Bioinformatics Advances, 2024, **00**, vbae116
https://doi.org/10.1093/bioadv/vbae116
Advance Access Publication Date: 20 August 2024
Original Article

OXFORD

Data and text mining

CoNECo: a Corpus for Named Entity recognition and normalization of protein Complexes

Katerina Nastou [1,*], Mikaela Koutrouli [1], Sampo Pyysalo [2], Lars Juhl Jensen [1]

[1]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen 2200, Denmark
[2]TurkuNLP Group, Department of Computing, University of Turku, Turku, Finland

*Corresponding author. Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3, Copenhagen 2200, Denmark. E-mail: katerina.nastou@cpr.ku.dk

Associate Editor: Shanfeng Zhu

Abstract

Motivation: Despite significant progress in biomedical information extraction, there is a lack of resources for Named Entity Recognition (NER) and Named Entity Normalization (NEN) of protein-containing complexes. Current resources inadequately address the recognition of protein-containing complex names across different organisms, underscoring the crucial need for a dedicated corpus.

Results: We introduce the Complex Named Entity Corpus (CoNECo), an annotated corpus for NER and NEN of complexes. CoNECo comprises 1621 documents with 2052 entities, 1976 of which are normalized to Gene Ontology. We divided the corpus into training, development, and test sets and trained both a transformer-based and dictionary-based tagger on them. Evaluation on the test set demonstrated robust performance, with F-scores of 73.7% and 61.2%, respectively. Subsequently, we applied the best taggers for comprehensive tagging of the entire openly accessible biomedical literature.

Availability and implementation: All resources, including the annotated corpus, training data, and code, are available to the community through Zenodo https://zenodo.org/records/11263147 and GitHub https://zenodo.org/records/10693653.

1 Introduction

Improved deep-learning methodologies (Milošević and Thielemann 2023), such as Transformer-based models (Vaswani *et al.* 2017) pre-trained on large corpora, coupled with efforts toward annotation of biomedical text corpora (Kim *et al.* 2003, Pyysalo *et al.* 2007, Krallinger *et al.* 2008, Herrero-Zazo *et al.* 2013, Li *et al.* 2016, Luo *et al.* 2022, Luoma *et al.* 2023), have recently led to major advances in the field of information extraction. The aforementioned corpora have facilitated the development of methods that can accurately recognize a variety of entity types, including chemicals (Krallinger *et al.* 2015), genes/proteins (Smith *et al.* 2008), organisms (Luoma *et al.* 2023), and diseases (Doğan *et al.* 2014). Leveraging these resources, deep learning-based methods have made great progress in Named Entity Recognition (NER) tasks for these entities (Lee *et al.* 2020).

Despite their biological and pharmacological importance (Santos *et al.* 2017, Harding *et al.* 2024), there remains a conspicuous lack of a corpus specifically designed to evaluate NER and Named Entity Normalization (NEN) of protein-containing complexes (henceforth referred to as `complex`). While there are resources available for recognition of human `complex` names (Bachman *et al.* 2018), or annotation of mentions of type `complex` as part of a BioNLP Shared Task focused on Relation Extraction (Bossy *et al.* 2015), the development of a comprehensive corpus that facilitates the training and evaluation of dictionary-based or deep learning-based NER systems for `complex` names across multiple organisms is notably absent. This gap highlights a critical area of need, given the biological importance of complexes.

In this work, we propose the annotation of a new corpus, named CoNECo (Complex Named Entity Corpus), to serve the purposes of NER and NEN of protein complexes. Specifically, we have annotated 1621 documents with 2052 `complex` named entities, normalized in Gene Ontology (GO) (Ashburner *et al.* 2000, Aleksander *et al.* 2023). We have split the CoNECo documents into training, development, and test sets. We used the training and development sets to train a transformer-based tagger (Luoma *et al.* 2023) and to improve a dictionary-based tagger (Pafilis *et al.* 2013). We evaluated both taggers on the held-out test set, achieving F-scores of 73.7% and 61.2%, respectively, and used the best taggers for large-scale tagging of all PubMed abstracts and open-access articles from PubMed Central. All data used and produced in this project, along with the code to reproduce the results, are openly accessible via Zenodo and GitHub.

2 Methods

2.1 The CoNECo corpus

2.1.1 Document selection

The first step toward the generation of a corpus for the annotation of `complex` named entities was document selection. To benefit from existing resources, we initially focused our efforts on previously annotated corpora where work was

Received: May 29, 2024; **Revised:** July 10, 2024; **Editorial Decision:** July 27, 2024; **Accepted:** August 4, 2024

already done for the annotation of complex, which largely fitted the definition introduced above. As complexes play crucial roles in cellular signaling, we decided to improve the balance of the corpus by expanding it with documents related to this topic.

