Towards Confidence Estimation for Typed Protein-Protein Relation Extraction

Camilo Thorne and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Stuttgart, Germany
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

Systems which build on top of information extraction are typically challenged to extract knowledge that, while correct, is not yet well-known. We hypothesize that a good confidence measure for relational information has the property that such interesting information is found between information extracted with very high confidence and very low confidence. We discuss confidence estimation for the domain of biomedical protein-protein relation discovery in biomedical literature. As facts reported in papers take some time to be validated and recorded in biomedical databases, such task gives rise to large quantities of unknown but potentially true candidate relations. It is thus important to rank them based on supporting evidence rather than discard them. In this paper, we discuss this task and propose different approaches for confidence estimation and a pipeline to evaluate such methods. We show that the most straight-forward approach, a combination of different confidence measures from pipeline modules seems not to work well. We discuss this negative result and pinpoint potential future research directions.

1 Introduction

The ever increasing body of biomedical literature has motivated a growing interest over the past 20 years in natural language processing (NLP) and information extraction (IE) techniques to retrieve, organize and index the knowledge it contains (Rodriguez-Esteban, 2009; Subramaniam et al., 2003). It has also spurred a number of (shared) tasks and system competitions of which the best known are the BioNLP Shared Task[1] and the BioCreative challenge[2]. Relevant subtasks include named entity recognition (NER, Leaman and Gonzalez, 2008), entity linking and normalization to unique database identifiers (Zheng et al., 2014), event (EE, Björne and Salakoski, 2015) and relation extraction (RE, Tymoshenko et al., 2012; Airola et al., 2008; Choi, 2016). The overall goal is to identify biomedical entity mentions, disambiguate them w.r.t. biomedical databases and to identify mentioned biomedical relations and, crucially, *discover* new relations which are not available in structured resources yet.

When solving biomedical RE and IE tasks, the standard focus is to build systems that achieve high precision and recall at identifying *known relations* in gold standards or in biomedical databases and ontologies. This focus usually overlooks a key dimension for relation discovery: extraction *relevance* or *trust*. Indeed, when applied to new text in the form of, *e.g.*, recently published biomedical papers or papers from transversal domains such as bioinformatics, most discovered relations can arguably be expected to come up as "false", without being *per se* false – but unrecorded in gold standards. In other words, discovered relations fall under one of three categories: (1) plainly true relations (as per biomedical gold standards) (2) *interesting* relations that might be true or false. (3) plainly false relationships (as per biomedical gold standards). Our hypothesis is that a useful confidence measure estimates the quality of relations in this order, as we exemplify in Figure 1.

In such a scenario, rather than dismissing all such unknown (but interesting) relations, the goal is to return a ranking based on extraction *confi-*

[1] http://2016.bionlp-st.org/
[2] http://www.biocreative.org/

dence (Cullota and McCallum, 2004). Confidence typically refers to some kind of scoring – for instance a real number. This brings forth the problem of *confidence estimation*. While it is clear that the confidence of a relation extracted from biomedical text should be a function of the different sources of evidence on which it relies, it is unclear (Q1) how to define a global confidence estimator for biomedical relation extraction, and (Q2) how to evaluate it.

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

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

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 paper we outline a first attempt to answer questions (Q1) and (Q2) for the domain of protein-protein relations and events, focusing on approach S1. The main contributions of this paper are: (1) We build a distantly supervised RE pipeline based on BANNER (Leaman and Gonzalez, 2008) for NER, TEES (Björne and Salakoski, 2015) for EE and RE, and GNAT and Gnorm (Hackenberg et al., 2011; Wei et al., 2015) to link protein mentions to the STRING protein interaction database (von Mering et al., 2005), to distantly determine the truth and falsity of the discovered typed protein-protein relations. (2) We define confidence measures for each component of our pipeline and analyze their impact on relation prediction. (3) Finally, we propose and compare several global confidence estimators that aggregate over these scores.

2 Evidence Sources for Confidence

As said in the introduction, there are main sources of evidence for biomedical relation discovery and extraction, namely: prediction confidence (S1), graph analytics of the discovered relations (S2) and domain knowledge gathering (S3). We discuss these in the following.

