BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 9 2012, pages 1253–1261
doi:10.1093/bioinformatics/bts125

Data and text mining

Advance Access publication March 13, 2012

Harmonization of gene/protein annotations: towards a gold standard MEDLINE

David Campos[1,*], Sérgio Matos[1], Ian Lewin[2], José Luís Oliveira[1] and Dietrich Rebholz-Schuhmann[2,*]

[1]University of Aveiro, IEETA/DETI, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The recognition of named entities (NER) is an elementary task in biomedical text mining. A number of NER solutions have been proposed in recent years, taking advantage of available annotated corpora, terminological resources and machine-learning techniques. Currently, the best performing solutions combine the outputs from selected annotation solutions measured against a single corpus. However, little effort has been spent on a systematic analysis of methods harmonizing the annotation results and measuring against a combination of Gold Standard Corpora (GSCs).

Results: We present Totum, a machine learning solution that harmonizes gene/protein annotations provided by heterogeneous NER solutions. It has been optimized and measured against a combination of manually curated GSCs. The performed experiments show that our approach improves the *F*-measure of state-of-the-art solutions by up to 10% (achieving ≈70%) in exact alignment and 22% (achieving ≈82%) in nested alignment. We demonstrate that our solution delivers reliable annotation results across the GSCs and it is an important contribution towards a homogeneous annotation of MEDLINE abstracts.

Availability and implementation: Totum is implemented in Java and its resources are available at http://bioinformatics.ua.pt/totum

Contact: david.campos@ua.pt; rebholz@ebi.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 1, 2011; revised on March 7, 2012; accepted on March 8, 2012

1 INTRODUCTION

In the last decades, we have witnessed an explosion of publicly available data, a consequence of the deep integration of computerized solutions in society. This rapid growth was also observed in biomedicine, with an overwhelming amount of data resulting from high-throughput methods, accompanied by a corresponding increase of textual information. For instance, MEDLINE contains over 18 million references to journal papers covering various biomedical fields (e.g. medicine and dentistry). MEDLINE and other biomedical resources are manually curated by expert annotators, in order to correctly identify biological entities (e.g. genes and proteins) and the relations between them (e.g. protein–protein interactions) from texts. However, manual annotation of large amounts of data has become a very demanding and expensive task. This situation naturally led to the development of computerized systems to perform these steps automatically.

The goal of information extraction (IE) is to extract structured and unambiguous information from unstructured data (e.g. natural language texts). Named entity recognition (NER) is a crucial initial task of biomedical IE, which intends to extract chunks of text that refer to specific entities of interest. It is one of the most important tasks, as the identified entities will be used as input to the following steps in the IE pipeline. However, gene and protein names have several characteristics that make difficult their identification in texts (Zhou *et al.*, 2004).

- many entity names are descriptive (e.g. 'normal thymic epithelial cells');
- two or more entity names sharing one head noun (e.g. '91 and 84 kDa proteins' refers to '91 kDa protein' and '84 kDa protein');
- one entity name with several spelling forms (e.g. 'N-acetylcysteine', 'N-acetyl-cysteine' and 'NAcetylCysteine');
- ambiguous abbreviations are frequently used (e.g. 'TCF' may refer to 'T cell factor' or to 'Tissue Culture Fluid').

Various systems were developed using different approaches and techniques, which can be categorized as being based on rules, dictionaries or machine learning. However, the most recent results clearly indicate that better performance can be achieved by using an ensemble of NER systems. As an example, the top five systems of the BioCreative II gene mention challenge (Smith *et al.*, 2008) used ensembles of NER solutions. In these systems, each approach identifies entity mentions with different characteristics and based on different knowledge. Moreover, most of the NER solutions are trained and/or tested in only one corpus, which is usually focused in a specific biomedical domain and provides specific gene/protein names and contexts. As a consequence, when the system is applied to a corpus from a different domain, the global performance drops significantly. Although this occurs with machine learning approaches, it also affects dictionary-based solutions, depending on the specificity of the used lexical resource. This is not only a consequence of the different domains, but also a result of the different annotation guidelines and their interpretation by human annotators. For instance, Colosimo *et al.* (2005) presents a study

[*]To whom correspondence should be addressed.

with 5000 abstracts, obtaining an inter-annotators agreement of 87% for Fly, 91% for Yeast and 69% for Mouse.

In summary, various sources of variability can be identified in human annotated corpora: specific biomedical domain or sub-domain of the documents, annotation guidelines and human annotators. Moreover, the different characteristics of NER systems introduce another source of variability for the harmonization task. As a result, considering the different underlying biological domains and the diversity of annotation types, combining gene/protein annotations from various systems is not a straightforward task. The harmonization method should take advantage of this variability, benefiting from the distinct background knowledge encoded by each system on each corpus, in order to obtain a more general solution, able to cope with the diversity of data found on a large-scale text repository such as MEDLINE.

This article presents Totum, a harmonization solution that addresses the problems of heterogeneous annotations. Section 2 presents the background of this work, and existent solutions to the combination problem. In Section 3 we present the proposed approach, and in Section 4 a comparison with state-of-the-art solutions, discussing the advantages and limitations. Finally, Section 5 presents some concluding remarks.

2 BACKGROUND

Nowadays, the annotation of biomedical documents is mainly performed manually by domain experts. Consequently, only small sets of documents have been manually annotated and made publicly available. The CALBC (Collaborative Annotation of a Large Biomedical Corpus) project intends to minimize this problem, providing a large-scale biomedical text corpus automatically annotated through the harmonization of several NER systems. This large corpus will contain annotations of several biological semantic groups, such as diseases, species, chemicals and genes/proteins (Rebholz-Schuhmann *et al.*, 2010).