The selection of the documents consisted of three steps, which are detailed below:

1) ComplexTome corpus (Mehryary *et al.* 2024): All documents from ComplexTome, a corpus designed for training a deep learning-based relation extraction system for physical molecular interactions were selected. These documents contained named entity annotations for complexes, fitting with our definition, and were kept during annotation. Moreover, since this is a corpus for extraction of physical protein interactions, it inherently pertains to the topic of protein complexes.
2) Expansion with 100 additional Reactome abstracts: As the ComplexTome contains 300 documents used for the annotation of pathways in the Reactome pathway knowledgebase (Gillespie *et al.* 2022), we selected 100 extra abstracts from Reactome to increase the representation of signaling-related documents in CoNECo.
3) Event Extraction for Post-Translational Modifications corpus from the BioNLP 2010 workshop (Ohta *et al.* 2010): 234 signaling-related abstracts were selected from the 388 total abstracts in this corpus based on the existence of at least one post-translational modification event and more than one entity in a document.

The corpus was split into training, development, and test sets, keeping the original split for ComplexTome and assigning the extra 334 documents to keep a 60%/20%/20% document-level split. All documents in CoNECo were annotated within the BRAT rapid annotation tool (Stenetorp *et al.* 2012).

2.1.2 Named entity annotation

CoNECo is a corpus aiming at NER and NEN of protein-containing complexes. As such, there is a single annotated entity type, namely "protein-containing complex." To annotate and normalize such entities in text, we primarily built on GO (Ashburner *et al.* 2000, Aleksander *et al.* 2023). GO has a substantial representation of protein-containing complexes, rooted at the sub-ontology of the homonymous GO term (*GO:0032991*) in the GO Cellular Component (GO:CC) ontology, and contains 2103 terms. The definition of these entities can be summarized as *a stable set of interacting proteins which can be co-purified by an acceptable method, and where the complex has been shown to exist as an isolated, functional unit in vivo*. An example annotation is given in Fig. 1.

It should be noted that NEN is decoupled from NER, allowing for the annotation of entity mentions that align with the definition introduced above, even if they are not present in GO, and thus cannot be normalized. This approach enriches the corpus with a more comprehensive annotation, enhancing its usability across various applications. Our decision to normalize to GO, as opposed to other specialized resources like Complex Portal (Meldal *et al.* 2022), is driven by the widespread adoption and organism-agnostic nature of GO, making it a versatile and universally applicable standard in the community.

Our previous experience in annotating a deep learning-ready corpus for NER (Luoma *et al.* 2023), has highlighted the importance of boundary consistency in entity annotations. Extra care has also been taken in CoNECo, to ensure that the minimal span containing the full name of the entity mentioned in the text is annotated and that the marked span starts and ends on a boundary between an alphanumeric string and a non-alphanumeric character. When a complex name coincides with the name of a protein or its protein constituents or the name of a protein family, separated by any punctuation, the complex entity is not annotated. Moreover, no overlapping annotations between complex names are allowed, and the leftmost longest matching entity is annotated in cases where there are overlaps between different complex names.

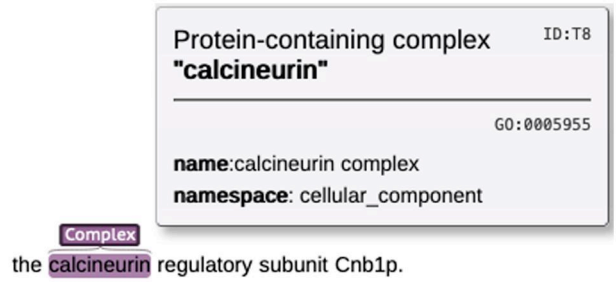

Figure 1. Illustration of a complex entity representation in CoNECo. The named entity ("calcineurin") has been identified by the annotators and normalized to the *calcineurin complex* term in GO (*GO:0005955*).

Another important factor to ensure the quality of the corpus is ensuring the consistency of mentions in text and the names and synonyms that these normalize to. We have performed a semi-automated check to evaluate this consistency, which produced a list of complex names that do not have a clear match in GO. All these names were manually checked against alternative Complex Portal (Meldal *et al.* 2022) and Reactome (Gillespie *et al.* 2022), to assess whether a link to GO entries could be obtained via them. If this was still not possible, but a link to alternative resources or the literature could be found, a comment was left in the corpus, and the entities were annotated solely for NER.