S1 Modules (*e.g.*, NER or EE systems) in state-of-the-art systems are typically underpinned by

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

Figure 1: Discovered protein-protein (typed) relations. Notice how the bottom example is plainly false (it states a regulation among hormones), and the top one is plainly true (a known regulation). We assume the one in the middle to represent a more interesting result, as this fact is not in the STRING database; however, PubMed/MEDLINE abstract 28186089 mentions it ("(...) TIPE2 (...) downregulated (...) Snail2 (...)").

supervised classifiers that in addition to a prediction, return a probability (*e.g.*, logistic classifiers) or a so-called margin (*e.g.*, linear discriminant classifiers). We would expect interesting relations whose individual components (*e.g.*, entities and events) were identified with a higher score, to stand a higher chance of being true. To this end, one can employ confidence mixtures (Iversen et al., 2008; Dawid et al., 1995). Given k experts, each returning a confidence value $c_i \in \mathbb{R}$, $i = 1 \ldots k$, a *confidence aggregation* is a function $\varphi(\cdot)$ such that $c = \varphi(c_1, \ldots, c_k)$, where $c \in \mathbb{R}$ is the *global confidence* score. Global confidences thus aggregate partial confidences assigned to the partial tasks into which a complex task such as relation discovery can be broken down, to produce a global score. This method, the one actually described in this paper, can be seen as a baseline confidence estimator for biomedical RE.

S2 Graph-based confidence estimation techniques on the other hand rely on the graph-theoretical structure of extracted or discovered protein interactions and interaction networks. This makes sense because RE and EE systems (as the ones we rely on in this paper) actually return such graphs and interaction networks. In particular, one can leverage literature in biomedical and cross-domain *link prediction* (Lichtenwalter et al., 2010; Peng et al., 2017; And et al., 2003). Such tech-

niques generally aim at predicting new edges (binary relations) in entity graphs via techniques such as similarity computation, weighted by properties such as the centrality or prominence of the connected entities – a measure that can be seen as a kind of confidence score. One can also exploit shortest path statistics among detected proteins (lengths of the paths, number of paths), as, intuitively the more relations between two proteins, the more likely that a specific relation holds.

S3 Last but not least, external knowledge sources can be used to, alone or in combination with the previous two methods, derive confidence estimators for biomedical relation discovery. Indeed, the STRING database itself describes a network of protein interactions, which can be combined with the interaction network built at discovery and extraction time to gather further, gold standard graph theoretical evidence for discovered interactions. Another possibility is to use techniques from the knowledge base population and enrichment communities such as Wick et al. (2013), reasoning over domain constraints and on whether discovered interactions satisfy or violate them (*e.g.*, the third example in Figure 1 is clearly false because its arguments are not proteins). Finally, one can also gather *textual evidence*, using techniques borrowed from cognitive systems such as IBM Watson (Murdock et al., 2012), exploiting PubMed/MEDLINE itself to derive lexical evidence.

3 Experiments

In this section, we describe our confidence estimation experiments for typed protein-protein interaction extraction and discovery. We refer to an *ordered* triple $rel = (p_1, r, p_2)$, where r is an event or relation type denoting a *directed* relation (*e.g.*, an expression, an inhibition) between proteins p_1 and p_2 as *typed interaction*. Note that this task is a subtask of event extraction (Kim et al., 2009) and an extension to protein-protein interaction detection (PPIs), where we want to predict if a protein pair (in any order) interacts in some way (Choi, 2016; Airola et al., 2008). Our whole pipeline is depicted in Figure 2.

3.1 Datasets

We used two main datasets in our experiments: Firstly, a large subset of MEDLINE from May 1992 to May 2017 (PMIDs 1376980 to 28211214). We ignore languages other than English, and entries without abstract. We also disregarded abstracts that do not contain any mentions to protein or genes. This corpus consists of 40,911,675 tokens in 1,939,915 abstracts.

Secondly, to distantly evaluate discovered relations, we use the STRING database (von Mering et al., 2005), which describes protein-protein relations. STRING was built by integrating different databases (including the Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG)) and expert-curated text-mining-based information. STRING covers around 9.6 million protein entries and 1.3 billion interaction entries of 2031 unique organisms species. We focus on the subset of human proteins and their interactions. For network analysis, we use a Neo4j[3] graph database. From STRING, we use 20,458 unique genes/proteins with 6,013,567 unique typed interactions. They refer to 17,538 EntrezGene IDs.

3.2 Relation Extraction

To extract and discover relations in the MEDLINE subset, we rely on two well-known state-of-the-art systems for protein and gene detection and protein-protein event and relation extraction, namely BANNER and TEES (Leaman and Gonzalez, 2008; Björne and Salakoski, 2015).

