CRAFT Shared Tasks 2019 Overview — Integrated Structure, Semantics, and Coreference

William A Baumgartner Jr.[1], **Michael Bada**[1], **Sampo Pyysalo**[2], **Manuel R. Ciosici**[3], **Negacy Hailu**[1], **Harrison Pielke-Lombardo**[1], **Michael Regan**[4] **and Lawrence Hunter**[1]

[1]University of Colorado Anschutz Medical Campus
[2]Turku NLP Group, University of Turku, Finland
[3]UNSILO A/S
[4]University of New Mexico

Abstract

As part of the BioNLP Open Shared Tasks 2019, the CRAFT Shared Tasks 2019 provides a platform to gauge the state of the art for three fundamental language processing tasks — dependency parse construction, coreference resolution, and ontology concept identification — over full-text biomedical articles. The structural annotation task requires the automatic generation of dependency parses for each sentence of an article given only the article text. The coreference resolution task focuses on linking coreferring base noun phrase mentions into chains using the symmetrical and transitive identity relation. The ontology concept annotation task involves the identification of concept mentions within text using the classes of ten distinct ontologies in the biomedical domain, both unmodified and augmented with extension classes. This paper provides an overview of each task, including descriptions of the data provided to participants and the evaluation metrics used, and discusses participant results relative to baseline performances for each of the three tasks.

1 Introduction

With its multiple layers of annotation, the Colorado Richly Annotated Full Text (CRAFT) corpus provides a unique foundation for integrating natural language processing (NLP) tasks involving structure, semantics, and coreference. As part of the BioNLP Open Shared Tasks 2019, the CRAFT corpus was used for the evaluation of three fundamental NLP tasks: dependency parse construction, coreference resolution, and ontology concept annotation. Each of these tasks is a foundational element to many NLP systems and their performances can propagate downstream and directly affect overall system accuracy. Dependency parses have been successfully employed for information extraction, e.g. from clinical records (Gupta et al., 2018), relation extraction, e.g. identifying protein post-translational modifications (Sun et al., 2017), and used as features for machine learning tasks, e.g. gene mention detection (Smith and Wilbur, 2009), among other uses. By linking noun phrases to a referent entity, coreference systems serve as annotation multipliers, amplifying results of entity recognition systems (Cohen et al., 2017), and have been shown to improve information extraction in biomedical text (Choi et al., 2016). The concept annotation task, also known as named entity recognition (NER), is a prerequisite for many biomedical NLP applications. Its importance is buttressed by the many previous shared tasks that have included aspects of NER (Hirschman et al., 2005; Smith et al., 2008; Krallinger et al., 2013) . Measuring the state of the art of these foundational tasks will inform the BioNLP community by resetting the performance benchmarks and demonstrating optimal methodologies.

The CRAFT Shared Tasks (CRAFT-ST) 2019 mark the inaugural use and subsequent release of thirty articles annotated in CRAFT that had previously been held in reserve. All 97 articles and accompanying annotations of the CRAFT corpus are now available in the public domain. To augment the results of the CRAFT-ST 2019, and to account for the relatively low participation rate, baseline systems for each task were evaluated in the same manner as the participant systems. The CRAFT-ST 2019 made use of the CRAFT v3.1.3 release[1]. Original task descriptions are available on the CRAFT-ST website [2]. An integrated scoring platform capable of supporting the evaluation of all three sub tasks of the CRAFT-ST 2019 is

[1]https://github.com/UCDenver-ccp/CRAFT/releases/tag/v3.1.3; doi:10.5281/zenodo.3460908

[2]https://sites.google.com/view/craft-shared-task-2019

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 174–184
Hong Kong, China, November 4, 2019.

also available as a standalone system[3], and as a pre-built Docker container[4].

2 The CRAFT Structural Annotation Task

For the structural annotation task (CRAFT-SA), participants were asked to automatically parse full-length biomedical journal articles of the CRAFT Corpus into dependency structures for each sentence. The CRAFT-SA task targets dependency parses as opposed to constituency parses in order to emphasize differences that directly affect the meaning of a parsed sentence; differences in constituent parse conventions can result in parse differences that do not affect the resultant meaning of a parsed sentence (Clegg and Shepherd, 2007).

There have been previous shared tasks in the general domain NLP community to evaluate dependency parse construction using both the CoNLL-X (Buchholz and Marsi, 2006) and CoNLL-U (Zeman et al., 2018) file formats. Although the dependency parses initially distributed with the CRAFT corpus more closely resemble the older CoNLL-X format, the CRAFT dependency data was transformed into a quasi-CoNLL-U format to allow the input provided to participants to be only the text of the documents making for a more realistic scenario compared to the CoNLL-X shared tasks which required participants to match gold standard tokenization for evaluation purposes.