To evaluate the general quality of annotations, we randomly selected approximately 5% of abstracts from the training set and provided them to two curators in two rounds for independent annotation. Subsequently, we calculated the F-score of their agreement to determine the consistency of annotations and the overall corpus quality. The two annotators (K.N. and M.K.) have a background in biology, and have previous experience in corpora annotation (Luoma *et al.* 2023, Mehryary *et al.* 2024, Nourani *et al.* 2024, Nastou *et al.* 2024) and are knowledgeable on protein-containing complexes due to previous and on-going projects they have worked on.

Detailed annotation guidelines are provided through Zenodo and this link https://katnastou.github.io/annodoc-CoNECo/.

In addition to human annotators we have examined the possibility of using an annotation assistant for annotating documents for the CoNECo corpus by providing it with the annotation guidelines. To assess if this is possible, we decided to use the documents from the second round of inter-annotator agreement (IAA) and provide those as prompts to custom versions of ChatGPT 4.0. The entire process is described in Supplementary Section S1.

2.2 Dictionary-based NER and NEN

The JensenLab tagger (Jensen 2016) is a fast, dictionary-based method for the recognition and normalization of

several biomedical entity types. Since dictionary-based tagging is widely used and adopted in several biological resources—including for tagging protein and organism names for the influential database of protein–protein interactions STRING (Szklarczyk *et al.* 2023)—we wanted to ensure that the CoNECo corpus is suitable for evaluating dictionary-based methods, like the JensenLab tagger. We built a dictionary for tagger using all terms of GO: CC below "protein-containing complex." Additional `complex` names from Complex Portal were added to the dictionary following the process below:

1) The mapping between entries in Complex Portal and complex terms in Gene Ontology was obtained by properly filtering the mapping file available through the Complex Portal FTP: http://ftp.ebi.ac.uk/pub/databases/intact/complex/current/go/complex_portal.v2.gpad.
2) Complex Portal entries with multiple mappings to GO entries were ignored, after manual checking of a random sample of these entries, where several cases where synonyms seemed to match only one of the mapped entries were identified.
3) Complex Portal names that were already present in GO were removed.
4) Complex Portal names that included names of their protein constituents were removed either automatically or manually and the final list of extra names was obtained.

The dictionary was expanded with orthographic variants of the names coming from GO:CC and those added from Complex Portal. Specifically, forms where the word "complex" is not included were generated as well as plural and adjectival forms of existing names. Orthographic variation related to hyphenation or spacing is handled internally by the JensenLab tagger as described in Pafilis *et al.* (2013) and no such variants were thus added.

The JensenLab tagger software was run on the combined training and development set to identify potential issues with the dictionary generated following the process above and the dictionary files—including the blocklist—were updated accordingly.

After the dictionary build was complete, the dictionary-based tagger was run on the corpus test set and an evaluation of both NER and NEN was performed.

2.3 Transformer-based NER

Despite the widespread use of dictionary-based methods, it would be an omission not to assess whether the corpus is also useful for deep learning-based NER, considering that the majority of the biomedical text-mining community has now migrated to such methods, especially Transformer-based ones (Miranda-Escalada *et al.* 2023). We have selected the RoBERTa-large-PM-M3-Voc (hereafter RoBERTa-biolm) (Lewis *et al.* 2020) model, as it has been repeatedly shown to outperform all other models in biomedical tasks (Luoma *et al.* 2023, Miranda-Escalada *et al.* 2023).

We used the method described in (Luoma and Pyysalo, 2020) for training and evaluation of a RoBERTa-biolm model casting NER as a sequence labeling task. We attached a single fully connected layer on top of the Transformer architecture and fine-tuned the model to detect `complex` entities in the training data by classifying individual tokens in input samples. We selected hyperparameters by doing a grid search with three repetitions for each parameter set—to minimize the effect of initial random weights on evaluation scores (Mehryary *et al.* 2016)—training on the documents in the training set, and validating on those in the development set. The best mean entity level F-score on the development set was used to select the set of hyperparameters for training the model for evaluation on the test set. All training and development data are used in fine-tuning the model with the best set of hyperparameters. The process was repeated three times and the results shown in this article are expressed as a mean and standard deviation of the exact and overlapping match F-score. The latter allows us to compare the performance with the dictionary-based tagger. The character encoding for the files in the corpus is UTF-8, and the NER labeling scheme we used is IOB2 (Ratnaparkhi and Marcus 1998).

