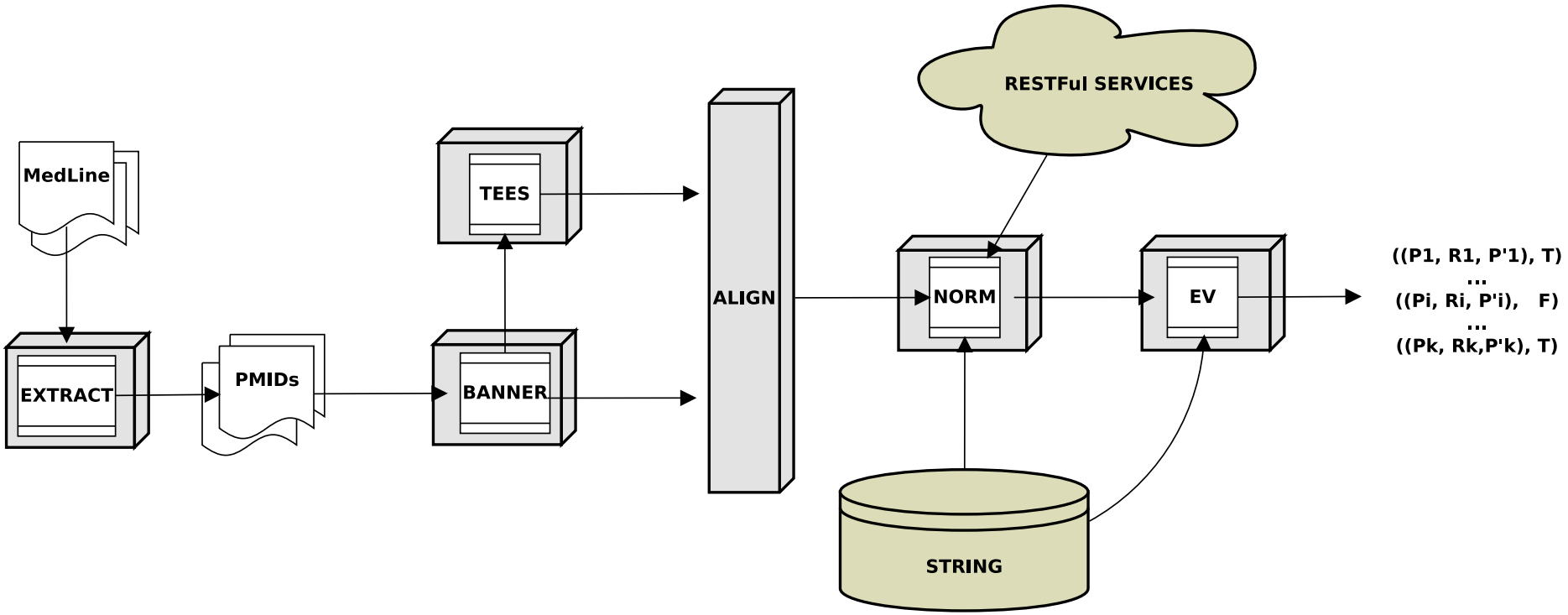
PMIDs 1376980 to 28211214
40,911,675 tokens in 1,939,915 abstracts
https://www.nlm.nih.gov/bsd/pmresources.html

STRING Human gene/protein subset of STRING database, which integrates many biomedical databases

20,458 unique genes/proteins with 6,013,567 unique typed interactions
https://string-db.org/

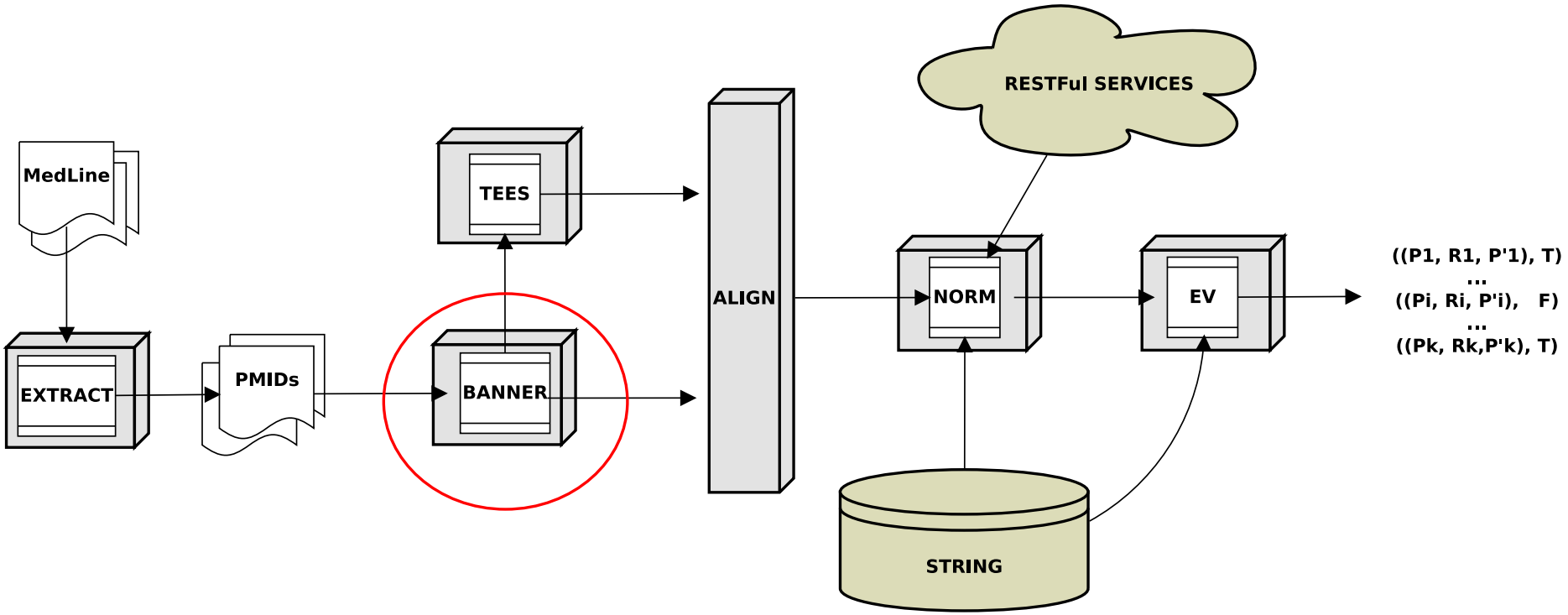
Institut für Maschinelle Sprachverarbeitung

