Overview of TAC-KBP2019 Fine-grained Entity Extraction

Heng Ji[1], **Avirup Sil**[2], **Hoa Trang Dang**[3], **Ian Soboroff**[3], **Joel Nothman**[4]
[1] Computer Science Department, University of Illinois at Urbana-Champaign
hengji@illinois.edu
[2] IBM Research AI
avi@us.ibm.com
[3] National Institute of Standards and Technology
{hoa.dang,ian.soboroff}@nist.gov
[4] Sydney Informatics Hub, University of Sydney
joel.nothman@gmail.com

Abstract

In the past several years TAC KBP Entity Discovery and Linking (EDL) track has only focused on five major coarse-grained entity types: person (PER), geo-political entity (GPE), location (LOC), organization (ORG) and facility (FAC). However, many real-world applications in scenarios such as disaster relief and technical support require us to significantly extend EDL capabilities to a wider variety of fine-grained entity types (e.g., technical terms, lawsuits, disease, crisis, vehicles, food, biomedical entities). In this overview paper we give an outline of the Ultra-Fine-Grained Name Tagging for Entity Types task (along with system participation) at the Knowledge Base Population (KBP) track at TAC 2019. We will also sketch out remaining challenges and future research directions.

1 Introduction

The Entity Discovery and Linking (EDL) track at TAC-KBP has experienced eleven years of joy and prosperity, thanks to the successful community efforts and DARPA and NIST's support at creating valuable resources and shared tasks. In addition to improved quality at each subtask (mention extraction, linking and NIL clustering), the major recent accomplishment lies in the dramatically enhanced portability. State-of-the-art EDL techniques today can take an arbitrary large-size corpus as input, extract entities from hundreds of languages (Pan et al., 2017), and link them to English knowledge bases with either rich properties (e.g., DBPedia) or scarce properties (e.g., World Fact Book or simply a product name list).

The goal of TAC-KBP EDL is to extract mentions of pre-defined entity types from unstructured text in any language, and link them to the entities in an English knowledge base (KB). In the past several years we have only focused on five major coarse-grained entity types: person (PER), geo-political entity (GPE), location (LOC), organization (ORG) and facility (FAC). However, many real-world applications in scenarios such as disaster relief and technical support require us to significantly extend EDL capabilities to a wider variety of fine-grained entity types (e.g., technical terms, lawsuits, disease, crisis, vehicles, food, biomedical entities). For example, the most frequent questions people ask Amazon's Alexa often involve new products, movies and actors.

EDL systems must be capable of assigning fine-grained types from any ontology, with a limited amount of training data for each entity type, because they must ultimately operate in specific domains for applications that require different ontologies. For example, in the following sentence, "*If **Rogers** is in the game, the Huskies will be much better equipped to match the Cougars in that aspect*", the mention "Rogers" should be labeled as `athlete` in addition to `person` according to the context (e.g., game, Huskies). These fine-grained entity types are proven to be effective in supporting a wide range of downstream applications such as relation extraction (Yao et al., 2010), question answering (Lin et al., 2012), and coreference resolution (Recasens et al., 2013).

The NLP community has done a lot of work on fine-grained entity typing (Ling and Weld, 2012; Choi et al., 2018; Xin et al., 2018). In TAC-KBP2019 we propose a new fine-grained entity extraction task that incorporates the following new and exciting changes:

- **Low-resource:** We will extend the number of types from 5 types to over 187 fine-grained types and evaluate the extraction capabilities for ultra-fine-grained entity typing with very little training data for each type. These fine-grained entity types have been defined for the TAC 2019 SM-KBP track, which evaluates tasks for the DARPA AIDA program. In an AIDA scenario, there is much informational conflict about the entities that participate in crucial newsworthy events and relations, and both fine-grained entity types and mention boundaries are needed to represent different hypotheses about the scenario (e.g., Is the shooter in an attack event a Military Personnel, or a Demonstrator?).

- **Human-in-the-loop:** We will provide limited “feedback” on system output, which should be used to improve each system. The feedback will be provided based on a user model of how analysts might interact with the system while giving feedback on system output.

- **Gold-standard annotation and resources:** Previous efforts have been compared on many different silver-standard annotations derived from Encyclopedia resources. In this task we have developed various resources including LDC human annotated 500 documents, UIUC human annotated 144 documents, and many silver-standard annotations derived from Wikipedia and YAGO.

2 Task Definition and Evaluation Metrics

This section summarizes the fine-grained Entity Extraction task conducted at TAC-KBP 2019. More details regarding data format and scoring software can be found in the task website[1].