3 Results and discussion

3.1 Corpus statistics

CoNECo comprises 1621 documents with a total of 398 718 words (calculated using the BERT pre-tokenizer) and a total number of 2052 `complex` named entities 443 of which are unique names. From a first glance at these numbers, the density of annotated mentions in CoNECo (0.5%) is low in comparison to other biomedical NER corpora. Specifically, we compared the density to S1000 (Luoma *et al.* 2023) which was 2.4% (6328 total mentions in 262 293 words), BC2GM (Smith *et al.* 2008) with 4.3% (24 596 mentions in 569 912 words), BC5CDR (Li *et al.* 2016) with 4.4% and 3.5% density for drugs/chemicals and diseases, respectively (15 915 drug/chemical and 12 617 disease mentions in 360 373 words), and NCBI Disease (Doğan *et al.* 2014) with 3.7% density (6892 mentions in 184 552 words). In all cases, the number of words in the corpus is calculated using the BERT pre-tokenizer. This level of sparsity in CoNECo means that precision could be severely affected for any tagging method, because mentions are rare, and thus result in lower F-scores for protein-containing `complex` recognition in comparison to other biomedical NER tasks. There is an additional challenge for machine learning-based methods, as the number of positive examples in the training set might not be sufficient to learn the real-world data distribution, and this could significantly affect performance on the test data.

Table 1 shows detailed statistics for the training, development, and test sets, as well as the entire corpus.

We attained a 92.5% IAA after two rounds of IAA on 50 documents, showcasing the high quality of CoNECo. Additional guidelines were added between the first and second rounds, which allowed us to reach an above 90% IAA. In the first round, our IAA was 78.8% and the inconsistencies derived mostly from the fact that one annotator annotated `complex` names when those coincided with the names of protein families, while the other did not. A new rule was thus added before the second round to clarify that in such cases `complex` names should not be annotated. After checking the differences in the second round all inconsistencies were attributed to annotation errors, which were fixed, and there was no need for a further update of the annotation guidelines. Moreover, the annotators were in full agreement about the normalization of entities to GO: CC from the first round of IAA. Our experiments using ChatGPT 4.0 (https://openai.com/chatgpt/) as an annotator, produced mediocre results in comparison to human annotators, with the best

Table 1. CoNECo corpus statistics.[a]

Category	Total documents	Total mentions	Unique names	Total normalized mentions	Unique normalizations
Train set	983	1360	330	1310	138
Devel set	320	409	126	387	63
Test set	318	283	103	279	58
Total	1621	2052	443	1976	183

[a] The number of unique names and normalizations in the train, devel, and test sets does not sum up to the total number of unique names and normalizations in the corpus, as there can be different unique names and normalizations in each set, which are duplicates once the entire corpus is considered.

model attaining an F-score of 63.7%. Results of these experiments are presented in detail Supplementary Section S1.

3.2 Dictionary-based NER

The original dictionary was built using all terms of GO: CC below "protein-containing complex" (GO:0032991) and 989 extra names were added from Complex Portal. We used this dictionary to run the JensenLab tagger on the combined training and development sets and quickly identified several issues with GO: CC terms that clash with our annotation guidelines. Specifically, 78 GO: CC terms were excluded based on manual review, as they do not represent individual complexes but rather groups of complexes with a common function (e.g. *GO:1902494, catalytic complex*) or localization (e.g. *GO:0140513 nuclear protein-containing complex*). The full list of excluded GO: CC terms is provided in the annotation guidelines documentation (https://katnastou.github.io/annodoc-CoNECo). On top of these, more specific issues were identified with clashes between `complex` names and protein names or `complex` names and protein family names, that required an update to our list of filtered names. After all the dictionary cleaning, we ran the tagger on the combined training and development set and obtained an F-score of 68.0% (precision 74.5%, recall 62.5%). All dictionary files to run the JensenLab tagger are available through Zenodo.

The dictionary was used to run the Jensenlab tagger on the CoNECo test set. Since the JensenLab tagger finds left-most longest matches of the names in its dictionary, the **overlapping matching** criterion is used to evaluate both this and the Transformer-based tagger. The tagger reached a precision of 57.5% (185/322), a recall of 65.4% (185/283), and an F-score of 61.2% on the test set. Our hypothesis that the sparsity of mentions in this corpus could severely affect the precision of the tagger in the test set, turned out to be true, as there is a 17.5% drop in precision from the dictionary run on the combined training and development sets to the test set. A detailed analysis and comparison of the recognition errors produced by both methods is presented in Section 3.4 below.