2.1 Data

2.1.1 Data preparation – CoNLL-X

The dependency parses distributed as part of the CRAFT corpus are automatically derived (Choi and Palmer, 2012) from the manually annotated Penn Treebank style data, which identifies the syntactic structure of each sentence. During the course of data preparation and testing, several updates were made to the Treebank data. The constituency parses for two sentences that were missing from the Treebank data were added. Also, in cases where the automatically derived dependency parse contained multiple ROOT nodes, the corresponding syntactic parse was edited, usually by dividing into multiple sentences, to ensure each dependency parse contained only a single ROOT node. Once the errors were fixed and the CoNLL-X formatted data was finalized, the data was transformed into a quasi-CoNLL-U form.

2.1.2 Data preparation – CoNLL-U

The CoNLL-U format[5] is a revised version of the CoNLL-X format that adds a number of features such as universal part-of-speech tags, language-specific part-of-speech tags, and a standardized multi-language dependency format. It includes representations of the original raw text in addition to its segmented and tokenized form. This is required for training systems that address sentence boundary detection and tokenization as part of extracting syntactic dependencies from raw text.

The CoNLL-X representation of the CRAFT dependency parses was converted into CoNLL-U format using scripts that 1) introduce document, paragraph, and sentence boundary markers and include the original untokenized text of each sentence, 2) supplement the Penn Treebank part-of-speech tags with their corresponding universal tags following the mapping proposed by the Universal Dependencies (UD) project[6], and 3) introduce morphological features based on the same part-of-speech migration guide. Spacing and paragraph information is added to the CRAFT CoNLL-U files by aligning the CoNLL-X files with the raw text for each article.

We note that while the resulting data is in the CoNLL-U format and includes UD part-of-speech tags and features, it retains the Stanford Dependency structure and labels from the CoNLL-X files and thus, does not fully conform to the UD representation in terms of its content.

2.2 Scoring

Scoring of the CRAFT-SA task made use of the scoring software provided for the CoNLL 2018 Shared Task (Zeman et al., 2018). Dependency parse performance is measured using three metrics, LAS, MLAS, and BLEX. We provide brief definitions of these metrics in the following and refer to Zeman et al. (2018) for details.

2.2.1 LAS

The Labeled Attachment Score (LAS) metric is the de facto standard metric for evaluating de-

[3] https://github.com/UCDenver-ccp/craft-shared-tasks; doi:10.5281/zenodo.3460928

[4] https://cloud.docker.com/u/ucdenverccp/repository/docker/ucdenverccp/craft-eval

[5] https://universaldependencies.org/format.html

[6] https://universaldependencies.org/tagset-conversion/en-penn-uposf.html

pendency parsing performance, and is commonly defined simply as the fraction of tokens for which the predicted head and dependency relation type (label) match the gold standard, i.e. `#correct/#tokens`. In the CoNLL 2018 setting applied in the CRAFT-SA task, this definition is generalized to account for cases where the predicted tokenization does not fully match the gold standard tokenization, and LAS is defined over aligned predicted (`pred-tokens`) and gold standard tokens (`gold-tokens`) as the harmonic mean (F1-score) over the precision `#correct/#pred-tokens` and recall `#correct/#gold-tokens`.

2.2.2 MLAS

The Morphology-aware Labeled Attachment Score (MLAS) is a modification of LAS that focuses on content words – ignoring e.g. punctuation and determiners – while also taking into account the part-of-speech, aspects of morphology, and associated function words. For a predicted token to be considered correct according to the MLAS criteria, it must match the gold standard values for the head and dependency label (as in LAS), and also the universal POS tag, selected morphological features (e.g. `Case`, `Number`, and `Tense`) and function words attached with particular dependency relations (e.g. `aux` and `case`). Similarly to LAS, MLAS is defined for system-predicted tokenization in terms of precision, recall and F1-score.

2.2.3 BLEX

Like MLAS, the Bilexical Dependency Score (BLEX) is a modification of LAS that focuses on content words, emphasizing lemmas instead of morphology. A predicted token is correct according to BLEX criteria if it matches the head, dependency relation, and lemma of the corresponding gold token. BLEX accounts for differences between the predicted and gold tokenization similarly to LAS and MLAS.

2.3 Baseline system

SyntaxNet (Andor et al., 2016), a transition-based neural network framework built using TensorFlow was used as the baseline system for the structural annotation task. The system was composed of two models of similar architecture: a part of speech (POS) tagger and a dependency parser. The Python NLTK punkt (Bird et al., 2009) sentence tokenizer was used to segment the articles into sentences which where used as input to the POS tagger model to generate POS annotations. The dependency parser model uses the POS annotations as input and generates dependency parses for each sentence. Each of the models was trained using the CRAFT training data as a gold standard.