The CALBC corpus is focused in the immunology biomedical sub-domain, which abstracts were collected from MEDLINE using the query ‘immunol*’. To generate the first version of this corpus, four different NER and normalization systems were used:

- System 1: implements a dictionary-based approach that takes morphological variability into consideration. It uses several publicly available resources, such as Swiss-Prot (Boutet *et al.*, 2007) and ChEBI (Degtyarenko *et al.*, 2008);
- System 2: applies a dictionary-based approach using Entrez Gene (Maglott *et al.*, 2005), Swiss-Prot, Genew (Wain *et al.*, 2004), GDB (Letovsky *et al.*, 1998) and OMIM (Hamosh *et al.*, 2005) as terminological resources;
- System 3: implements a machine learning approach using Conditional random fields (CRFs), receiving orthographic and morphological features as input. It also integrates a dictionary-based step to identify gene mentions that were missed by the CRF. This system was trained using data from several corpora, including GENIA (Kim *et al.*, 2003), PennBioIE (Kulick *et al.*, 2004), GENETAG (Tanabe *et al.*, 2005), PIR (Mani *et al.*, 2005) and AlMed (Bunescu *et al.*, 2005). In the end, the system performs normalization to provide identifiers for each gene/protein name;
- System 4: implements a dictionary-based solution, performing fuzzy matching and disambiguation to remove false positives.

These systems use different approaches to process the text, implementing different tokenization methods and/or strategies to deal with stopwords. Thus, we can argue that each system provides annotations with different characteristics, varying with the used techniques and resources. In order to take advantage of this variability, it is necessary to implement a method that will combine the several annotations, providing only one gene/protein name per chunk of text. To make this combination process possible, the several systems need to ‘speak and understand the same language’. IeXML (Rebholz–Schuhmann *et al.*, 2006) facilitates such task, by defining an XML standard for abstracts, sentences and annotations representation. Using this cross corpus standard, we can combine the heterogeneous annotations, either by unifying and/or intersecting the annotations, or through the implementation of machine learning-based solutions.

Intersection requires the agreement of at least two systems for accepting an annotation, which improves precision but degrades recall. For instance, Torii *et al.* (2009) presents a typical intersection solution, combining the annotations from four machine learning-based NER systems. Kuo *et al.* (2007) presents other interesting solution to combine two CRF models, by intersecting the top 10 adjacent annotations of each model and selecting the intersection with the best score. Union approaches, on the other hand, provide annotations performed by either one of the systems, improving recall but degrading precision. For instance, Ando (2007) performs the union of two CRF models, removing annotations that overlap with longer ones.

Intersection and union solutions are widely used, due to the simplicity and positive outcomes of such methods. For instance, in the BioCreative II gene mention task (Smith *et al.*, 2008), most of the participating systems that used an ensemble of systems applied intersection or union to combine the heterogeneous annotations. There are also solutions that use both techniques, Li *et al.* (2009) and Hsu *et al.* (2008) get the best performance by intersecting the annotations of similar models and then unifying the results of the intersections.

Machine learning-based solutions intend to learn the tokens’ boundaries by experience, using manually annotated data to this purpose. The annotated data provides curated knowledge, which makes the decisions more accurate and supported. However, what makes this solution unique is also its biggest limitation, because manually annotated data is sparse in comparison with unannotated data, which could limit the learning window. Wilbur *et al.* (2007) presents a machine learning solution to combine the annotations from the 19 NER systems that participated in the BioCreative II gene mention task, using a first-order CRF with a simple set of features (tokens and systems’ matches). Mika and Rost (2004) present a different approach, a weighted Support Vector Machine (SVM) to perform the harmonization of three SVMs and one dictionary-based system. Both solutions presented positive results, by obtaining better performance in comparison with each system used in isolation. Even the systems with low performance contributed to an improved harmonization result, adding variability that was not provided by other NER systems. However, both approaches trained the models on the same corpus being annotated, which demanded the use of a cross-validation strategy. During this process, both systems used almost the complete corpus for training purposes, which may create a model that is highly fitted to specific features of the training data. Consequently, the model could deviate from its target function, making it less effective when used in corpora with different characteristics.

The goal of the harmonization solution presented in this article is to provide automatic annotations for a large set of abstracts (almost one million) from MEDLINE, covering several sub-domains and organisms in immunology. Since machine learning-based solutions improve both precision and recall through the usage of curated knowledge, our goal is to develop a solution less dependent on a specific corpus and able to annotate most of MEDLINE abstracts with high accuracy.

3 METHODS

In order to develop a harmonization solution based on supervised machine learning, it is crucial to collect manually annotated data for the training procedures. To avoid the single corpus dependency, we used four of the biggest gold standard corpora (GSCs), which cover different biomedical domains and organisms:

- **FSUPRGE** (Hahn *et al.*, 2008): is a set of 3236 abstracts extracted from MEDLINE focused on gene regulation and expression, namely on regulatory events and all the components that are involved. The annotation process was semi-automatic, using a NER system that supports active learning (AL) to speed up the annotation process with no loss of annotation quality. During the AL process, the system selects the sentences that are expected to be more informative to the classifier, in order to be annotated by human experts;
- **JNLPBA** (Kim *et al.*, 2004): this corpus is a sub-set of the GENIA corpus, containing 2399 abstracts extracted from MEDLINE using the MeSH terms 'human', 'bloodcell' and 'transcription factor'. These abstracts were manually annotated based on the GENIA ontology, which makes each annotation independent of the context. The JNLPBA corpus includes only five classes (protein, DNA, RNA, cell line and cell type) from the 36 available in the GENIA ontology. Only the protein, DNA and RNA classes were used in this work;
- **PennBioIE**: is composed of several MEDLINE abstracts of two highly specialized biomedical sub-domains: the molecular genetics of cancer, and the inhibition of cytochrome P-450 enzymes. We use the oncology sub-set, which contains 1414 abstracts with annotations of proteins and RNAs;
- **GENETAG**: is composed of 20 000 sentences extracted from MEDLINE abstracts, not being focused in any specific domain. It contains annotations of proteins, DNAs and RNAs, which were performed by experts from biochemistry, genetics and molecular biology. This corpus was used in the BioCreative II challenge (Smith *et al.*, 2008), providing 15 000 sentences for training and 5000 sentences for testing. For this work, since the used systems implement normalization, it was necessary to find the original abstracts for each sentence. At the end, to avoid ambiguity problems, only 17 590 sentences were used.