In our evaluation of the normalization for JensenLab tagger, an impressive 94.6% F-score was achieved, since 175 out of 185 normalizations for matching spans were found to be identical, underscoring the efficacy of our method in accurately mapping entities to their corresponding standardized terms. The 10 instances where mismatches occurred are documented in Supplementary Table S1. It appears that the majority of these mismatches (80%) can be attributed to the tagging of names that are either more general or more specific than the expected normalization. This outcome, while not perfectly aligned with the intended normalization, is still acceptable. Among the remaining cases, one instance involves a conflicting narrow synonym in Gene Ontology that is less preferred to the annotated normalization and an annotation error, where we neglected to assign a normalization to a specific entity.

Table 2. NER performance comparison between dictionary-based and transformer-based methods on the CoNECo test set.

Method	Precision	Recall	F-score
Dictionary-based NER	57.5%	65.4%	61.2%
Transformer-based NER	66.3%	83.0%	73.7%
Union of matches	57.3%	87.0%	69.0%
Intersection of matches	89.6%	50.8%	64.8%

3.3 Transformer-based NER

We fine-tuned a pre-trained RoBERTa-biolm using the training set of the CoNECo corpus and used the development set to identify a set of hyperparameters where the mean average F-score is the highest on the task of `complex` NER. We obtained best results with models trained for 9 epochs, with a learning rate of 2E-5, a batch size of 2, and a maximum sequence length of 128. The full set of hyperparameters tested is presented in Supplementary Table S2. The F-score is 73.1% (std = 0.27%), the precision is 80.5% (std = 1.18%), and the recall is 67.0% (std = 0.91%). A model trained using this set of hyperparameters is available through Zenodo.

We then tested the performance of this model against the CoNECo test set and obtained an F-score of 73.7%, a precision of 66.3% (234/353), and a recall of 83.0% (235/283). The performance of this model on the dev and test sets is similar, in terms of F-score, but we can see once more that the precision is affected in the test set run (14% drop). A comparison between this method and the dictionary-based method is presented in Table 2. Moreover, as we inferred from the sparsity statistics, the results for this task are worse in comparison to other biomedical NER tasks, where Transformer-based methods tend to reach above 90% F-score (Lee *et al.* 2020, Luoma *et al.* 2023, Miranda-Escalada *et al.* 2023).

We made an additional comparison to check the overlap of matches between the two NER methods in the CoNECo test set. We found 186 overlapping matches, 167 matches unique to the Transformer-based tagger, and 136 matches unique to the JensenLab tagger. If we make a union of all the matches and evaluate against the CoNECo test set we get a precision of 57.3%, a recall of 87.0%, and an F-score of 69.0%. If on the other hand, we use only the consensus of the two methods we get the following results: precision = 89.6%, recall = 50.8%, and F-score = 64.8%. Including all results leads to an increase in recall while taking the more strict approach and using only the consensus of the two methods leads to an almost 90% precision, but a huge drop in recall to ~50%. In summary, if one needs normalization, the consensus of the two methods will yield better results than the dictionary-based tagger alone. On the other hand, if normalization is not needed best results are obtained by using the

Transformer-based method alone. In the next section, we look into the errors produced by both methods and further assess the strengths and shortcomings of each approach.

3.4 Error analysis

We have looked at the errors produced by both methods in detail and grouped them in categories as presented in Table 3. Supplementary Tables S3 and S4 provide a detailed overview of errors for the JensenLab tagger and Transformer-based tagger, respectively.