Team	LAS	MLAS	BLEX
T013 - Run 1	65.994	0	45.618
T013 - Run 2	69.318	0	54.798
T014 - Run 1	89.695	85.549	86.631
T014 - Run 2	89.65	85.441	86.596
T014 - Run 3	89.536	85.318	86.545
Baseline	56.68	44.22	0.0

Table 1: Results showing the average score over all test documents for each metric from the structural annotation (dependency parse construction) task for all participating teams.

2.4 Results

Two teams submitted five runs in total for the CRAFT-SA task (Table 1). Team T013 used the SpaCy dependency parser with (Run1) and without (Run2) the OGER NER system to test whether adding semantic information in the form of named entities can improve resultant dependency parses. In the case of this evaluation, the incorporation of an NER system caused a drop in performance, however this decrease in performance is confounded by tokenization differences resulting from their system grouping entities as single tokens. Using a neural approach and custom biomedical word embeddings, Team T014 demonstrated state of the art performance in dependency parsing over biomedical text, achieving high marks for all submitted runs. Both submitted systems out-performed the baseline by a large margin.

3 The CRAFT Coreference Resolution Task

Coreference resolution, linking strings of text that have the same referent, is a challenging NLP task that offers potential benefit to downstream tasks if done successfully. The challenge arises in linking strings of text over long distances across a document, or possibly between documents. The benefit of doing so can be substantial as coreference resolution has the ability to amplify results of upstream

tasks such as concept recognition, thereby potentially improving the performance of downstream tasks, e.g. information extraction, that require explicitly represented entities. It has been estimated that successful coreference resolution would inherently add over 106,000 additional concept annotations to the CRAFT corpus through referent linkages (Cohen et al., 2017).

Coreference resolution is an active area in the NLP research community, and the most relevant previous shared task on coreference resolution is the CoNLL-2012 Shared Task (Pradhan et al., 2012), which evaluated identity chains curated in the OntoNotes project (Hovy et al., 2006). The OntoNotes corpus consists of text from conversational speech, broadcast conversations, broadcast news, magazine articles, newswire, and web data in three languages (English, Arabic, and Chinese), covering 1M words per language. The CRAFT corpus presents some unique challenges to the coreference resolution task. While slightly smaller than the OntoNotes corpus in regards to word count (1M), 620k works is still substantial, and scientific text is a domain not covered in OntoNotes explicitly. Further, CRAFT equals the highest median token count (24.0) per sentence (news wire) and the second highest median sentence count per document (318 vs. 565 for broadcast conversations) in the OntoNotes corpus. The combination of longer sentences and more sentences per document allows for an increase in the potential distances between coreference mentions within the sentences themselves and within each document. Adding further complexity to the task is CRAFT's use of discontinuous mentions, i.e. coreference mentions that have intervening text (see example of a discontinuous mention in Figure 1). Discontinuous mentions comprise 5.7% of all identity chain mentions in the CRAFT corpus. This is the first task on coreference resolution that allows for discontinuous mentions as far as the task organizers are aware.

3.1 Data

Annotation of the identity chains in the CRAFT corpus is described in (Cohen et al., 2017). For the purposes of the CRAFT-CR task, the strings of text (referred to as `mentions` below) that are linked to form coreference chains must exist in the same document, but can be localized any distance from one another. Some mentions may be found to be adjacent while others may exist only in the document title and conclusion, for example. Two types of coreference have been resolved for all base noun phrases in the CRAFT corpus. Identity chains link mentions of the same referent, and can span the entire document. Apposition relations link adjacent noun phrases that have the same referent and are not linked by a copula. The CRAFT-CR task focuses on reproducing the manually curated identity chains.

Statistic	Training	Test
Min IC length	2	2
Max IC length	187	157
Median IC length	3	2
Average IC length	4.77	4.70
Total IC	16,302	7,185
IC per document	243.3	239.5
Total mentions	77,755	33,749
Discont. mentions	4,485	1,845

Table 2: Descriptive statistics of the coreference resolution annotations in the CRAFT training and test sets. IC = identity chain

3.1.1 Data preparation

During the course of data preparation for the CRAFT-CR task, some errors in the coreference annotations were discovered, and subsequently fixed. The most common error involved two identity chains sharing a single base noun phrase mention. Each shared mention was manually reviewed, and the two identity chains were merged in cases where the chains were deemed to be about the same referent. In cases where the presence of a shared mention in one chain was clearly an error, it was removed and the identity chains remained distinct. The CRAFT-CR training and test data are summarized in Table 2.

3.1.2 Data format

The CRAFT-CR task makes use of the CoNLL-2011/2012 data format for representing identity chains[7], with a modification to enable representation of discontinuous mentions. Discontinuous mentions are denoted by the addition of a character or characters (non-digit) after the chain identifier (integer) as depicted in Figure 1.