Since each corpus is focused on a different goal and biomedical domain, the annotated entity names differ from corpus to corpus. The 10 most frequent annotations (Fig. 1) reflect this variability, presenting annotations that only appear in one corpus (e.g. overall, 'KIR' only appears on FSUPRGE), and annotations shared by the corpora with significantly different proportions of occurrences (e.g. 'NF-KappaB' is the most frequent annotation on JNLPBA, but only the eighth most frequent annotation on GENETAG and FSUPRGE). Moreover, the percentage of unique entity names is also different, which shows the entities sparseness and specificity of each corpus. For instance, since PennBioIE is focused on a very specialized sub-domain, the number of unique annotations in this corpus corresponds to just 18% of the complete set of annotations. On the other hand, since GENETAG is not focused on any sub-domain, 65% of its annotations are unique. Even when the proportion of unique entity names is not high, each corpus provides a unique set of names that is not available in any other corpora, delivering an extensive set of contexts where the gene/protein name could be found.

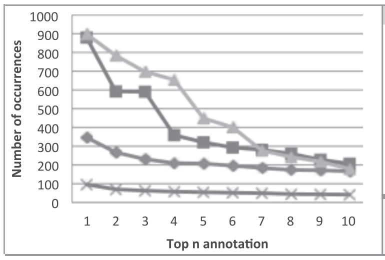
In order to obtain performance results, the corpora were divided into train and test sets. JNLPBA and GENETAG were already divided by the providers, using ≈17 and 25% of the data for testing, respectively. On the other hand, PennBioIE and FSUPRGE were not divided, so we left 30% of the data for testing purposes. Since each corpus is provided in a specific format, all the data were converted to the IeXML format, creating one large corpus with ≈6566 abstracts for training and 2242 for testing.

After annotating the corpus using the four systems described in Section 2 (S1–S4), there were several points of disagreement. Figure 2 shows some examples that reflect this variability. For instance, some systems include the organism name in the gene/protein names and others do not (Fig. 2: Example 1), which remains a point of active discussion among expert annotators. Other point of disagreement is the inclusion of the tokens 'protein' or 'gene' as suffix or prefix, making the systems to have a different behaviour (Fig. 2: Example 3). Finally, there is also variability regarding the inclusion of greek letters in the entity names (Fig. 2: Example 2).

The observed annotations variability also result in different performance results. Thus, it is important to understand the performance behaviour of the used systems, comparing them against typical and publicly available solutions with similar characteristics. Consequently, we annotated the four corpora using six solutions, three based on machine-learning and three based on dictionaries. Kuo *et al.* (2007) presents a CRF-based solution trained on GENETAG corpus, using orthographic and morphological features. It implements a bidirectional strategy, by combining two CRF models: one parsing the sentences from left to right (forward), and other parsing the sentences from right to left (backward). Another system is ABNER (Settles, 2005), which also applies CRFs trained on GENETAG corpus, using orthographic and morphological features. For the last ML-based solution, we trained ABNER on JNLPBA. Regarding dictionary-based solutions, the first one uses exact matching and BioThesaurus 7.0 (Liu *et al.*, 2005) as the gene/protein names dictionary after removing uninformative terms that are not used in the scientific literature. The identification of the terms uses orthographic variability (e.g. 'HZF[-]1' and '[Hh]zf[-]1') as described in Kirsch *et al.* (2006). The second solution is similar to the previous one, however it uses the Swiss-Prot subset of UniProt as the dictionary. After the matching process, basic disambiguation is performed through a specific term frequency associated with the term. The last solution also uses a disambiguation layer, but using BioThesaurus 7.0 instead.

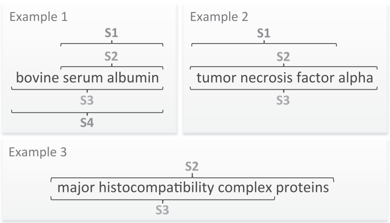
Figure 3 compares these six public solutions with the four systems used in this work (S1–S4), considering the four human annotated corpora and exact matching evaluation. In FSUPRGE two systems are above the average of public solutions, and the remaining are outside of the standard deviation (SD) range. Considering JNLPBA, one system is above the average, two are within the SD and one is outside that range. For GENETAG, one system is above the average of the public solutions, one is within the SD range and the remaining two are outside that range. Finally, all the used systems are above the average of the public solutions on PennBioIE. Remember that the ML-based solution that we use performs normalization, which does not happen on ML-based public systems. Thus, it is expectable that the ML-based public solutions provide better results, since the normalization step discards some names that were not possible to relate with unique identifiers.

A brief analysis indicates that our set of systems follow the average behaviour of the other solutions. In fact, a statistical analysis of the performance results show no significant difference between the two sets of systems. A two-tailed non-parametric Mann–Whitney test was performed, considering each corpus separately, resulting in *P*-values in the interval [0.2571; 0.9143].



FSUPRGE	JNLPBA	PennBioIE	GENETAG	
KIR	NF-kappa B	K-ras	insulin	
cytokine	IL-2	beta-catenin	Ras	
IFN-gamma	NF-kappaB	p53	Sp1	
cytokines	IL-4	ras	p53	
IL-6	AP-1	N-myc	AP-1	
insulin	transcription factors	N-ras	CAT	
kinase	TNF-alpha	polymerase	JNK	
NF-kappaB	IFN-gamma	MYCN	NF-kappaB	
RNA polymerase	cytokines	APC	MAPK	
TNF-alpha	IL-10	H-ras	CREB	
57261	**45635**	**17524**	**21677**	**Corpus annotations (#)**
3.76%	**8.79%**	**27.39%**	**2.65%**	**Top 10 (%)**
26.57%	**36.38%**	**17.75%**	**64.59%**	**Unique annotations (%)**

Fig. 1. Ten most frequent annotations on each curated corpus, reflecting the variability between the corpora. The percentage of unique annotations indicates the variability within each corpus. The highlighted annotation appear only on that specific corpus.