BANNER is linear-chain conditional random field (CRF) NER system, that relies on an array of pre-trained models, dictionary and gold corpora for training and prediction. It uses the BIO format to spot the beginning (B) and constituent words (I) of a protein mention, and tokens that lie outside (O) mentions. For this paper, we use a gene detection model trained on the GNormPlus[4]gene gold corpus (Wei et al., 2015), which achieves 83 % F_1. Please note that we run BANNER separately from TEEs and realign the results in a separate step in the pipeline (see Figure 2).

TEES is a biomedical RE system, underpinned by a multiclass support vector machine (SVM). It relies on BANNER as a subcomponent to detect entities and on the BioNLP 2013 shared task data to estimate SVMs that detect (1) event triggering words and their GENIA event types: regulations, positive regulations, negative regulations, (de)phosphorilations (2) the arguments of

[3] https://neo4j.com/

[4] We used this model for consistency with the GN systems that we describe below, which also use models trained over this corpus created for the BioCreative II GN shared task.

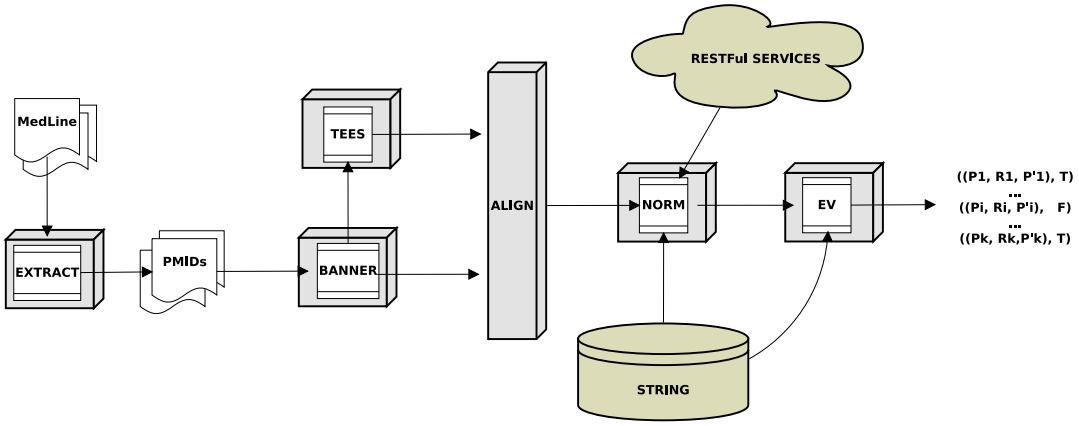

Figure 2: Overview of the relation extraction pipeline described in this paper (full pipeline).

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

Table 1: Event/Relation extraction statistics. "Events" refers to event trigger words, "Relations" refers to a relational structure connecting *typed* events to cause–theme protein pairs by TEES, as in Figure 3. All counts are unique counts.

the event or relation: its first argument (cause) and its second argument (theme). It can also detect complex event structures, event structures containing nested events, which we currently disregard. For each of its predictions, TEES returns a confidence value in the form of an SVM margin (distance of the trigger or protein to its separating hyperplane). TEES achieves 50.74 % F_1.

3.3 Relation Normalization

In order to verify if a candidate typed relation occurs in STRING, and to build (silver) standards for experimental analysis, we define a mapping from (M1) a protein mention p to a *canonical* form (norm(p)), *i.e.*, STRING protein unique identifiers (UIDs), and (M2) a relation/event type r to a STRING interaction type (ev(r)).

Protein matching Task (M1) is known in biomedical literature as the *protein normalization* task. It has been object of active research since the early 2000s, giving rise to the BioCreative Gene Normalization (GN) shared task. In this paper, we use two state-of-the-art GN systems, GNAT (Hackenberg et al., 2011), with a performance of 86.7 % F_1 and GNorm (Wei et al., 2015), with a performance of 86.4 % F_1. We denote this method by $\mathrm{gn}_\mathrm{N}(p)$, for each GN system N and protein mention p. GNAT and GNorm normalize gene/protein mentions to EntrezGene UIDs, which cover a subset of STRING protein UIDs.