There are three error categories that affect both methods. "Annotation errors" result in all cases in False Positives and thus refer to cases where the annotators have missed an entity annotation, that is otherwise correctly recognized by either the dictionary- or Transformer-based tagger (e.g. annotators did not annotate *APC/C* in the following sentence from document 21700221_12 in the corpus "We found that mutation of charged residues to alanine interfered with the *APC/C*-dependent ubiquitination and degradation of geminin"). Recalculation of the performance of the JensenLab tagger without these errors results in a 1.5% increase in F-score (62.7%) due to an increased precision of 60.3%. The result is even more prominent for the Transformer-based tagger, where fixing the identified annotation errors results in 77.6% F-score due to a 72.8% precision. Both approaches face issues with "ambiguous names" that can denote either a protein-containing complex, a gene/protein or a family or large groups of proteins (e.g. *PCNA* which is not annotated in CoNECo, but detected by the Transformer-based tagger in this sentence from document 24768535: "Modification of *PCNA* by ISG15 plays a crucial role in termination of error-prone translesion DNA synthesis"). According to our annotation guidelines, protein and family named entities have priority over `complex` during annotation, and thus the ambiguous entities have remained unannotated. However, given the inherent ambiguity of biomedical entity names in such cases, it becomes evident how these can severely impact the effectiveness of either method. One special subcategory of ambiguity is "part-of longer entity" errors, where longer more specific entities are either not part of the dictionary or have failed to be recognized by the Transformer-based method [e.g. *IKK* and *IkappaB kinase* are detected by both taggers but are not annotated as they are part of the longer protein family name *IkappaB kinase (IKK)-related kinases*] in the following sentence from document 17015689: "Involvement of the *IkappaB kinase (IKK)-related kinases* tank-binding kinase 1/IKKi."

The other error categories can only be attributed to one of the two methods. "Dictionary errors" are responsible for approximately half of the mistakes made by the dictionary-based tagger. This error category indicates that the effectiveness of dictionary-based approaches hinges on the quality of their source dictionary. The main impact of dictionary errors is on recall, primarily due to the names missing from the dictionary. It is important to acknowledge that this problem stems from the fact that normalization is an inherent part of dictionary-based NER, and these cases would also challenge the Transformer-based method if results were normalized, and normalization relied on Gene Ontology as a source. Although there are strategies to mitigate these problems, the absence of names from the normalization source is a severely limiting factor. An example of such an error that results in a false negative can be seen in document 12815069, where in the sentence "Elimination of endogenous Omi by RNA interference abolishes c-IAP1 cleavage and desensitizes cells to apoptosis induced by *TRAIL*.," *TRAIL* is not detected as a `complex`, as the synonym is missing from GO. False positives are most commonly caused by tagging consecutive protein names as `complex`. For example in document 19228687, *MSH2/MSH6* and *MSH2/MSH3* are tagged as single `complex` entities in the sentence "MutSalpha (*MSH2/MSH6*) and MutSbeta (*MSH2/MSH3*) are eukaryotic mismatch recognition proteins." However, according to the annotation guidelines, these are not annotated as their constituent protein named entities have prevalence over the `complex` entity annotation.

While "discontinuous entity" could affect either method, in CoNECo it has only affected the JensenLab tagger. Such an error is exemplified in this excerpt extracted from document 20427570: "Deltamba1/Deltamdm38 mitochondria show severe defects in *complexes III* and *IV* of the respiratory chain," where the Transformer-based tagger has recognized *complexes III* and *IV* as a `complex`, while this is not possible for the dictionary-based method. In general "discontinuous entities" pose significant challenges for dictionary-based methods and are one of the issues that deep learning-based methods can resolve better. Finally, "Unidentified model error" is an error category that affects only the Transformer-based tagger and refers to errors for which we could not properly interpret what has led the model into making these mistakes (e.g. not detecting *DDR* in this sentence: "The manner in which RDM1 acts in both the *DDR* complex and as a factor bridging DRM2 and AGO4 remains unclear" from document 24498436).

3.5 Large-scale tagging

Results on tagging of over 36 million PubMed abstracts (as of January 2024) and 6 million articles from the PMC open

Table 3. Error analysis for the JensenLab and the transformer-based taggers.

	Total		FN		FP	
Error categories	**dict-based**	**TF-based**	**dict-based**	**TF-based**	**dict-based**	**TF-based**
Annotation error	9	26	0	0	9	26
Ambiguous name	100	124	0	35	100	89
Dictionary error	112	–	97	–	15	–
Discontinuous entity	1	0	1	0	0	0
Part-of longer entity	13	4	0	0	13	4
Unidentified model error	–	13	–	13	–	0
Total	235	167	98	48	137	119

FN: false negative; FP: false positive; dict-based: dictionary-based tagger; TF-based: transformer-based tagger.

access subset (as of November 2023) for both the JensenLab tagger and the Transformer-based method are provided via Zenodo. Tagging with the JensenLab tagger yielded 26 657 127 `complex` matches, covering 42 120 unique surface forms. Due to the nature of JensenLab tagger, these matches all map back to specific GO:CC cellular component entries, as GO was used to generate the dictionary for tagging. Considering the impressive results we got for normalization on the CoNECo test set (F-score = 94.6%), we expect that mapping to be accurate also in the large-scale tagging. Tagging with the Transformer-based model produced a total of 19 654 180 matches, out of which 105 242 names are unique.