Example 1
S1
S2
bovine serum albumin
S3
S4
Example 2
S1
S2
tumor necrosis factor alpha
S3
Example 3
S2
major histocompatibility complex proteins
S3

Fig. 2. Examples of the annotations' variability provided by the four systems. S*n* indicates the annotation performed by system *n*.

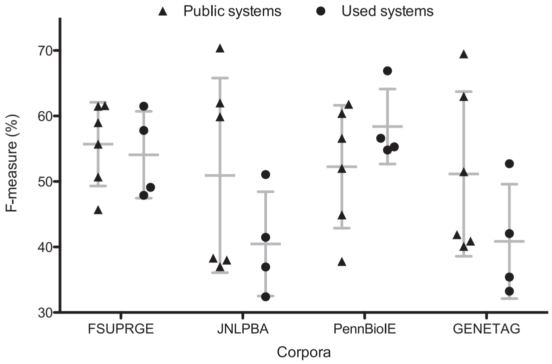


Fig. 3. Comparison of systems S1–S4 against publicly available solutions, considering the four GSC, namely the whole set of FSUPRGE and PennBioIE and only the test parts of JNLPBA and GENETAG. The bars illustrate the mean and SD of each set.

After annotating the corpora with the four systems, since each system uses its own tokenization technique, we created a tokenization method compatible with all strategies, allowing the creation of a single data source that contains the systems' contributions and gold standard annotations. Such data source is in a CoNNL-like format (Sang and De Meulder, 2003), where each line contains six columns: token, BIO[1] tags for each of the four systems, and gold standard BIO tag (Fig. 4).

Using the data in the CoNNL-like format, we are able to train a machine learning method, which can be supervised or semi-supervised. Semi-supervised solutions use both annotated and unannotated data, in order to obtain features of the entity names that are not present in the annotated data. Specifically for this task, the usage of unannotated data could contribute to a better abstract learning of the named entities boundaries. However, the application of such techniques is computationally heavy and could be performed as an extension to an equivalent supervised solution. Thus, we decided to use a supervised method, through the application of CRFs (Lafferty *et al.*, 2001), since it presents several advantages over other methods. At first, CRFs avoid the label bias problem (Lafferty *et al.*, 2001), a weakness of maximum entropy Markov models (MEMMs). On the other hand, CRFs also have advantage over hidden Markov models (HMMs), a consequence of its conditional nature that results in the relaxation of the independence assumptions (Wallach, 2004). Finally, to compare against SVMs, we analyzed the algorithms' complexity. For training, linear CRFs have quadratic complexity (Sutton and McCallum, 2006). However, such complexity increases exponentially with the used CRF order. On the other hand, SVMs training complexity could be cubic in the worst case (Burges, 1998). Regarding the prediction phase, both algorithms have linear complexity. Overall, we can argue that both algorithms provide positive outcomes, but SVMs could require more time to train complex models.

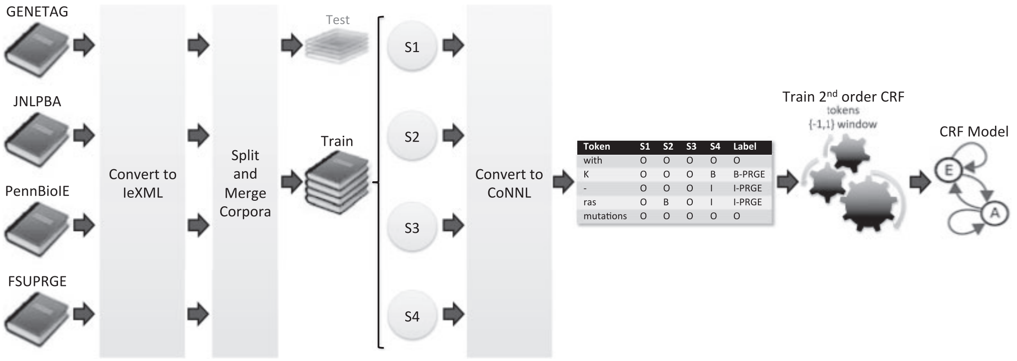
CRFs were first introduced by Lafferty *et al.* (2001). Assuming that we have an input sequence of observations (represented by X), and a state variable that needs to be inferred from the given observations (represented by Y), a CRF is a form of undirected graphical model that defines a single log-linear distribution over label sequences (Y) given a particular observation sequence (X). This layout makes it possible to have efficient algorithms to train models, in order to learn conditional distributions between Y_j and feature functions from training data. To accomplish this, it is necessary to determine the probability of a given label sequence Y given X, and consequently the most likely label. First, the model assigns a numerical weight to each feature, then those weights are combined to determine the probability of a certain value for Y_j. This probability is calculated as follows:

$$p(y|x,\lambda)=\frac{1}{Z(x)}\exp\left(\sum_j \lambda_j F_j(y,x)\right), \tag{1}$$

where λ_j is a parameter to be estimated from training data and indicates the informativeness of the respective feature, $Z(x)$ is a normalization factor and $F_j(y,x)=\sum_{i=1}^{n} f_j(y_{i-1},y_i,x,i)$, where each $f_j(y_{i-1},y_i,x,i)$ is either a state function $s(y_{i-1},y_i,x,i)$ or a transition function $t(y_{i-1},y_i,x,i)$. CRFs can be extended into higher-order models, which makes each y_i dependent on a fixed number o of previous variables $y_{i-o},\ldots,y_i$. Accordingly, the probability will consider not only the previous observation and its features, but o-previous observations and features.