Therefore, in order to increase normalization recall, we resorted to a disambiguation-based method. We relied on a STRING RESTful web-service[5] that returns, given a protein mention p, a list of possible STRING canonical matches, together with a gloss (a small textual definition), to build a custom bag-of-words disambiguation method, that ranks candidates by computing the cosine similarity of the gloss and the sentence in which the mention occurs. We denote this method by $\mathrm{lk}(p)$.

This gave rise to a protein normalization method for protein mentions p summarized by:

$$\mathrm{norm}_\mathrm{N}(p) = \begin{cases} \mathrm{gn}_\mathrm{N}(p), & \text{if } \mathrm{gn}_\mathrm{N}(p)\downarrow, \\ \mathrm{lk}(p), & \text{if } \mathrm{gn}_\mathrm{N}(p)\uparrow, \mathrm{lk}(p)\downarrow, \\ \mathrm{NA}, & \text{if } \mathrm{gn}_\mathrm{N}(p)\uparrow, \mathrm{lk}(p)\uparrow, \end{cases} \quad (1)$$

for $\mathrm{N} \in \{\mathrm{GNAT}, \mathrm{GNorm}\}$. By $\uparrow$ (resp. $\downarrow$) we mean that the method returns no canonical (resp. returns a canonical) STRING UID for mention p.

[5] https://string-db.org/cgi/help.pl?&subpage=api

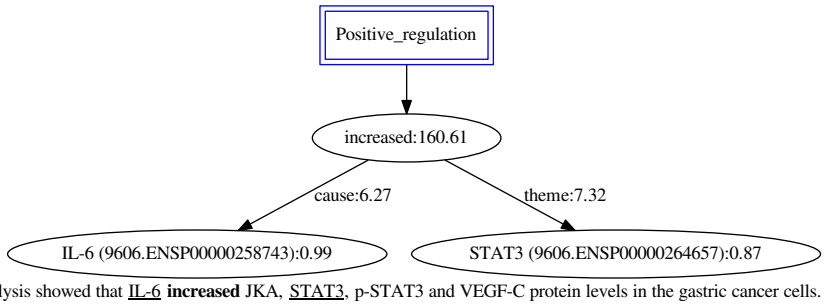

Figure 3: Protein-protein relational structure extracted by our pipeline for the first relation from Figure 1. The leave nodes represent the protein entities, labeled with their STRING UID and BANNER confidence. The internal node represents the event in which they participate as arguments, labeled with its TEES recognition confidence. The labels on its outgoing edges represent cause–theme TEES labeling of its protein arguments, and its TEES confidence. Finally, the root represents the predicted event type.

Note that GNAT and GNorm were tuned for distinct, though related GN subtasks, namely human GN and cross species GN, and can produce different results. If no normalization method returns a STRING UID, we consider the canonical protein for mention p undefined (NA).

Event type matching To deal with (M2), we relied on the other hand on a simple rule-based method, that maps the three GENIA event types returned by TEES GENIA event types to typed and directed interactions in STRING: protein inhibitions, activations and expressions. As a GENIA event type r may correspond to more than one STRING interaction, we map them to *sets* of interactions with

$$\text{ev}(r) = \begin{cases} \{\text{inhibitits}\}, & \text{if } r = R^-, \\ \{\text{expresses,activates}\}, & \text{if } r = R^+, \\ \{\text{expresses,activates}\} \cup \{\text{inhibits}\}, & \text{if } r = R. \end{cases} \quad (2)$$

In other words, a TEES relation type r (a regulation R, a negative regulation R^-, or a positive regulation R^+) will be mapped to (sets of) STRING protein inhibitions, expressions and activations.

Relation matching For $\text{N} \in \{\text{GNAT}, \text{GNorm}\}$, we determine a positive match for a candidate relation (triple) (p_1, r, p_2) if for at least a value $t \in \text{ev}(r)$ the triple $(\text{norm}_\text{N}(p_1), t, \text{norm}_\text{N}(p_2))$ occurs in the STRING database, negative otherwise. If $\text{norm}_\text{N}(p_i)$, for $i \in \{1, 2\}$, returns no canonical STRING UID, we discard the candidate altogether. As GNAT and GNorm produce different results, rather than aggregating results, we generated two separate silver standards, summarized by Table 2. Both cover around 1/2 of the original dataset of candidates and both are skewed towards negative matches.

normalizer	norm. relations	positive
GNAT*	11723	973
GNormPlus*	8639	544

Table 2: Silver standards obtained with our normalization methods. By the asterisk we mean the GN system plus our backoffs. By "norm. relations" we mean the number of relational structures for which protein pairs and event types could be normalized to STRING interaction types and protein UIDs and by "positive" to those that actually match interactions in STRING.