[7] See the `*_conll File Format` heading: http://conll.cemantix.org/2012/data.html

```
48141 0 7  high JJ - ... - (64a)
48141 0 8  and  CC - ... - -
48141 0 9  low  JJ - ... - (65
48141 0 10 IOP  NN - ... - (64a)|65)
```

Figure 1: Sample representation of two coreference mentions, `high..IOP` and `low IOP`. Note the use of the character `a` in the chain identifier (`64a`) to indicate a discontinuous mention for the `high..IOP` mention. Empty columns 7-11 have been elided for figure layout consideration.

3.2 Scoring

There are a wide range of coreference resolution scoring metrics available. For historical purposes, the five reference metrics (MUC, B^3, CEAFE, CEAFM, BLANC) of Pradhan et al. (2014) are used to score the CRAFT-CR task. Due to their apparent unreliability and their low agreement rate, the Link-based Entity-Aware (LEA) metric proposed by Moosavi and Strube (2016) is also used to measure coreference system performance. The LEA metric was designed specifically to address the shortcomings of the previously used metrics. By taking into account all coreference links and evaluating resolved coreference relations instead of resolved mentions, the LEA metric accurately assesses recall and precision.

The coreference scoring implementations were modified in two ways for the CRAFT-CR task. First, because the CRAFT-CR data allows for mentions with discontinuous spans, the implementations were augmented to take as input the modified CoNLL-Coref 2011/2012 file format. Second, the implementations were updated to allow overlapping mentions to match instead of enforcing strict mention boundary matching. This option was added to allow for a slightly more flexible, permissive evaluation. The augmented implementations of all metrics used in the CRAFT-CR task have been made publicly available[8].

3.3 Baseline system

For comparison purposes, we evaluated the Berkeley coreference resolution system using the CRAFT-CR task test data (Durrett and Klein, 2013). The Berkeley system is an english coreference system predicated on learning using simple, but large numbers of lexicalized features. This baseline evaluation made use of the built-in preprocessing machinery for sentence splitting, tokenization, and parsing, and their pre-trained CoNLL 2012 model. Prior to evaluation, results from the Berkeley system were post-processed to adjust for some system idiosyncrasies, e.g. replacing "-LRB-" in the 'word' column with the "(" or "[" that is found in the actual text, and then the coreference information was mapped onto the gold standard tokenization provided with the test data.

3.4 Results

One team submitted three runs for evaluation in the CRAFT-CR task (Table 3). They augmented the state-of-the-art end-to-end neural coreference resolution system of Lee et al. (2017) by incorporating extra syntactic features including grammatical number agreements between mentions, as well as semantic features using MetaMap to identify entity mentions. They also investigated the use of PubMed word vectors (Chiu et al., 2016) (Run1) and SciBERT word vectors (Beltagy et al., 2019) (Run2, Run3) as inputs to their model. As implemented, the system of Team T010 performed admirably compared to the baseline. F-scores are in line with some previous coreference systems used on CRAFT (Cohen et al., 2017), thus emphasizing the challenge of coreference resolution in general, and of coreference resolution over biomedical text in particular. While the baseline system and Run1 of the participant system produced on average shorter chains than those in the evaluation set ($p<0.01$, Mann-Whitney U test), Run2 and Run3 of the participant system were both able to generate distributions of coreference chain lengths that were not significantly different from the evaluation set (Run2: p=0.94, Run3: p=0.79, Mann-Whitney U test) suggesting that inclusion of the SciBERT embeddings helps to achieve the proper chain length distribution.

4 The CRAFT concept annotation task

Concept annotation has been a mainstay in BioNLP shared tasks dating back to the very first BioCreative, which involved the detection of gene/protein mentions in abstracts and their subsequent normalization to gene identifiers from model organism databases (Hirschman et al., 2005). Detecting biomedical concepts is a foundational NLP task, and performance of this task

[8] https://github.com/bill-baumgartner/reference-coreference-scorers; doi:10.5281/zenodo.3462790