Using the MALLET's CRF implementation (McCallum, 2002), we used a second-order CRF. At the beginning, we applied a simple set of features: tokens, systems annotations tags and a $\{-1,1\}$ window of tokens to model local context. In order to optimize the set of features, we performed several experiments using part-of-speech, stemming, different window sizes and different CRF orders. However, the performance always dropped and the

[1] The BIO encoding scheme is used to represent the annotations, were each token should be in beginning ('B'), inside ('I') or outside ('O') of an entity name.



Token	S1	S2	S3	S4	Label
with	O	O	O	O	O
K	O	O	O	B	B-PRGE
-	O	O	O	I	I-PRGE
ras	O	B	O	I	I-PRGE
mutations	O	O	O	O	O

Fig. 4. Illustration of the required steps to train the CRF model using the several corpora.

initial set of features was kept. Figure 4 illustrates the workflow to convert the data and train the model.

This model can change the annotations' boundaries, remove incorrect annotations, and generate new annotations in comparison with the ones provided by the systems. However, if the systems being combined perform normalization, i.e. provide unique identifiers for the entities, creating new annotations may not be desirable, since assigning identifiers to such annotations will not always be possible. In order to create a Totum solution that does not create new annotations, we changed the training portion of the GSC, removing manual annotations, i.e. replacing the corresponding entity labels by 'O', in those cases that were not identified as an entity by any of the four systems. Accordingly, we end up with a different GSC, adjusted to a different goal, which only contains gold standard entity labels where at least one system produced an entity output. This filtered version of the corpus contains 78% of the original gold standard annotations. Performing the CRF training in this new corpus, we get a new solution focused on changing the annotations boundaries or removing incorrect ones. Furthermore, we also built a post-processing filter to remove new annotations, which could happen (not in significant proportions) since the model uses tokens as features to learn the boundaries. In the end, we provide two different solutions: one, identified as Totum, optimized for harmonizing annotations from NER systems, and the other, identified as TotumID, guided towards harmonizing the annotations and respective identifiers provided by normalization systems.

4 RESULTS

4.1 Experimental setting

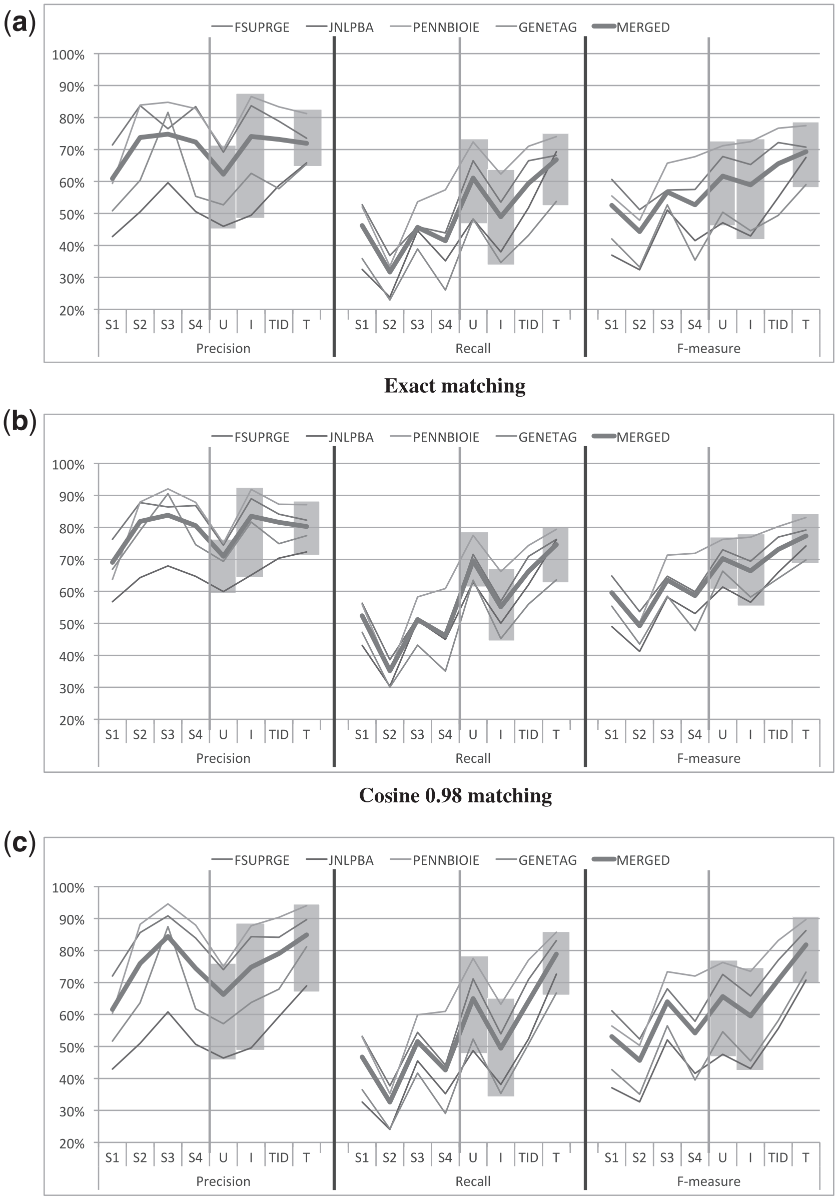
In order to obtain F-measure, precision and recall results that reflect the behaviour of the several solutions, we have applied four matching techniques: exact, nested and approximate using two different similarity thresholds. This detailed analysis is important since some post-NER tasks can be performed even if imprecise names are provided (e.g. relation extraction). Thus, we first perform exact alignment, which requires the boundaries of the entities to match exactly. Then, to perform approximate alignment, IDF (inverse document frequency) scores of the tokens were calculated using the corpus of one million MEDLINE abstracts about immunology. With these scores, we can calculate a similarity value using the cosine between the two vectors of the tokens. For example, if annotator A1 annotates the phrase Pa = 'T1 T2' and the annotator A2 the phrase Pb = 'T1 T2 T3', there is no exact match. Thus, we consider the IDF scores of each token $fx = \mathrm{idf}(Tx)$, calculating the cosine similarity between the vectors $v1 = <f1,f2,0>$ and $v2 = <f1,f2,f3>$. A match is accepted if $\cos(v1,v2)$ is equal or higher than a predefined value. In this work we use two different thresholds, 0.98 and 0.90. Finally, we also use nested alignment, in order to check when an annotation contains the boundaries of another.