3.4 Confidence Estimation

In this subsection we describe our global confidence estimation models. These models aggregate confidence values returned by the key components of our pipeline, namely, BANNER and TEES for proteins, and event/event types, as shown in Figure 3.

Component-wise confidence For every RE candidate triple $rel = (p_1, r, p_2)$ we compute the following confidence values:

Entity-level (marginal) confidence (BANNER): we return the so-called *product gamma probability* (Cullota and McCallum, 2004) of protein theme (resp. cause) mentions p starting at position

t in an abstract with BIO labels $(s_t, \ldots, s_{t+k})$, defined by:

$$\text{cf}_{\gamma_t}(p) = \prod_{i=t}^{k} \gamma_i(s_i) \quad (3)$$

(resp., $\text{cf}_{\gamma_c}(p)$ for cause mentions) where $\gamma_i(s_i) = \alpha_i(s_i) \cdot \beta_i(s_i)/P(w_0, \ldots, w_i; \Lambda)$ is the normalized product of the forward and backward Viterbi lattice probabilities of label $s_i \in \{\text{B}, \text{I}\}$ at position i, computed from BANNER's underlying CRF model Λ, and w_i is a word token. This measure basically characterizes the likelihood that a given span of MEDLINE tokens is indeed a protein.

From TEES, we use event-level confidence based on the margins in the SVM, namely $\text{cf}_{ev}(r)$, $\text{cf}_c(p_1)$ and $\text{cf}_t(p_2)$ for event (type), cause, and theme predictions.

In summary, we use *five* component-wise confidence features for a relation triple $rel = (p_1, r, p_2)$, namely, the BANNER product gamma probability of theme proteins, the TEES margin value for theme proteins, the BANNER product gamma probability of cause proteins, the TEES margin value for cause proteins, and the TEES margin value for events/event types.

Confidence aggregation Different confidence sources will have a different impact on their global aggregate (Iversen et al., 2008; Dawid et al., 1995). Such impact can be quantified as a *weight*, set *a priori* or *a posteriori* by training a classifier over gold (or silver) data and plugging into the aggregates the inferred weights. For the experiments described in this paper we chose the latter, and trained a logistic classifier over our silver MEDLINE datasets (see the next subsection for a detailed description), and used its coefficients $\vec{\theta}$ to weight confidences.

We propose two fundamentally different methods to aggregate the separate confidence values to one measure for a triple rel.

The first method assumes that global confidence is a linear combination of component-wise confidences for a relation *rel* (cf_m, with $m \in \{\gamma_t, \gamma_c, ev, c(p_1), t(p_2)\}$), namely, their *(weighted) average*:

$$\text{cf}_{avg} = \frac{1}{5} \cdot \sum_m \theta_m \cdot \text{cf}_m \quad (4)$$

The second method assumes that each component-wise confidence is totally independent of each other (and hence independence for each pipeline prediction), and defines global confidence as a *(weighted) product*:

$$\text{cf}_{prod} = \prod_m \theta_m \cdot \text{cf}_m \quad (5)$$

We considered also unweighted versions of the confidence aggregators, by considering unit weights $\vec{\theta} = (1, 1, 1, 1, 1)^T$, that assign the same importance to all component-wise confidences.

Evaluation To evaluate our approach, we relied on a number of different strategies and combinations thereof. In particular, we split our two silver GNAT and GNORM datasets $\mathcal{S}_\text{N}$, into two disjoint train $\mathcal{T}_\text{N}$ and test $\mathcal{E}_\text{N}$ subsets. Given how unbalanced our data is, we, in addition, resampled the training sets by (1) oversampling positive matches, and (2) undersampling negative matches until we obtained two balanced training sets $\mathcal{S}_\text{GNAT}$ and $\mathcal{S}_\text{GNorm}$ each of 2000 relations. For testing, we kept a set of 1000 unresampled relations each.