Metric	Run	P_M	R_M	F_M	P_{CR}	R_{CR}	F_{CR}
B^3	T010 - Run 3	0.731	0.578	0.646	0.517	0.384	0.440
	Baseline	0.552	0.294	0.384	0.379	0.195	0.257
$B^3{}_{APM}$	T010 - Run 3	0.779	0.615	0.687	0.538	0.406	0.462
	Baseline	0.685	0.364	0.476	0.435	0.224	0.296
BLANC	T010 - Run 3	0.731	0.578	0.646	0.506	0.473	0.489
	Baseline	0.552	0.294	0.384	0.413	0.193	0.263
$BLANC_{APM}$	T010 - Run 3	0.779	0.616	0.688	0.513	0.480	0.496
	Baseline	0.686	0.365	0.476	0.447	0.209	0.284
CEAFE	T010 - Run 3	0.731	0.578	0.646	0.454	0.354	0.398
	Baseline	0.552	0.294	0.384	0.334	0.195	0.247
$CEAFE_{APM}$	T010 - Run 3	0.779	0.615	0.688	0.484	0.377	0.424
	Baseline	0.685	0.364	0.476	0.393	0.230	0.290
CEAFM	T010 - Run 3	0.731	0.578	0.646	0.555	0.439	0.490
	Baseline	0.552	0.294	0.384	0.429	0.228	0.298
$CEAFM_{APM}$	T010 - Run 3	0.779	0.615	0.688	0.574	0.453	0.507
	Baseline	0.685	0.365	0.476	0.487	0.259	0.338
LEA	T010 - Run 3	0.731	0.578	0.646	0.475	0.345	0.400
	Baseline	0.552	0.294	0.384	0.335	0.171	0.226
LEA_{APM}	T010 - Run 3	0.779	0.615	0.687	0.491	0.360	0.415
	Baseline	0.685	0.364	0.476	0.376	0.193	0.255
MUC	T010 - Run 3	0.731	0.578	0.646	0.644	0.511	0.570
	Baseline	0.552	0.294	0.383	0.450	0.233	0.307
MUC_{APM}	T010 - Run 3	0.779	0.616	0.688	0.665	0.527	0.588
	Baseline	0.685	0.365	0.476	0.530	0.275	0.362

Table 3: Results for the coreference resolution task. Runs achieving highest coreference F-score are shown. The APM subscript indicates that partial mention matches were allowed. P_M: mention precision; R_M: mention recall; F_M: mention F-score; P_{CR}: coreference precision; R_{CR}: coreference recall; F_{CR}: coreference F-score

impacts many potential downstream applications. Mapping textual mentions of ontology concepts presents its own set of challenges. Well-known among these are conceptual synonymy, by which a given represented concept may be indicated by multiple unique textual mentions, and textual polysemy, by which a given text string may refer to multiple represented concepts. Particularly prevalent in the biomedical literature are acronyms and other abbreviations of represented concepts. Additionally, some ontologies employ standard patterns for concept labels, but some of these may result in long, complex labels that are infrequently seen in the literature (Ogren et al., 2005; Funk et al., 2014).

The CRAFT corpus is uniquely positioned to gauge the state of the art in ontological concept recognition as it comprises over 159,000 concept annotations spanning ten ontologies from the Open Biomedical Ontologies (OBO) (Smith et al., 2007) collection. Participants in the CRAFT concept annotation (CRAFT-CA) task were provided the plain-text version of each article and a file containing each ontology in the OBO format[9]. The CRAFT-CA task was further subdivided into two subtasks. The first subtask involved recognition of concepts in the original OBO files. The second subtask involved the recognition of concepts in the original OBO files augmented with *extension classes*, which are classes created by CRAFT developers but defined in terms of proper OBO classes. These extension classes were created for various reasons[10]: Some were created to capture mentions of concepts different from, but corresponding to, concepts represented in the ontologies, e.g., functionally defined entities corresponding to represented molecular functionalities. Oth-

[9] https://github.com/owlcollab/oboformat

[10] https://github.com/UCDenver-ccp/CRAFT/blob/master/concept-annotation/README.md

ers are semantically broadened forms of the represented concepts, while others were created to unify classes from different ontologies that were semantically equivalent so that there would not be multiple concept annotations for the same text spans if disparate annotation sets are aggregated.

4.1 Data

Concept annotations in the CRAFT corpus span ten Open Biomedical Ontologies (Smith et al., 2007), including the Chemical Entities of Biomedical Interest (ChEBI) ontology (Degtyarenko et al., 2007), the Cell Ontology (CL) (Bard et al., 2005), the Biological Process (GO_BP), Cellular Component (GO_CC) and Molecular Function (GO_MF) subontologies of the Gene Ontology (Ashburner et al., 2000), the Molecular Process Ontology (MOP)[11], the NCBI Taxonomy (NCBITaxon) (Federhen, 2011), the Protein Ontology (PR) (Natale et al., 2010), the Sequence Ontology (SO) (Eilbeck et al., 2005), and the Uberon cross-species anatomy ontology (UBERON) (Mungall et al., 2012). Note that concept annotations in the CRAFT corpus are permitted to have discontinuous spans with intervening text; e.g., for the phrase `somatic and germ cells`, the combination of the two substrings `somatic` and `cells` is annotated with the concept for `somatic cells` (CL:0002371) even though `somatic` and `cells` are not adjacent to one another in the text. There are over 2,300 concept annotations with discontinuous spans in the CRAFT corpus. The ontologies provided for the CRAFT-CA task were the same versions used during the annotation of CRAFT. As with the other tasks, the data is divided into a training set consisting of 67 full-text articles from the PMC Open Access subset, and a test set of 30 full-text articles chosen using identical selection criteria. Concept annotation of the CRAFT articles is described in detail in Bada et al. (2012) and Bada et al. (2017). Summary statistics showing total annotation counts for the ten ontologies used in the CRAFT corpus are shown in Table 4.