To evaluate the performance of the two Totum solutions, we trained the harmonization model using the train part of the merged corpus, either with the original complete annotation set or with filtered annotations as explained in the previous section. This then allows us to check the results on the unaltered test part of each corpus and on the merged test set, providing accurate information regarding the behaviour of both solutions. Such solutions will be compared against the two most common and state-of-the-art harmonization approaches: intersection (two vote agreement) and union (one vote agreement).

4.2 Performance analysis

Figure 5 presents an overview of the results obtained in the experiments, focusing on the comparison of Totum against Union and Intersection. Annex 1 in the Supplementary Material presents detailed and precise results. Overall, the harmonization solutions present better results than the average of the four systems. On the other hand, when comparing with the best performing system, the two state-of-the-art approaches have a better performance only on FSUPRGE, PennBioIE and Merged. Both Totum solutions present better results, with the exception of TotumID on GENETAG, which is outperformed in exact matching.

Comparing the harmonization solutions, Totum significantly outperforms the other approaches. TotumID also presents better results than the two state-of-the-art solutions. Finally, union also presents better results than intersection. To analyze the improvements of both Totum approaches, we studied in detail the results achieved on the merged corpus, since it reflects better the global systems' behaviour. Moreover, there is no big difference between the results of the two approximate matching techniques. Consequently, we will only consider the cosine 0.98 alignment, which expresses better the process of discarding less informative tokens during the alignment. Therefore, comparing Totum with union, F-measure improvements of 7.61, 7.06 and 16.17% were obtained for exact (69.30%), approximate (77.34%) and nested (81.77%) matching, respectively. Against intersection, Totum achieved better performance by 10.34% for exact, 10.91% for approximate and 22.25% for nested alignment. Comparing TotumID



Exact matching

Cosine 0.98 matching

Nested matching

Fig. 5. Overview of the results achieved by systems S1–S4 and harmonization solutions on the test parts of each corpus and on the merged test set, considering exact, cosine 0.98 and nested matching. The filled boxes indicate the range of performance results for Union, Intersection and Totum, across the five test sets. (S*n*, System *n*; U, Union; I, Intersection; T, Totum and TID, TotumID).

with union, it presents an improvement of 3.89% (65.58%) on exact, 2.83% (73.11%) on approximate and 5.22% (70.83%) on nested matching. Against intersection, TotumID presented better results, with improvements of 6.62, 6.68 and 11.30% for exact, approximate and nested alignment, respectively. Considering the other corpora, Totum presents the best improvements on JNLPBA and less on PennBioIE. On the other hand, TotumID performs better on JNLPBA and worst on GENETAG, where it is slightly outperformed by union. Surprisingly, the best final results were achieved in the corpora that we used smaller amounts of data for the training procedures (FSUPRGE and PennBioIE), which is a direct consequence of the better results achieved by the systems. In summary, both Totum approaches present significant improvements in comparison with the two state-of-the-art solutions. However, the best results are achieved on nested alignment, which indicates that both Totum solutions provide longer names than the other approaches.

Regarding precision and recall, intersection presents better precision than union, since it uses two system votes to reach an agreement. On the other hand, union has better recall than intersection, because it only uses one vote. However, Totum presents better recall in all experiments. Thus, we can conclude that our solution is more sensitive than the other approaches, recognizing

Table 1. Number of annotations generated by each system and harmonization solution in comparison with manually curated data, considering the test parts of the corpora

	FSUPRGE	JNLPBA	PennBioIE	GENETAG	Merged
Gold	**17 181**	**6142**	**5285**	**5716**	**34 324**
System 1	12 691	4670	4636	4034	26 031
System 2	7562	2899	2109	2172	14 742
System 3	10 290	4599	3346	2725	20 960
System 4	9043	4269	3663	2687	19 662
Union	16 524	6447	5460	5233	33 664
Intersection	10 986	4722	3805	3165	22 678
TotumID	14 025	5349	4660	4127	27 866
Totum	16 431	6467	5074	4694	31 906

The highlighted values indicate the solution (and the harmonization method) that provided the higher number of annotations for each corpus.

more entity names correctly. Regarding precision, Totum always performs better on nested matching. However, in the other matching techniques, intersection presents better precision. This means that our approach has increased specificity in comparison with the used systems and union. Overall, Totum significantly improves recall (sensitivity) in comparison with other approaches, with a small drop of precision (specificity) in comparison with intersection. Thus, we can argue that our solution deals better with heterogeneous annotations and features, considerably improving recall and with no precision loss.

4.3 Annotations analysis

To understand the improved results provided by both Totum solutions, we have to study the generated annotations. Table 1 presents the number of annotations provided by the systems and harmonization solutions, when annotating the test parts of the GSC. System 1 provides more annotations than the other systems, which does not mean that it delivers the best results. Analyzing Figure 5, we can see that System 1 is outperformed by Systems 3 and 4 in most of the corpora. The same pattern is verified in the harmonization solutions, where union presents the largest amount of annotations in almost all corpora. However, Totum provides the best trade-off between precision and recall, generating approximately the same number of annotations as in the GSC, and with fewer mistakes.