To learn the weights $\vec{\theta}$ of the confidence aggregation models and hence of component-wise confidences, we trained a logistic classifier over each of our silver standards:

$$P(t' = 1 | \vec{c}) = (1 + \exp(-\sum_m \theta_m \cdot \text{cf}_m))^{-1} \quad (6)$$

where $t' = 1$ if *normalized* triple *rel* is in STRING, $m \in \{\gamma_t, \gamma_c, ev, c(p_1), t(p_2)\}$ and $\vec{c} = (\text{cf}_{\gamma_t}, \text{cf}_{\gamma_c}, \text{cf}_{ev}, \text{cf}_{c(p_1)}, \text{cf}_{t(p_2)})^T$. The parameters $\vec{\theta}$ were learned by maximizing the likelihood $\mathcal{L}(\vec{\theta}; \mathcal{T}_\text{N}) = \prod_j \pi(\vec{c}^{(j)}; \vec{\theta})^{r'^{(j)}} \cdot (1 - \pi(\vec{c}^{(j)}; \vec{\theta}))^{1-r'^{(j)}}$ via iterative weighted least squares. We tested each of the two ensuing logistic models over each of the two test datasets,

train dataset $\mathcal{T}$	test dataset $\mathcal{E}$	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

Table 3: Evaluation of logistic models over the different possible train/test combinations of our various silver standards. In bold, the combination with the best performance. We used the best model (gnat_train) to derive the logistic model (Equation 6) and the weights used in the weighted confidence aggregation models.

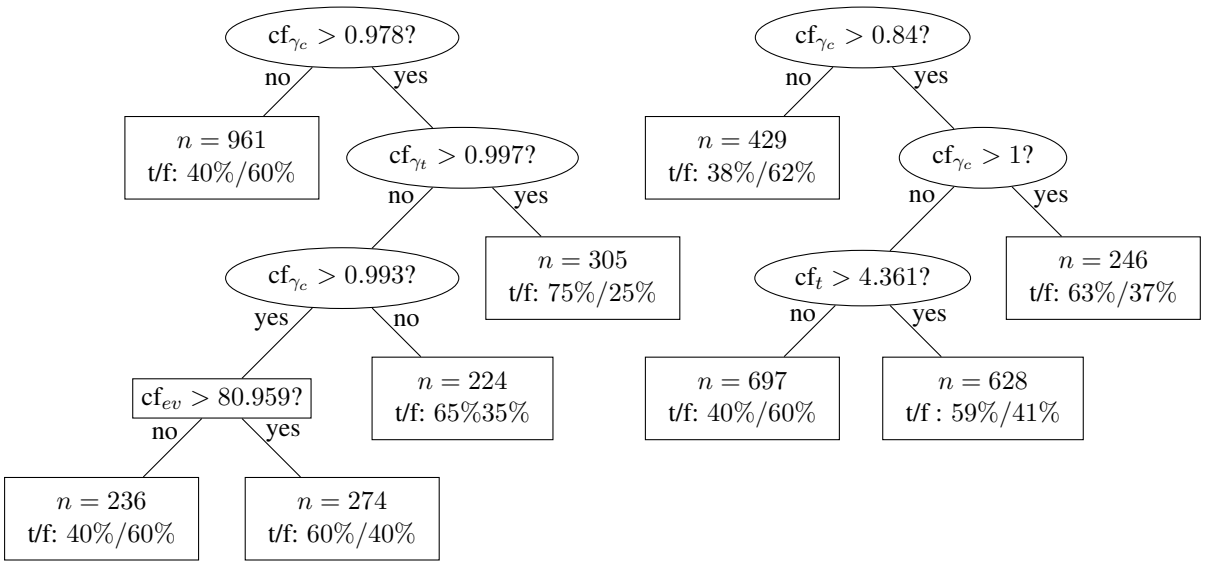

Figure 4: Right: J48 decision tree for the GNAT silver standard (training set). Left: J48 decision tree for the GNorm silver standard (training set). Nodes correspond to the component-wise confidence features defined in Section 3.4. The higher a component-wise confidence, the higher its information gain. Notice how, in general, we observe a higher gain for TEES confidence scores, plus some contribution coming from the BANNER confidence of theme proteins. We used for both models a pruning setup whereby we imposed each tree leave to contain at least 150 relations. In the visualization, the leaves describe also the distribution of positive (t) and negative (f) matches for each bin, and their size n (triples per bin).