These results suggest that many synonyms are missing from the reference resources (GO, Complex Portal) used to generate the dictionary of `complex` names, which leads to a lower total count compared to the deep learning-based method. To investigate this, we compared the overlap of matches between the two systems. The total number of common matches between the two is 8 544 366—which leaves 18 112 761 `complex` matches found only by JensenLab tagger and 11 109 814 found only by the Transformer-based method. The most frequent common matches are unambiguous `complex` names, as expected from the consensus between two completely different approaches. The matches produced only by the Transformer-based tagger confirm that GO and Complex Portal lack synonyms for known complexes, since notable omissions like "major histocompatibility complex" as a synonym for "MHC protein complex" (GO:0042611) or "TFIIH" as a synonym for "transcription factor TFIIH holo complex" (GO:0005675) result in many false negatives in the scientific literature. Additionally, several ambiguous names (e.g. SNARE), which correspond to both `complex` and protein/protein family names, are unique to the Transformer-based tagger large-scale results because they are intentionally blocked in the dictionary-based tagger. While the results unique to the JensenLab tagger in many cases are correct, they also include the types of false positives already described in the error analysis. When looking at the results of tagging the entire scientific literature more cases of clear false positives due to dictionary errors are observed. Some notable examples are names like "gait" or "coma" which are tagged due to being synonyms for "GAIT complex" (GO:0097452) and "COMA complex" (GO:0000817), respectively. These could be a good starting point for manual curation to improve the block list of the dictionary-based tagger. The files with the frequencies for all matches for both methods can be found on Zenodo, as well as the common and unique matches for each method.

4 Conclusions

In this work, we present CoNECo, the first corpus specifically designed for training and evaluating NER methods for `complex` recognition. The entities in the corpus are normalized to Gene Ontology, allowing for the evaluation of NEN methods on top of NER. CoNECo consists of 1621 documents and 2052 named entity annotations (1976 normalized to GO), 443 of which are unique. Despite the sparsity of the corpus, it has been shown that it can be used for training and evaluating Transformer-based language models (F-score = 73.7%, 77.6% correcting for annotation errors), as well as dictionary-based methods (F-score = 61.2%, 62.7% correcting for annotation errors). Error analysis results have shown that ambiguity in entity names is the main issue faced by both dictionary- and deep learning-based methods. Moreover, dictionary-based methods are severely affected by the omission of a plethora of synonyms in reference resources like GO and Complex Portal.

Acknowledgements

We thank the CSC—IT Center for Science for generous computational resources.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

No competing interest is declared.

Funding

This work was supported by the Novo Nordisk Foundation [NNF14CC0001, NNF20SA0035590 to M.K.], the Academy of Finland [332844], and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie [101023676 to K.N.].

Data availability

All data utilized in this project, the code to replicate the results and large-scale tagging results of the biomedical literature are available under open licenses from Zenodo https://zenodo.org/records/11263147 and GitHub https://zenodo.org/records/10693653.

References

Aleksander SA, Balhoff J, Carbon S *et al.*; Gene Ontology Consortium. The gene ontology knowledgebase in 2023. *Genetics* 2023;**224**:iyad031.

Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.

Bachman JA, Gyori BM, Sorger PK. Famplex: a resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining. *BMC Bioinformatics* 2018;**19**:248.

Bossy R, Golik W, Ratkovic Z *et al.* Overview of the gene regulation network and the bacteria biotope tasks in bionlp'13 shared task. *BMC Bioinformatics* 2015;**16 Suppl 10**:S1–16.

Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Inform* 2014;**47**:1–10.

Gillespie M, Jassal B, Stephan R *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res* 2022;**50**:D687–92.

Harding SD, Armstrong JF, Faccenda E *et al.* The IUPHAR/BPS guide to pharmacology in 2024. *Nucleic Acids Res* 2024;**52**:D1438–49.

Herrero-Zazo M, Segura-Bedmar I, Martínez P *et al.* The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions. *J Biomed Inform* 2013;**46**:914–20.

Jensen LJ. One tagger, many uses: illustrating the power of ontologies in dictionary-based named entity recognition. In: *Proceedings of the Joint International Conference on Biological Ontology and BioCreative*, Corvallis, Oregon, United States, 2016:1747.

Kim J-D, Ohta T, Tateisi Y *et al.* Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* 2003;**19 Suppl 1**:i180–2.