To analyze the generated annotations, we developed a tool to compare the exact annotations provided by two solutions, in order to study the changes promoted by solution 2 against solution 1. We considered seven different categories of agreement and disagreement: Matched (the annotation is the same in the two solutions); New (the second solution adds an annotation that does not exists in the first one); Removed (the second solution removes an annotation provided by the first one); Add left (one or more tokens were added to the left side of the annotation); Add right (one or more tokens were added to the right side of the annotation); Remove left (one or more tokens were removed from the left side of the annotation); and Remove right (one or more tokens were removed from the right side of the annotation).

Additionally, for each annotation, we performed exact matching with the GSC to find if the change was correct or not. Figure 6 presents the results of comparing Totum with the other harmonization solutions, considering the merged test corpus. Overall, there is a high level of agreement between the several solutions, with an average of 85% correct annotations. The biggest sources of disagreement are *new*, *remove*, *add right* and *add left* categories. The addition of new annotations is one of the most important, since it adds annotations that were not considered by other approaches. On average, 61% of these annotations are correct according to the gold standard. Considering nested alignment, >72% of those new annotations are correct. The impact of this task is reflected in the comparison with the intersection approach (Fig. 6a). Ultimately, this task adds more true positives than false positives which contributes to a better precision, and reduces the number of false negatives contributing to a better recall. Another important category is *remove*, which discards false positives provided by other solutions. We can see the impact of this task in the comparison with union (Fig. 6b), where >76% of the deletions are correct. Adding tokens to the right side is the category where Totum performs worst. In average, it changes 40% of the annotations to correct, 40% to incorrect and 20% were wrong and remain wrong after the change. Finally, adding tokens to the left side presents a small positive contribution, by changing in average almost 50% to correct, 33% to incorrect and 17% that are still wrong after the change.

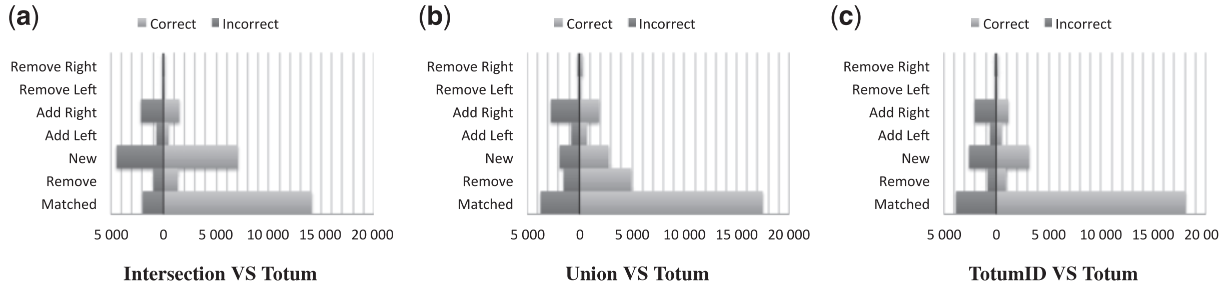
The only difference between our two solutions is the compatibility with normalization systems. Thus, there is a high level of agreement between the two approaches, differing only on the generation of new annotations (Fig. 6c). Remove left and right did not present any significant results, which reinforces the idea that our solutions provide longer names in comparison with other approaches.

Due to the generation of longer names, Totum considers that the suffixes and prefixes ‘gene’, ‘protein’, and the ones relative to species and greek letters, always make part of the annotations. However, this is not consistent with the annotations on all corpora. For instance, in comparison with intersection, Totum corrects ‘IL-2’ to ‘IL-2 gene’, but changing ‘RFX-B’ to ‘RFX-B protein’ makes the annotation to be wrong according to the gold standard. Regarding the addition of greek letters, it corrects ‘SDF1’ to ‘SDF1 alpha’. Our solution also adds organism names on annotations, converting ‘CD81’ to ‘mouse CD81’ and ‘AML1’ to ‘human AML1’, which are not correct according to the manually annotated data. Furthermore, Totum could consider the same chunk of text as being an annotation or not, which could be correct or not depending on the corpus. For instance, in comparison with intersection, Totum removes the annotation ‘CD4’, which is correct 51 times and wrong 22 times. The same occurs with the addition of the annotation ‘cytokine’, which is correct 88 times and wrong 67 times. This behaviour does not mean that Totum is completely wrong, since some corpora were annotated focusing in very specialized biomedical sub-domains, and consequently, some gene/protein names were discarded since they were not related with that sub-domain.

In summary, we can argue that Totum maintains a constant global behaviour, allowing the annotation of large amounts of data following the same guidelines, which were obtained training a machine learning model on several GSC.

5 CONCLUSION

In this article, we presented Totum, a new cross-corpus solution to harmonize heterogeneous gene/protein names from several NER or normalization systems. This approach uses CRFs to take advantage of the variability existent in several corpora from different domains,



Intersection VS Totum

Union VS Totum

TotumID VS Totum

Fig. 6. Comparison of the annotations provided by Totum against the other harmonization solutions.

learning the correct tags for the tokens and making the final result more precise and reasoned. In comparison with traditional harmonization solutions, which only allow fixing the annotations boundaries (by adding or removing tokens), our solution also allows creating new annotations or removing incorrect ones, which extends the traditional harmonization behaviour. Totum is also compatible with normalization systems (TotumID), preserving the provided identifiers and avoiding the creation of new annotations which would not have an identifier assigned.

Analysing the annotations provided by Totum, we concluded that improved results are achieved due to the deletion of incorrect annotations, the recognition of annotations discarded by other approaches, and the usage of the knowledge provided by the systems' annotations to create new entity names. In the end, we may conclude that Totum provides longer annotations than the other approaches, presenting a similar behaviour regarding the boundaries definition of the different gene/protein names.