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**

Table 4: Correlation-based evaluation of the confidence aggregation models. In bold, the model with the highest τ value. No test was statistically significant (although one came close to $p = 0.05$). In all cases, this indicates absence of correlation with linking judgments. Unweighted models were obtained by considering uniform weights (viz., $\vec{\theta} = (1, 1, 1, 1, 1)^T$).

feature	deviance	p-value
cf$_{ev}$	6.200	0.013
cf$_t$	**17.803**	$\mathbf{2.451 \cdot 10^{-05}}$
cf$_{\gamma_t}$	2.370	0.124
cf$_{\gamma_c}$	**22.667**	$\mathbf{1.926 \cdot 10^{-06}}$

Table 5: ANOVA/Analysis of deviance table for the best logistic model from Table 3 (χ^2-test). In bold, the features with greater impact, both statistically significant with $p < 0.01$.

and chose the model with the highest F_1, as seen in Table 3.

The confidence estimation models themselves were evaluated following a methodology proposed by Cullota and McCallum (2004) to evaluate entity-level confidence measures: measure the correlation between matching judgments and relation confidence. Ideally, one would expect a bias in confidence towards positive matches. In this paper, we considered Kendall's τ correlation, which is rank-sensitive and robust to ties.

Last, but not least, we used the logistic model and the balanced datasets to conduct an exploratory analysis on the component-wise confidence themselves, to understand which, from all of our pipeline's components has a bigger impact on global confidence estimation. To this end we relied on two separate methodologies: On the one hand, we conducted an analysis of variance/deviance[6] over the (optimal) logistic model's

[6]Logistic models are known in statistical literature as *generalized linear models*; in such cases rather than analyzing error variance as for linear models, one analyzes *deviance*, viz., prediction error.

features. On the other hand, we inferred two decision trees over our two training sets. Decision trees rank component-wise confidences cf_m, w.r.t. *information gain*. We used the J48 decision tree classifier[7], that discretizes continuous variables.

4 Results and Discussion

The results of our confidence aggregation experiments are summarized by Tables 3–5 and Figure 4.

As Table 3 shows, the best logistic model was obtained over the GNAT training dataset. Interestingly, the best result arose from cross-testing, when be tested it of the GNorm dataset test corpus. We conjecture that this might be due to a slightly better generalization capacity of the GNAT normalizer, as opposed to GNorm.

Regarding our confidence estimation models however, as Table 4 shows, our analysis returned no observable correlation (all τ values are close to zero), but without reaching statistical significance. Furthermore, of all estimators, the best (albeit by a very small margin) estimator was average, weighted confidence. We interpret this negative result to mean that aggregating confidences alone, disregarding: (1) the performance and/or confidence of the different normalizations methods (2) the structural properties of discovered relations, and (3) additional evidence gathered from external sources is simply not enough to define meaningful confidence estimators.

Finally, as shown by Table 5 and Figure 4 both the ANOVA and decision-tree/information gain analysis point out that the most informative features were the BANNER and TEES confidences for the arguments – the theme (2nd argument) and the cause (1st argument) – of protein-protein relational structures. Interestingly event (TEES) confidences do not seem to play a major role. This however seems consistent with the fact that TEES models are optimized for recognizing theme-event and cause-event pairs (by leveraging on the dependency parse tree of the sentence), a harder task than that of event recognition.

It suggests that while the aggregation of component-wise confidences is not a good global confidence estimator, including them as features of a more complex model encompassing a wider array of evidence sources might still be useful. It also suggests that normalization confidence – the last step in the pipeline – should be taken into account as the confidence values coming from protein recognition have the most impact.

[7]Weka (http://www.cs.waikato.ac.nz/ml/weka/) for our logistic and J48 models.

5 Conclusions & Future Work

In this paper we have proposed a confidence estimation methodology for biomedical protein-protein typed interaction discovery from PubMed/MEDLINE abstracts. Measuring confidence or trust is important because in this setting not all false positives – interactions that are not known to occur in biomedical databases – may be necessarily false. This sorting by confidence should satisfy key criteria, namely that true matches should be scored high, clearly false matches low, and "interesting" relations somewhere in between.

To do so, we have proposed a pipeline that builds upon state-of-the-art protein NER, protein-protein EE and RE and GN systems, to discover and distantly evaluate against the STRING database protein-protein typed interactions. Then, we have described a number of baseline confidence estimation techniques that aggregate the confidence prediction scores of the pipeline's components.