System Architecture



Institut für Maschinelle Sprachverarbeitung

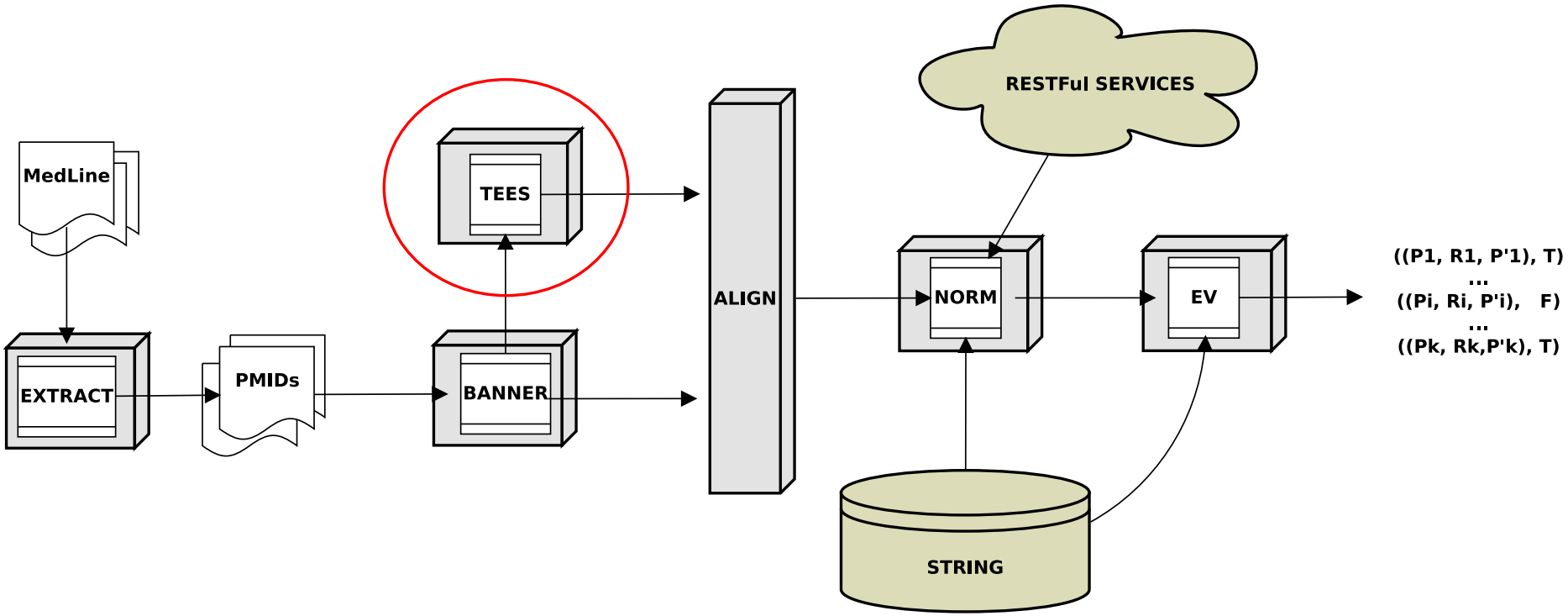
System Architecture



BANNER [LG08] for protein/gene recognition (CRF-based biomedical entity recognition tool)