The experiments demonstrate that both solutions outperform the most common and state-of-the-art approaches. Considering the merged corpus, and in comparison with an intersection approach, Totum presents F-measure improvements of up to 10.34, 10.91 and 22.25% on exact, approximate and nested alignment, respectively. Comparing against union, improvements of 7.61, 7.06 and 16.17% are achieved, regarding the same matching strategies.

Overall, Totum takes advantage of the annotations provided by several systems for different corpora, providing a solution that is not constrained to a specific corpus as the original systems are. In the end, the harmonized annotations provided by Totum present F-measures of 69.30, 77.34 and 81.77% for exact, approximate and nested alignment. With these results, we believe that this approach is a step towards a homogeneous annotation of MEDLINE abstracts, supporting several biomedical domains and organisms.

ACKNOWLEDGEMENT

We have to thank to the members of the CALBC project: David Milward, Peter Corbett, Udo Hahn, Ekatarina Buyko, Katrin Tomanek, Jan Kors, Erik van Mulligen and Ning Kang.

Funding: EU Support Action grant 231727 under the 7th EU Framework Programme within the theme 'Intelligent Content and Semantics' (ICT 2007.4.2) and 'Fundação Para a Ciência e a Tecnologia' (FCT) under the project PTDC/EIA-CCO/100541/2008.

Conflict of Interest: none declared.

REFERENCES

Ando,R. (2007) BioCreative II gene mention tagging system at IBM Watson. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop.* Madrid, Spain, pp. 101–103.

Boutet,E. *et al*. (2007) UniProtKB/Swiss-Prot. In: Edwards,D. (ed.) *Plant Bioinformatics: Methods and Protocols (Series: Methods in Molecular Biology)*, Vol. 406. Humana Press Inc., Totowa, NJ.

Bunescu,R. *et al*. (2005) Comparative experiments on learning information extractors for proteins and their interactions. *Artif. Intell. Med.*, **33**, 139–155.

Burges,C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.*, **2**, 121–167.

Colosimo,M. *et al*. (2005) Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*, **6** (Suppl. 1), S12.

Degtyarenko,K. *et al*. (2008) Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36** (Suppl. 1), D344–D350.

Hahn,U. *et al*. (2008) Semantic annotations for biology—a corpus development initiative at the Jena University Language & Information Engineering (JULIE) Lab. In *LREC 2008–Proceedings of the 6th International Conference on Language Resources and Evaluation.* Paris, France, pp. 28–30.

Hamosh,A. *et al*. (2005) Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33** (Suppl. 1), D514–D517.

Hsu,C. *et al.* (2008) Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics*, **24**, i286.

Kim,J. *et al*. (2003) GENIA corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**, 180–182.

Kim,J. *et al*. (2004) Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. Association for Computational Linguistics, Geneva, Switzerland, pp. 70–75.

Kirsch,H. *et al*. (2006) Distributed modules for text annotation and IE applied to the biomedical domain. *Int. J. Med. Inform.*, **75**, 496–500.

Kulick,S. *et al*. (2004) Integrated annotation for biomedical information extraction. In *Proceedings of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, Boston, Massachusetts, USA.

Kuo,C. *et al*. (2007) Rich feature set, unification of bidirectional parsing and dictionary filtering for high F-score gene mention tagging. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid, Spain, pp. 105–107.

Lafferty,J.*et al*. (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001)*. Williamstown, Massachusetts, USA.

Letovsky,S. *et al*. (1998) Gdb: the human genome database. *Nucleic Acids Res.*, **26**, 94.

Liu,H. *et al*. (2005) Biothesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103.

Li,L. *et al*. (2009) Integrating divergent models for gene mention tagging. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2009 (NLP-KE 2009)*, Dalian, China, pp. 1–7.

Maglott,D. *et al*. (2005) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, **33** (Suppl. 1), D54–D58.

Mani,I. *et al*. (2005) Protein name tagging guidelines: lessons learned. *Comp. Funct. Genom.*, **6**, 72–76.

McCallum,A.K. (2002) MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu

Mika,S. and Rost,B. (2004) Protein names precisely peeled off free text. *Bioinformatics*, **20** (Suppl. 1), i241.

Rebholz–Schuhmann,D. *et al*. (2006) IeXML: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules. In *Proceedings of BioLink, ISMB 2006.*Fortaleza, Brazil.

Rebholz-Schuhmann,D. *et al*. (2010) CALBC silver standard corpus. *J. Bioinform. Comput. Biol.*, **8**, 163–179.

Sang,E. and De Meulder,F. (2003) Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. Association for Computational Linguistics, Vol. 4, Edmonton, Canada, pp. 142–147.

Settles,B. (2005) Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, **21**, 3191.

Smith,L. *et al.* (2008) Overview of BioCreative II gene mention recognition. *Genome Biol.*, **9** (Suppl. 2), S2.

Sutton,C. and McCallum,A. (2006) An Introduction to Conditional Random Fields for Relational Learing. In: Getoor,L. and Taskar,B. (eds), *Introduction to Statistical Relational Learing*. MIT Press, Cambridge, MA.

Tanabe,L. *et al*. (2005) GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, **6** (Suppl. 1), S3.

Torii,M. *et al*. (2009) BioTagger-GM: a gene/protein name recognition system. *J. Am. Med. Inform. Assoc.*, **16**, 247.

Wain,H. *et al*. (2004) Genew: the human gene nomenclature database, 2004 updates. *Nucleic Acids Res.*, **32** (Suppl. 1), D255–D257.

Wallach,H. (2004) Conditional random fields: an introduction. *Rapport technique MS-CIS-04-21*, Vol. 50. Technical Report MS-CIS-04-21. Department of Computer and Information Science, University of Pennsylvania.

Wilbur,J. *et al*. (2007) Biocreative 2. Gene mention task. In *Proceedings of the Second Biocreative Challenge Evaluation Workshop*, Madrid, Spain, pp. 7–16.

Zhou,G. *et al*. (2004) Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, **20**, 1178–1190.