Our experiments and correlation analysis show that, while the prediction confidence of modules in later stages of the pipeline seems to influence more positive decisions, confidence aggregation is not enough to define estimation models satisfying the criteria mentioned. We conjecture that this is due to the fact that prediction confidence alone does not provide sufficient evidence to rank relations. Also, in this work, the confidence of normalization was not fully addressed. As further work we plan to focus on more complex evidence gathering methods.

Aknowledgements We thank Jörg Hackenberg for his help running and integrating GNAT into our pipeline, and Philippe Thomas for his comments on our event mappings. This work was supported by a grant from the Ministry of Science, Research and Arts of Baden-Württemberg to Roman Klinger.

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. All-paths graph kernel for protein-protein interac-

tion extraction with evaluation of cross-corpus learning. *BMC Bioinformatics* 9(Suppl 11):S2.

Dennis Wilkinson And, Dennis Wilkinson, and Bernardo A. Huberman. 2003. A method for finding communities of related genes. In *Proceedings of the National Academy of Sciences of the United States of America*. pages 5241–5248.

Jari Björne and Taio Salakoski. 2015. Tees 2.2: Biomedical event extraction for diverse corpora. *BMC Bioinformatics* 16(16):S4.

Sung-Pil Choi. 2016. Extraction of proteinprotein interactions (PPIs) from the literature by deep convolutional neural networks with various feature embeddings. *Journal of Information Science* .

Aron Cullota and Andrew McCallum. 2004. 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.

A. Dawid, M. DeGroot, J. Mortera, Roger Cooke, S. French, C. Genest, M. Schervish, D. Lindley, K. McConway, and R. Winkler. 1995. Coherent combination of experts' opinions. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* 4(2):263–313.

Jög Hackenberg, Marin Gerner, Maximilian Haeussler, Illés Solt, Conrad Plake, Martin Schroder, Graciela Gonzalez, Goran Nenadic, and Casey M. Bergman. 2011. The GNAT library for local and remote gene mention normalization. *Bionformatics* 27(19):2769–2771.

Edwin S Iversen, Giovanni Parmigiani, and Sining Chen. 2008. Multiple model evaluation absent the gold standard through model combination. *Journal of the American Statistical Association* 103(483):897–909.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP '09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*. BioNLP '09, pages 1–9.

Robert Leaman and Graciela Gonzalez. 2008. 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.

Ryan N. Lichtenwalter, Jake T. Lussier, and Nitesh V. Chawla. 2010. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '10, pages 243–252.

J. William Murdock, James Fan, Adam Lally, Hideki Shima, and Branimir Boguraev. 2012. Textual evidence gathering and analysis. *IBM Journal of Research and Development* 56(3):8.

Jiajie Peng, Kun Bai, Xuequn Shang, Guohua Wang, Hansheng Xue, Shuilin Jin, Liang Cheng, Yadong Wang, and Jin Chen. 2017. Predicting disease-related genes using integrated biomedical networks. *BMC genomics* 18(1):1043.

Raul Rodriguez-Esteban. 2009. Biomedical Text Mining and Its Applications. *PLoS Comput Biol* 5(12).

L. Venkata Subramaniam, Sougata Mukherjea, Pankaj Kankar, Biplav Srivastava, Vishal S. Batra, Pasumarti V. Kamesam, and Ravi Kothari. 2003. Information extraction from biomedical literature: Methodology, evaluation and an application. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. CIKM '03, pages 410–417.

Kateryna Tymoshenko, Swapna Somasundaran, Vinodkumar Prabhakaran, and Vinay Shet. 2012. Relation mining in the biomedical domain using entity-level semantics. In *Proceedings of the 20th European Conference on Artificial Intelligence*. ECAI '12, pages 780–785.

Christian von Mering, Lars J. Jensen, Berend Snel, Sean D. Hooper, Markus Krupp, Mathidel Foglierini, Nelly Jouffre, Martijn A. Huynen, and Peer Bork. 2005. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* 33(Suppl. 1):D433–D437.

Chih-Husuan Wei, Hung-Yu Kao, and Zhiyoung Lu. 2015. GNomrPlus: An integrative approach for tagging genes, gene families, and protein domains. *BioMed Research International* 2015(2015):ID 918710.

Michael L. Wick, Sameer Singh, Ari Kobren, and Andrew McCallum. 2013. Assessing confidence of knowledge base content with an experimental study in entity resolution. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*. AKBC@CIKM '13, pages 13–18.

Jin Guang Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2014. Entity linking for biomedical literature. In *Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics*. DTMBIO '14, pages 3–4.