Institut für Maschinelle Sprachverarbeitung

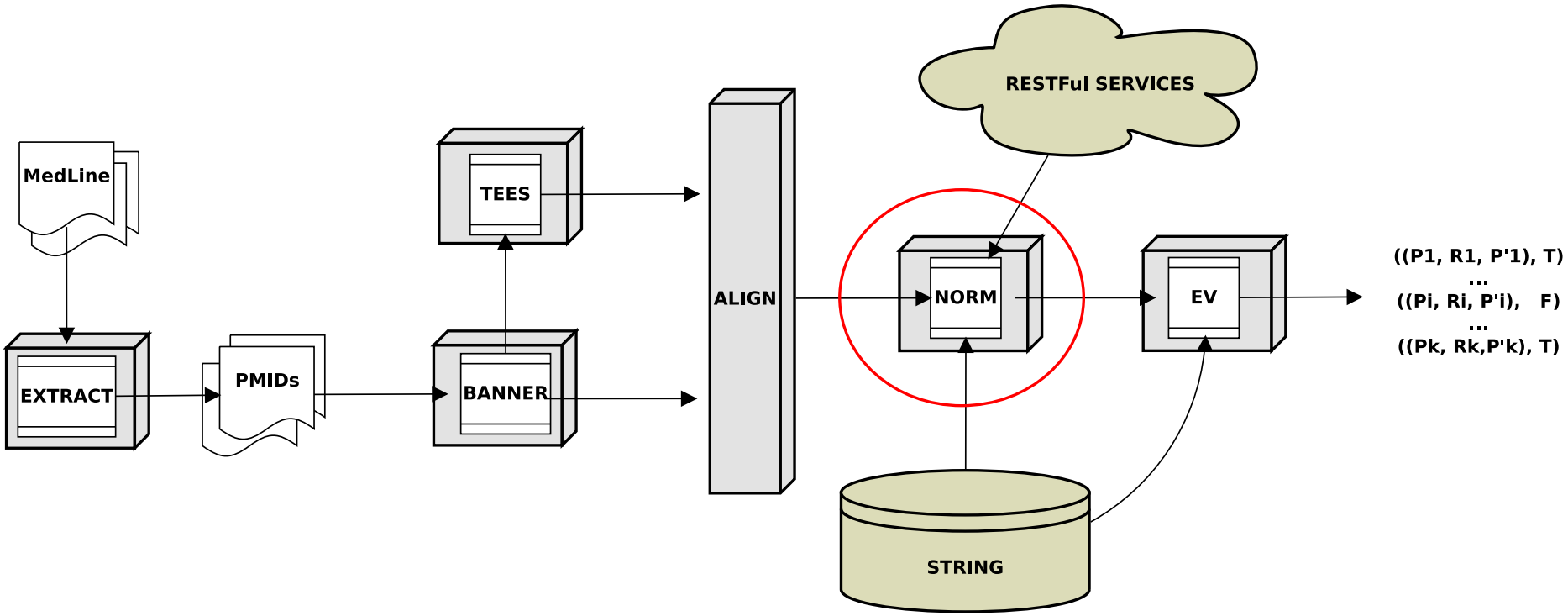
System Architecture



TEES [BS15] for protein-protein typed event recognition

Institut für Maschinelle Sprachverarbeitung

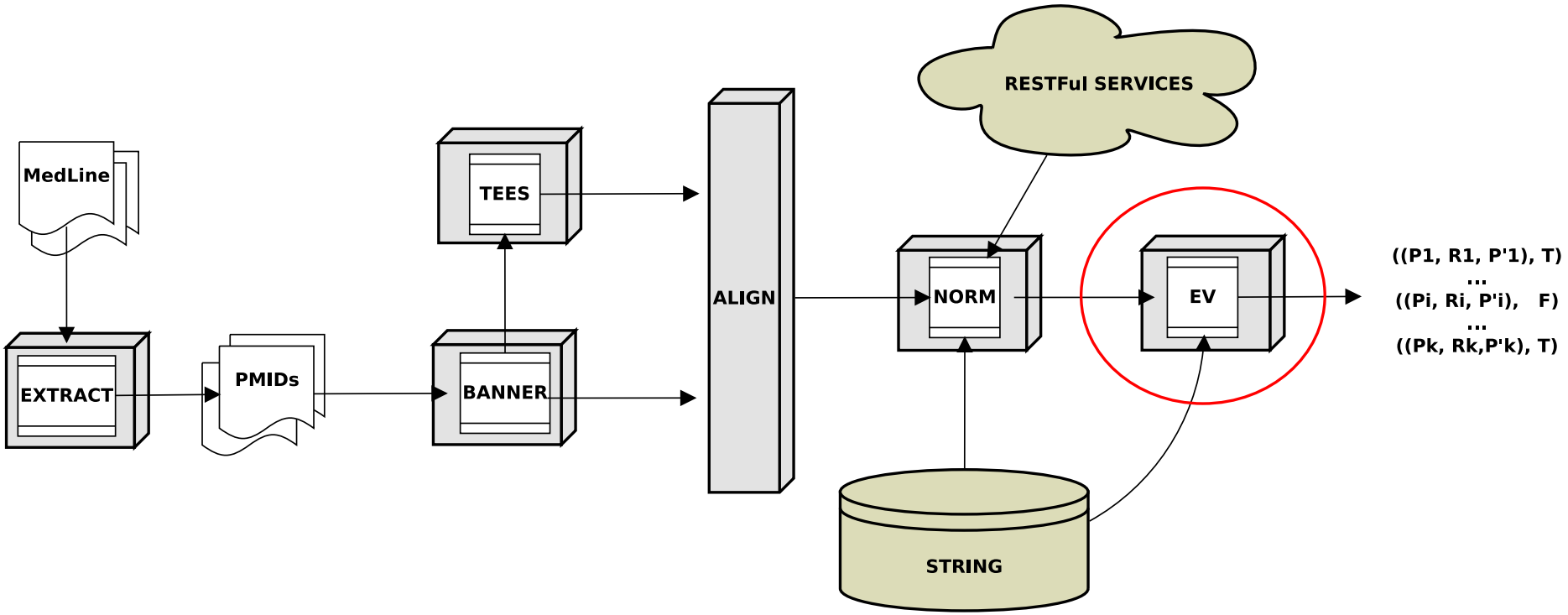
System Architecture



GNAT [HGH+11] and GNorm [WKL15] for protein normalization

Institut für Maschinelle Sprachverarbeitung

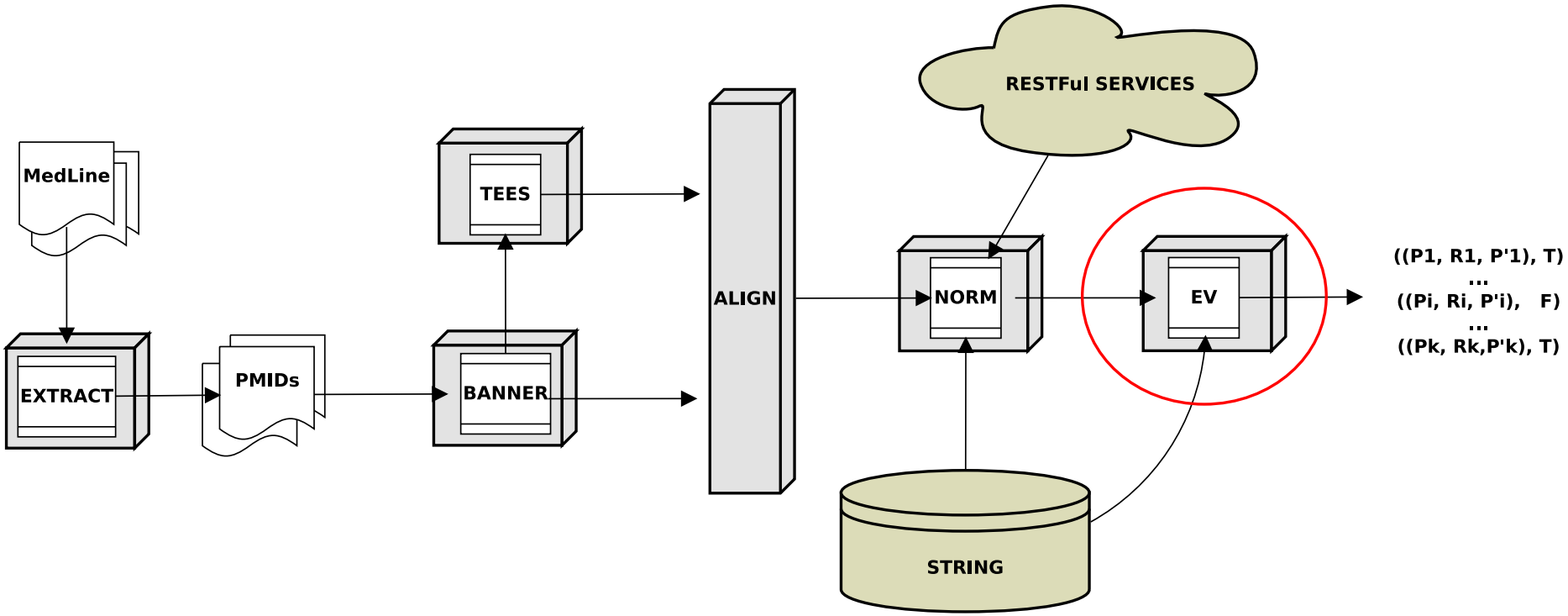
System Architecture



Extended the normalizers to normalize event types as well

Institut für Maschinelle Sprachverarbeitung

System Architecture



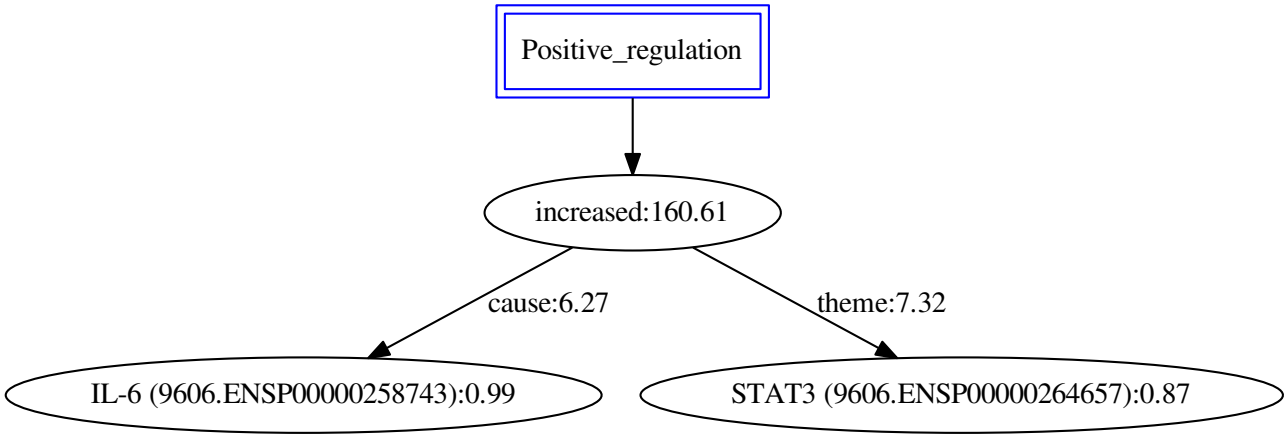