4.1.1 Data preparation

Some minor concept annotation errors were discovered and addressed during preparation for the CRAFT-CA task. These errors included an NCBITaxon concept that was found to not exist in the version of the NCBI Taxonomy used to annotate CRAFT, as well as some erroneous extension class prefixes used in the GO_MF extended ontology file. Errors were addressed prior to the commencement of the shared tasks.

4.1.2 Data format

The CRAFT corpus is distributed with a script that can convert its native annotation format to a variant of the BioNLP format[12] which is used for both input and output for the CRAFT-CA task. This format captures span information, the concept identifier, and the covered text for each annotation (See Figure 2).

4.2 Scoring

The method of Bossy et al. (2013) was used to measure performance of the concept annotation systems with respect to the CRAFT corpus. This method employs a hybrid measure taking into account both the degree to which the predicted annotation boundaries match the reference, as well as a similarity metric for scoring the concept match. The boundary match uses the modified Jaccard index scheme described in Bossy et al. (2012), which allows for flexible matching but prefers exact matches. The concept similarity metric of Wang et al. (2007) is used to score the predicted concepts. As suggested by Bossy et al. (2013) , the weight factor, w, was set to 0.65, which ensures that ancestor/descendant predictions always have a greater value than sibling predictions, while root predictions never yield a similarity greater than 0.5. An implementation of the scoring algorithm has been made publicly available[13].

4.3 Baseline system

We evaluated a baseline system on the CRAFT-CA data to use as a comparison for the participant-submitted runs. The baseline system is a two-stage machine learning system proposed in Hailu (2019) and trained only on the CRAFT corpus. The first stage makes use of NERSuite (Cho et al., 2010) to detect concept mention spans using a conditional random field (CRF) model. The CRF model was trained as described in Okazaki (2007), and uses as features words, parts of speech, and constituency parse information within a window of three tokens

[11] http://obofoundry.org/ontology/mop.html

[12] http://2013.bionlp-st.org/file-formats

[13] https://github.com/UCDenver-ccp/craft-shared-tasks; doi:10.5281/zenodo.3460928

Ontology	Training	Test	Ontology	Training	Test
CHEBI	4,548 (18)	2,200 (14)	CHEBI_EXT	11,915 (38)	5,248 (19)
CL	4,043 (244)	1,749 (175)	CL_EXT	6,275 (249)	2,872 (175)
GO_BP	9,280 (493)	3,681 (272)	GO_BP_EXT	13,954 (526)	5,847 (287)
GO_CC	4,075 (80)	1,184 (14)	GO_CC_EXT	8,495 (150)	3,217 (30)
GO_MF	375 (0)	94 (0)	GO_MF_EXT	4,070 (28)	1,822 (20)
MOP	240 (0)	101 (0)	MOP_EXT	386 (0)	111 (0)
NCBITaxon	7,362 (2)	3,101(0)	NCBITaxon_EXT	7,592 (2)	3,219 (0)
PR	17,038 (84)	6,409 (44)	PR_EXT	19,862 (110)	7,932 (44)
SO	8,797 (108)	3,446 (45)	SO_EXT	24,955 (182)	9,136 (72)
UBERON	12,269 (235)	6,551 (118)	UBERON_EXT	14,910 (255)	7,416 (133)

Table 4: Total and discontinuous (in parentheses) concept annotation counts by ontology for both the 67 article training and 30 article test sets.

```
T1   CL:0000540 83 89   neuron
T2   CL:0002613 239 247;259 265   striatal ... neuron
T3   CL:0002613 434 442;451 457   striatal ... neuron
T4   CL:0000540 703 709   Neuron
```

Figure 2: Sample annotations demonstrating the BioNLP format used as input and output for the CRAFT-CA task. Note the presence of two annotations with discontinuous spans. The document identifier is indicated in the filename for each annotation file.