Krallinger M, Leitner F, Rodriguez-Penagos C *et al.* Overview of the protein-protein interaction annotation extraction task of biocreative II. *Genome Biol* 2008;**9 Suppl 2**:S4–19.

Krallinger M, Rabal O, Leitner F *et al.* The chemdner corpus of chemicals and drugs and its annotation principles. *J Cheminform* 2015;**7**:1–17.

Lee J, Yoon W, Kim S *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40.

Lewis P, Ott M, Du J *et al.* Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online. Association for Computational Linguistics, 2020, 146–157. https://doi.org/10.18653/v1/2020.clinicalnlp-1.17

Li J, Sun Y, Johnson RJ *et al.* Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016;**2016**:baw068.

Luo L, Lai P-T, Wei C-H *et al.* BioRED: a rich biomedical relation extraction dataset. *Brief Bioinform* 2022;**23**:bbac282.

Luoma J, Pyysalo S. Exploring cross-sentence contexts for named entity recognition with BERT. In: *Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online)*. International Committee on Computational Linguistics, 2020, 904–914. https://doi.org/10.18653/v1/2020.coling-main.78

Luoma J, Nastou K, Ohta T *et al.* S1000: a better taxonomic name corpus for biomedical information extraction. *Bioinformatics* 2023;**39**:btad369.

Mehryary F, Björne J, Pyysalo S *et al.* Deep learning with minimal training data: TurkuNLP entry in the BioNLP shared task 2016. In: *Proceedings of the 4th BioNLP Shared Task Workshop, Berlin, Germany*. Association for Computational Linguistics, 2016, 73–81. https://doi.org/10.18653/v1/W16-3009

Mehryary F, Nastou K, Ohta T *et al.* STRING-ing together protein complexes: extracting physical protein interactions from the literature. *Bioinformatics* 2024;**40**:btae552.

Meldal BHM, Perfetto L, Combe C *et al.* Complex portal 2022: new curation frontiers. *Nucleic Acids Res* 2022;**50**:D578–86.

Milošević N, Thielemann W. Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *J Web Semant* 2023;**75**:100756. https://doi.org/10.1016/j.websem.2022.100756

Miranda-Escalada A, Mehryary F, Luoma J *et al.* Overview of drugprot task at biocreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database* 2023;**2023**:baad080.

Nastou K, Mehryary F, Ohta T *et al.* RegulaTome: a corpus of typed, directed, and signed relations between biomedical entities in the scientific literature. *Database* 2024;**2024**:baae095.

Nourani E, Koutrouli M, Xie Y *et al.* Lifestyle factors in the biomedical literature: comprehensive resources for named entity recognition. *Bioinformatics* 2024, in press.

Ohta T, Pyysalo S, Miwa M *et al.* Event extraction for post-translational modifications. In: Cohen KB, Demner-Fushman D, Ananiadou S, Pestian J, Tsujii J, Webber B (eds.), *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, Uppsala, Sweden*. Association for Computational Linguistics, 19–27, 2010. https://aclanthology.org/W10-1903.

Pafilis E, Frankild SP, Fanini L *et al.* The species and organisms resources for fast and accurate identification of taxonomic names in text. *PLoS One* 2013;**8**:e65390.

Pyysalo S, Ginter F, Heimonen J *et al.* Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 2007;**8**:50–24.

Ratnaparkhi A, Marcus MP. Maximum entropy models for natural language ambiguity resolution. 1998. https://repository.upenn.edu/entities/publication/cabbef43-08b0-4c55-85c4-429dd0751228/full (22 May 2023, date last accessed).

Santos R, Ursu O, Gaulton A *et al.* A comprehensive map of molecular drug targets. *Nat Rev Drug Discov* 2017;**16**:19–34.

Smith L, Tanabe LK, Ando RJ *et al.* Overview of biocreative II gene mention recognition. *Genome Biol* 2008;**9 Suppl 2**:S2–19.

Stenetorp P, Pyysalo S, Topić G *et al.* brat: A web-based tool for NLP-assisted text annotation. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France*. Association for Computational Linguistics, 2012, 102–107. https://aclanthology.org/E12-2021.

Szklarczyk D, Kirsch R, Koutrouli M *et al.* The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2023;**51**:D638–46.

Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Red Hook, NY, USA*. Curran Associates Inc., 2017, 6000–6010. ISBN 9781510860964.

Bioinformatics Advances, 2024, 00, 1–7
https://doi.org/10.1093/bioadv/vbae116
Original Article