≈ 20,000 relations from ≈ 10,000 PMIDs

≈ 10,000 normalized relations per normalizer (GNAT, GNorm)

Institut für Maschinelle Sprachverarbeitung

Typed Protein-Protein Relations

A typed interaction is a triple $rel = (p_1, r, p_2)$ denoting a directed relation r between proteins p_1 and p_2



Western blot analysis showed that IL-6 **increased** JKA, STAT3, p-STAT3 and VEGF-C protein levels in the gastric cancer cells. (pmid 26750536)

(typed interaction extracted by the pipeline
for the first relation from the motivating example)

Institut für Maschinelle Sprachverarbeitung

Component-wise Confidences

For every candidate rel we compute component-wise confidence values

This gives rise to five confidence values for each rel

Institut für Maschinelle Sprachverarbeitung

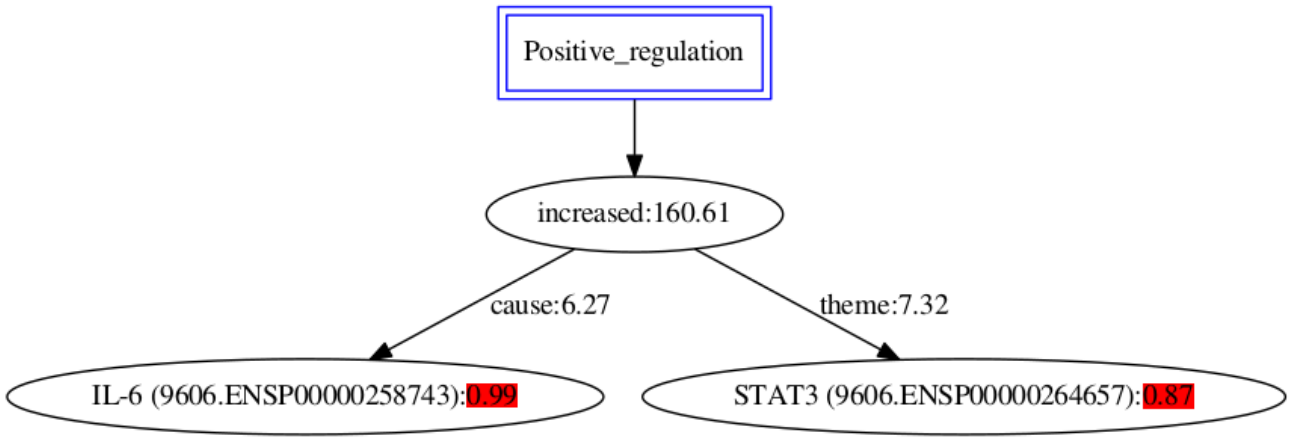
Component-wise Confidences

For every candidate rel we compute component-wise confidence values

This gives rise to five confidence values for each rel

▷ BANNER (CRF) gamma probability [CM04] of protein entities:

$\text{cf}_{\gamma_c}(p1)$ and $\text{cf}_{\gamma_t}(p2)$



Western blot analysis showed that IL-6 **increased** JKA, STAT3, p-STAT3 and VEGF-C protein levels in the gastric cancer cells. (pmid 26750536)