		Proper OBO				OBO + extension			
Ontology	**Submission**	**SER**	**P**	**R**	**F1**	**SER**	**P**	**R**	**F1**
CHEBI	T013 - Run 3/1	0.34	0.79	0.75	0.77	0.27	0.84	0.79	0.81
	Baseline	0.44	0.91	0.59	0.72	0.29	0.89	0.73	0.80
CL	T013 - Run 3/2a	0.56	0.68	0.62	0.65	0.35	0.77	0.67	0.72
	Baseline	0.53	0.83	0.48	0.61	0.33	0.79	0.67	0.73
GO_BP	T013 - Run 3/1	0.30	0.83	0.78	0.80	0.29	0.81	0.81	0.81
	Baseline	0.39	0.83	0.64	0.72	0.29	0.84	0.74	0.79
GO_CC	T013 - Run 1/2a	0.39	0.77	0.75	0.76	0.20	0.92	0.83	0.87
	Baseline	0.44	0.88	0.60	0.71	0.20	0.93	0.83	0.88
GO_MF	T013 - Run 2/2a	0.04	0.99	0.96	0.98	0.39	0.82	0.68	0.74
	Baseline	0.07	0.99	0.92	0.95	0.45	0.82	0.56	0.66
MOP	T013 - Run 3/2a	0.27	0.81	0.94	0.87	0.34	0.89	0.73	0.79
	Baseline	0.43	0.87	0.65	0.75	0.36	0.88	0.72	0.79
NCBITaxon	T013 - Run 3/2a	0.05	0.97	0.97	0.97	0.077	0.98	0.93	0.96
	Baseline	0.07	0.99	0.93	0.96	0.07	0.99	0.94	0.96
PR	T013 - Run 3/1	0.68	0.50	0.59	0.54	0.73	0.49	0.46	0.47
	Baseline	0.69	0.60	0.40	0.48	0.62	0.61	0.45	0.52
SO	T013 - Run 3/2a	0.16	0.90	0.88	0.89	0.13	0.92	0.91	0.92
	Baseline	0.21	0.91	0.82	0.86	0.18	0.92	0.85	0.89
UBERON	T013 - Run 1/2a	0.37	0.77	0.71	0.74	0.39	0.77	0.69	0.73
	Baseline	0.41	0.84	0.61	0.70	0.36	0.86	0.66	0.75

Table 5: Aggregate concept annotation results evaluated per ontology against the 30 CRAFT test documents. For Team T013, their highest scoring run is displayed based on SER. Run identifiers indicate (proper OBO/OBO_EXT). Note that Run 2a is an unofficial run as it was submitted after the deadline, however since there were no other teams participating, Run 2a is included in the official results. SER = Slot Error Rate; P = Precision; R = Recall; F1 = F1-score.

upstream and downstream of each concept mention. The second stage links each textual mention identified by the CRF to an ontology identifier using a stacked Bi-LSTM approach implemented by the OpenNMT system (Klein et al., 2018). By modeling concept normalization as sequence-to-sequence translation at the character level, the baseline system maps characters in the text spans identified in the first stage to characters in ontology identifiers to normalize concepts.

4.4 Results

One team submitted three runs to the CRAFT-CA task (Table 5). They used variants of two systems, one a modified ontology-specific BioBert[14] model with (Run3) and without (Run1) input from the OGER NER system (Furrer et al., 2019) and with weights pretrained on PubMed using identifiers from the ontologies as the tag set, and the other a BiLSTM with ontology pretraining (Run2). With regard to overall system performances, marked improvement in recognition of concepts from CHEBI, GO_BP, GO_MF, and SO was observed compared to past evaluations using the CRAFT public dataset (Funk et al., 2014). However, it is important to note that past evaluations were performed on CRAFT v1/2 concept annotations, whereas the testing of this shared task was performed on v3 concept annotations, which constitute a major update of the concept annotations relative to those of v1/2 (including first usage of extension classes), so we do not believe it is safe to directly compare evaluations performed on these substantially different versions of the concept annotations. The BioBert approach augmented with the OGER NER system (Run3) generally outperforms the other approaches when normalizing to proper OBO concepts, whereas the BiLSTM approach is generally better when the extension classes are used.

Neither the baseline system, nor any of the submitted runs identified annotations with discontinuous spans. Though annotations with discontinuous spans make up only a small percentage (1.46%) of the overall annotations, their exclusion from system output could represent potential low hanging fruit for improving overall system performance. Protein Ontology concept recognition remains a target for future work as system performances did not surpass an F-score of 0.55. Inclusion of the extension classes generally resulted in improvement of performance when compared to runs using only the proper ontology concepts, possibly attributable to the labels and synonyms that were provided for the extension classes. One exception is for GO_MF_EXT where performance is expected to suffer with inclusion of the extension class annotations as the proper ontology class count was limited to a very small subset of the original ontology. Overall, however, performance on the CRAFT-CA task demonstrated state-of-the-art performance for ontological concept recognition in biomedical text.

[14] https://github.com/dmis-lab/biobert

5 Conclusion

The CRAFT-ST 2019 provides a platform to gauge performance on three fundamental NLP tasks, automated dependency parse construction, coreference resolution, and ontology concept annotation against a high quality, manually annotated corpus of full-text biomedical articles. Submitted runs from participating systems demonstrate promising results, particularly with respect to automated dependency parse construction and some aspects of ontological concept annotation. Clear needs for improved extraction of protein ontology concepts remain, while the neural approaches used have addressed long standing deficiencies in the recognition of biological process concepts in text. Coreference resolution system performances highlight the existing challenges of coreference resolution in general, and of coreference resolution over biomedical text in particular.