Institut für Maschinelle Sprachverarbeitung

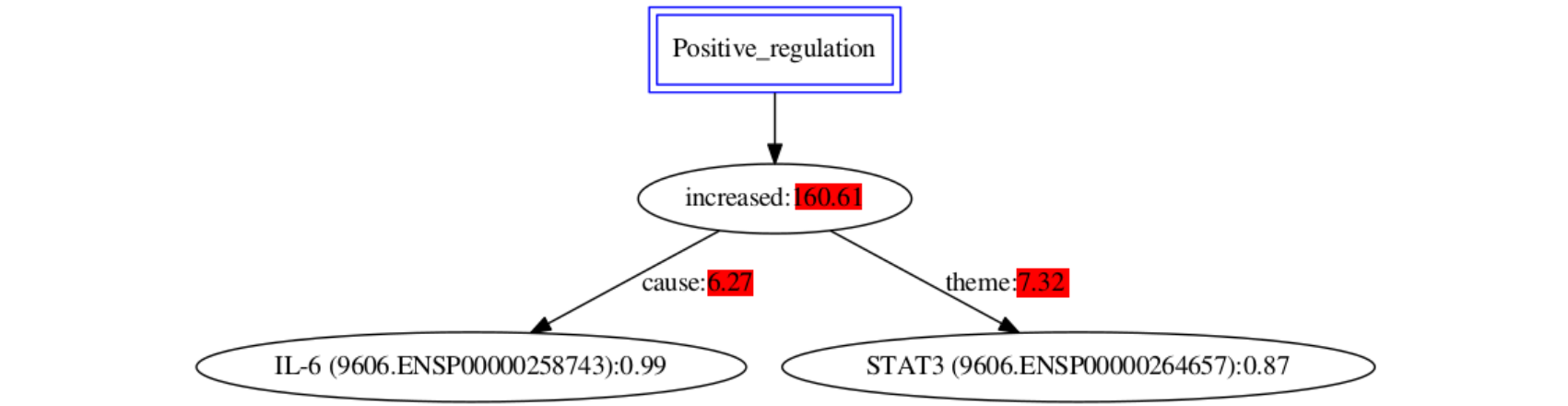
Component-wise Confidences

For every candidate rel we compute component-wise confidence values

This gives rise to five confidence values for each rel

- ▷ TEES (SVM) margins for resp., events and their causes and themes:

$\text{cf}_c(p1)$ and $\text{cf}_t(p2)$ and $\text{cf}_{ev}(r)$



Institut für Maschinelle Sprachverarbeitung

Aggregation Methods

- Global confidence as a linear combination of component-wise confidences, namely, a (weighted) average:

$$\mathsf{cf}_{avg}^{(i)} = \frac{1}{5} \cdot \sum_{m} we_m \cdot \mathsf{cf}_{m}^{(i)}$$

Institut für Maschinelle Sprachverarbeitung

Aggregation Methods

- Global confidence as a linear combination of component-wise confidences, namely, a (weighted) average:

$$\mathsf{cf}_{avg}^{(i)} = \frac{1}{5} \cdot \sum_m we_m \cdot \mathsf{cf}_m^{(i)}$$

- Global confidence as a (weighted) product of component-wise confidences:

$$\mathsf{cf}_{prd}^{(i)} = \prod_m we_m \cdot \mathsf{cf}_m^{(i)}$$

Institut für Maschinelle Sprachverarbeitung

Aggregation Methods

- Global confidence as a linear combination of component-wise confidences, namely, a (weighted) average:

$$\mathsf{cf}_{avg}^{(i)} = \frac{1}{5} \cdot \sum_m we_m \cdot \mathsf{cf}_m^{(i)}$$

- Global confidence as a (weighted) product of component-wise confidences:

$$\mathsf{cf}_{prd}^{(i)} = \prod_m we_m \cdot \mathsf{cf}_m^{(i)}$$

- Weights $\bar{we} = (\mathsf{cf}_{ev}, \mathsf{cf}_c, \mathsf{cf}_t, \mathsf{cf}_{\gamma_t}, \mathsf{cf}_{\gamma_c})^T$ were learned from our silver data via logistic regression
- We considered unweighted versions of the aggregation models via unit weights $\bar{we} = (1, 1, 1, 1, 1)^T$

Institut für Maschinelle Sprachverarbeitung

Train and Test Datasets

- GNAT and GNorm give rise to two silver standards, heavily biased towards false relations ($> 80\%$)
- We oversampled positive relations and undersampled negative relations to derive two balanced training sets

Institut für Maschinelle Sprachverarbeitung

Train and Test Datasets

- GNAT and GNorm give rise to two silver standards, heavily biased towards false relations ($> 80\%$)
- We oversampled positive relations and undersampled negative relations to derive two balanced training sets
- For each silver standard, we extracted a test set (disjoint from training sets)

GNAT

Train: 2000 relations
(balanced via resampling)

Test: 1000 relations

GNorm

Train: 2000 relations
(balanced via resampling)

Test: 1000 relations

Institut für Maschinelle Sprachverarbeitung

Weight $\bar{we}$ Learning

- We learned logistic models from the training datasets

$$P(\ rel^{(i)} = \mathsf{True} \mid \bar{\mathsf{cf}}^{(i)}\ ) = (1 + \exp(-\sum_{m} \theta_m \cdot \mathsf{cf}_m^{(i)}))^{-1}$$