The approaches taken by participants in the CRAFT-ST 2019 mirror the current themes in AI and NLP today. Neural approaches are unsurprisingly the preferred methodology for addressing these NLP tasks. The CRAFT ST 2019 have provided new benchmarks for these fundamental NLP tasks, setting the stage for the next evolution of system development.

Acknowledgments

The authors would like to thank Kevin Cohen and Karin Verspoor for their input during the early planning stages for the CRAFT Shared Task 2019, and Tiffany Callahan for help with the coreference chain length statistics. The authors gratefully acknowledge their support from NIH grants R01LM009254, R01LM008111 and T15LM009451.

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany. Association for Computational Linguistics.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the craft corpus. *BMC bioinformatics*, 13(1):161.

Michael Bada, Nicole Vasilevsky, William A Baumgartner, Melissa Haendel, and Lawrence E Hunter. 2017. Gold-standard ontology-based anatomical annotation in the craft corpus. *Database*, 2017.

Jonathan Bard, Seung Y Rhee, and Michael Ashburner. 2005. An ontology for cell types. *Genome biology*, 6(2):R21.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec. 2013. Bionlp shared task 2013–an overview of the bacteria biotope task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 161–169.

Robert Bossy, Julien Jourde, Alain-Pierre Manine, Philippe Veber, Erick Alphonse, Maarten Van De Guchte, Philippe Bessières, and Claire Nédellec. 2012. Bionlp shared task-the bacteria track. In *BMC bioinformatics*, volume 13, page S3. BioMed Central.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning*, pages 149–164. Association for Computational Linguistics.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174.

Han-Cheol Cho, Naoaki Okazaki, Makoto Miwa, and Jun?ichi Tsujii. 2010. Nersuite: a named entity recognition toolkit. *Tsujii Laboratory, Department of Information Science, University of Tokyo, Tokyo, Japan*.

Jinho D Choi and Martha Palmer. 2012. Guidelines for the clear style constituent to dependency conversion. *Technical Report 01–12*.

Miji Choi, Haibin Liu, William Baumgartner, Justin Zobel, and Karin Verspoor. 2016. Coreference resolution improves extraction of biological expression language statements from texts. *Database*, 2016.

Andrew B Clegg and Adrian J Shepherd. 2007. Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC bioinformatics*, 8(1):24.

K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. 2017. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC bioinformatics*, 18(1):372.

Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982.

Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. 2005. The sequence ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):R44.

Scott Federhen. 2011. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143.

Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K Bretonnel Cohen, Lawrence E Hunter, and Karin Verspoor. 2014. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC bioinformatics*, 15(1):59.

Lenz Furrer, Anna Jancso, Nicola Colic, and Fabio Rinaldi. 2019. Oger++: hybrid multi-type entity recognition. *Journal of cheminformatics*, 11(1):7.

Anupama Gupta, Imon Banerjee, and Daniel L Rubin. 2018. Automatic information extraction from unstructured mammography reports using distributed semantics. *Journal of biomedical informatics*, 78:78–86.

Negacy Degefa Hailu. 2019. *Investigation of traditional and deep neural sequence models for biomedical concept recognition*. Ph.D. thesis, University of Colorado.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of biocreative: critical assessment of information extraction for biology.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M Rush. 2018. Opennmt: Neural machine translation toolkit. *arXiv preprint arXiv:1805.11462*.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2013. Overview of the chemical compound and drug name recognition (chemdner) task. In *BioCreative challenge evaluation workshop*, volume 2, page 2. Citeseer.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642.

Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13(1):R5.

Darren A Natale, Cecilia N Arighi, Winona C Barker, Judith A Blake, Carol J Bult, Michael Caudy, Harold J Drabkin, Peter D?Eustachio, Alexei V Evsikov, Hongzhan Huang, et al. 2010. The protein ontology: a structured representation of protein forms and complexes. *Nucleic acids research*, 39(suppl_1):D539–D545.

Philip V Ogren, K Bretonnel Cohen, and Lawrence Hunter. 2005. Implications of compositionality in the gene ontology for its curation and usage. In *Biocomputing 2005*, pages 174–185. World Scientific.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2014, page 30. NIH Public Access.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. 2007. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):S2.

Larry H Smith and W John Wilbur. 2009. The value of parsing as feature generation for gene mention recognition. *Journal of biomedical informatics*, 42(5):895–904.

Dongdong Sun, Minghui Wang, and Ao Li. 2017. Mptm: A tool for mining protein post-translational modifications from literature. *Journal of bioinformatics and computational biology*, 15(05):1740005.

James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. 2007. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.