Weight $\bar{we}$ Learning

- We learned logistic models from the training datasets

$$P(\ rel^{(i)} = \text{True} \mid \bar{\text{cf}}^{(i)}\ ) = (1 + \exp(-\sum_{m} \theta_m \cdot \text{cf}_m^{(i)}))^{-1}$$

- Cross-evaluated models over train/test pairs to choose best model

train dataset	test dataset	F_1
gnat_train	test_gnat	0.688
gnorm_train	test_gnat	0.658
gnat_train	**test_gnorm**	**0.766**
gnorm_train	test_gnorm	0.733

- Normalized AOD values to $[0, 1]$ to infer $\bar{we}$

Institut für Maschinelle Sprachverarbeitung

Correlation-based Evaluation of Aggregation Models

estimator	Kendall τ	p-value
cf_{prod} (unweig.)	0.041	0.127
cf_{prod}	0.041	0.127
cf_{avg} (unweig.)	0.032	0.210
cf_{avg}	**0.050**	**0.056**

- In **bold** the model with the highest τ value
- No test was statistically significant $\Rightarrow$ absence of correlation with linking judgments

Institut für Maschinelle Sprachverarbeitung

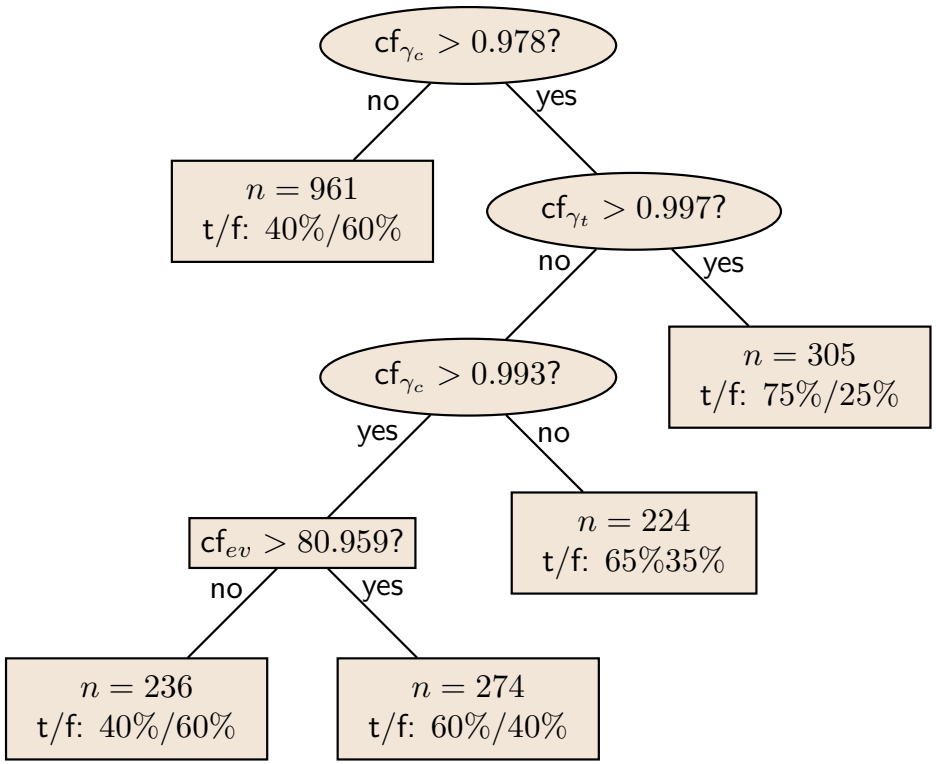
Confidence Feature Ranking - AOD

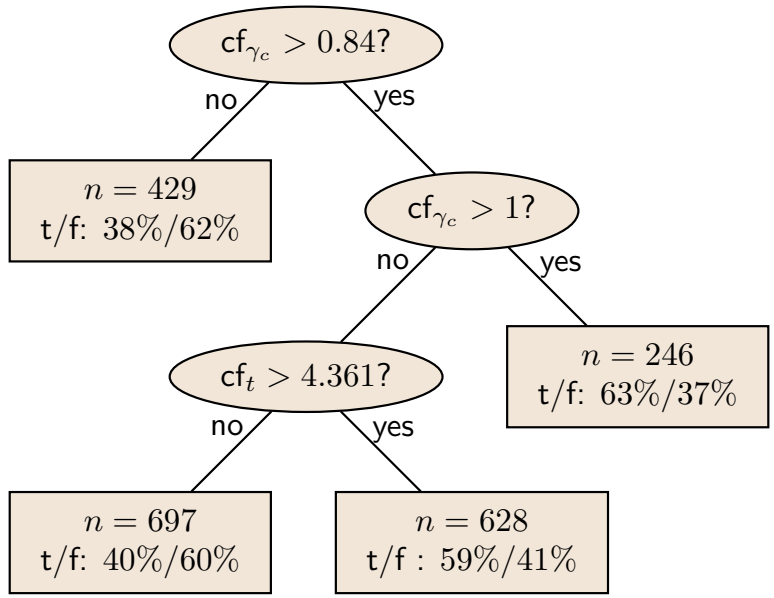
feature	deviance	p-value
cf_{ev}	6.200	0.013
$\mathbf{cf}_{t}$	**17.803**	$\mathbf{2.451 \cdot 10^{-05}}$
cf_{c}	4.858	0.028
cf_{γ_t}	2.370	0.124
$\mathbf{cf}_{\gamma_c}$	**22.667**	$\mathbf{1.926 \cdot 10^{-06}}$

- Analysis of deviance table for the best logistic model
- In **bold** the features with greater impact, and $p < 0.01$

Institut für Maschinelle Sprachverarbeitung

Confidence Feature Ranking - Decision Trees





- ▷ Right: J48 decision tree for the GNAT silver standard (training set)
- ▷ Left: J48 decision tree for the GNorm silver standard (training set)

Institut für Maschinelle Sprachverarbeitung

Conclusions

- The prediction confidence of some modules seems to influence positive decisions
- Confidence aggregation is not enough to define estimation models satisfying our criteria
- Confidence alone does not provide sufficient evidence to rank relations
- In this work the confidence of normalization was not fully addressed
- As further work we plan to focus on more complex evidence gathering methods

Institut für Maschinelle Sprachverarbeitung

Institut für Maschinelle Sprachverarbeitung

References

Jari Björne and Taio Salakoski.

Tees 2.2: Biomedical event extraction for diverse corpora.

BMC Bioinformatics, 16(16):S4, 2015.

Aron Cullota and Andrew McCallum.

Confidence estimation for information extraction.

In *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL '04, pages 109–112, 2004.

Jög Hackenberg, Marin Gerner, Maximilian Haeussler, Illés Solt, Conrad Plake, Martin Schroder, Graciela Gonzalez, Goran Nenadic, and Casey M. Bergman.

The GNAT library for local and remote gene mention normalization.

Bionformatics, 27(19):2769–2771, 2011.

Robert Leaman and Graciela Gonzalez.

BANNER: An executable survey of advances in biomedical named entity recognition.

In *Proceedings of the 2008 Pacific Symposium on Biocomputing*, PSB '08, pages 652–63, 2008.

Chih-Husuan Wei, Hung-Yu Kao, and Zhiyoung Lu.

GNomrPlus: An integrative approach for tagging genes, gene families, and protein domains.

BioMed Research International, 2015(2015):ID 918710, 2015.

Institut für Maschinelle Sprachverarbeitung

Appendix: Gamma Probability

- BANNER is a CRF-based protein recognizer
- We can leverage Viterbi matrix backward and forward probabilities to derive that of a complete entity span
- CRF gamma probability [CM04] of entity $p = w_i, \ldots, w_t$ starting at position t in a text with BIO labels $(s_t, \ldots, s_{t+k})$, $s_i \in \{\mathsf{B}, \mathsf{I}\}$ can be defined as

$$\mathrm{cf}_{\gamma_t}(p) = \prod_{i=t}^{k} \gamma_i(s_i) = \frac{\alpha_i(s_i) \cdot \beta_i(s_i)}{P(w_0, \ldots, w_i | \Lambda)}$$

this characterizes the likelihood that a given span of MEDLINE tokens is indeed a protein!

Institut für Maschinelle Sprachverarbeitung

Appendix: Matching

We match to STRING typed interaction as follows:

1. a protein mention p to a canonical form *i.e.*, STRING protein unique identifiers (UIDs)

$$\mathrm{norm}_{\mathsf{N}}(p) = \begin{cases} \mathrm{gn}_{\mathsf{N}}(p), & \text{if } \mathrm{gn}_{\mathsf{N}}(p)\downarrow \\ \mathrm{lk}(p), & \text{if } \mathrm{gn}_{\mathsf{N}}(p)\uparrow,\ \mathrm{lk}(p)\downarrow \\ \mathsf{NA}, & \text{if } \mathrm{gn}_{\mathsf{N}}(p)\uparrow,\ \mathrm{lk}(p)\uparrow \end{cases}$$

2. a relation/event type r to a STRING interaction type

$$\mathrm{ev}(r) = \begin{cases} \{\mathrm{inhibitits}\}, & \text{if } r \text{ is a neg. reg.} \\ \{\mathrm{expresses},\ \mathrm{activates}\}, & \text{if } r \text{ is a pos. reg.} \\ \{\mathrm{expresses},\ \mathrm{activates}\} \cup \{\mathrm{inhibits}\}, & \text{if } r \text{ is a reg.} \end{cases}$$

Institut für Maschinelle Sprachverarbeitung

Appendix: Pipeline Statistics

	Unit	Count
	PMIDs	11773
	Relations	21169
Elements	Proteins	11726
	Events	864
	Causes	5694
	Themes	6032
Regul.	General	4425
	Positive	9830
	Negative	6484

- "Events" refers to event trigger words
- "Relations" refers to a relational structure connecting typed events to cause–theme protein pairs by TEES

Institut für Maschinelle Sprachverarbeitung