2.1 Fine-grained Entity Extraction Task

Given a document collection, an entity extraction system is required to automatically identify and classify entity name mentions into one of the types defined in the schema (section 2.3). This year we only focus on mention extraction+classification and only on English source documents. In future years we expect to add entity linking and source documents in foreign languages. The output contains one line for each mention, where each line has the following tab-delimited fields:

- Field 1: system run ID

- Field 2: mention (query) ID: unique for each entity name mention.

- Field 3: mention head string: the full head string of the entity name mention.

- Field 4: document ID: mention head start offset mention head end offset: an ID for a document in the source corpus from which the mention head was extracted, the starting UTF-8 character offset of the mention head, and the ending UTF-8 character offset of the mention head.

- Field 5: reference KB link entity ID (or NIL link): entity linking is not required in 2019, so please put “NIL” label for this column.

- Field 6: entity type: a type indicator for the entity

- Field 7: mention type: only name mentions are required in 2019, so please put “NAM” for this column.

- Field 8: a confidence value. Each confidence value must be a positive real number between 0.0 (exclusive, representing the lowest confidence) and 1.0 (inclusive, representing the highest confidence), and must include a decimal point (no commas, please).

2.2 Scoring Metrics

We use the standard precision, recall and F-score to evaluate the performance of fine-grained entity etraction. We design the following weighted metric for entity mention typing (classification):

- 1 if the system’s type (sys) is identical to the gold type (gold)

[1] http://nlp.cs.rpi.edu/kbp/2019/

- If the system's type is a supertype of the gold type, then the system receives partial credit based on: $0.5^{(\text{depth(sys)}-\text{depth(gold)})}$, e.g.,
  - 0.5 if sys is parent of key
  - 0.25 if sys is grandparent of key
  - 0.125 if sys is great grandparent of key
- 0 otherwise

This scoring metric is implemented in: https://github.com/wikilinks/neleval with its `--type-weights` configuration.

2.3 Entity Ontology / Schema

LDC has defined a set of 187 entity types that are salient to the SM-KBP 2019 scenario (2014-2015 Russia-Ukraine conflict), for the purposes of representing informational conflict and different hypotheses about the entities, events, and relations in the scenario. For EDL 2019, we choose these 187 entity types as our target ontology.

However, we also set up an optional task to extract 7000+ entity types defined in the YAGO ontology to highly encourage teams to develop techniques for wider domains. To assist in system development, we selected 7,309 entity types from YAGO/WordNet. Each type has at least 10 entity entries in DBPedia. Full information about the types and schema are provided in the footnote urls below [2]

For EDL 2019, systems must be able to identify the 187 AIDA entity types that are defined for the SMKBP 2019 scenario: The reason for restricting the EDL evaluation to SM-KBP entity types is two-fold: First, to reduce the number of different types that a human annotator must keep track of and still be able to provide reliable annotations. Second, there is much overlap between the fine-grained types in YAGO (so a single mention in context can often have multiple types); the SMKBP entity types have been curated to avoid most of the type overlap in YAGO, focusing only on the type distinctions that are needed for the AIDA application and scenario.

The entity types are defined in LDC2019E07, as well as in the appendix of the annotation guidelines. The 187 entity types comprise a hierarchical ontology that has three levels (type, subtype, subsubtype):

- Type level: The most coarse-grained level. Types at this level have the form type (e.g., “per”)
- Subtype level: Types at this level have the form type.subtype (e.g., “per.politician” , having parent “per”)
- Subsubtype level: The most fine-grained level. Types at this level have the form type.subtype.subsubtype (e.g., “per.politician.governor” , having parent “per.politician”)

Annotators and participants should select a type at the finest-grained level they can confidently label, backing off to a higher level if necessary. A partial mapping from AIDA entity types to YAGO entity types is provided [3].

2.4 Human Feedback

Human feedback was provided to systems based on a user model of how analysts might interact with the system. In the model, the user is reviewing a document and working through it from the beginning to the end. During this sequential process, if at time t the user encounters an incorrect system-generated annotation (which could be a miss, an error in mention extent, or a type error), the user tags the error and provides the correction. The user has some tolerance value k for the amount of error they are willing to see before they lose confidence in the system ’ s annotations and stop reviewing the document. Let the amount of error encountered so far in document d at time t be a function Err(d,t). The user will stop reviewing the document as soon as Err(d,t) was greater than k. Err(d,t) could be a function of the number of errors seen up to time t, how egregious the errors are, and possibly how the errors are distributed across the document. For 2019, we used a simple error function: Err(d,t) = total number of errors seen up to time t. Feedback was given until k=10 errors were detected in the document.

All participants in EDL Evaluation Window 1 were given feedback on the same set of 10 documents. The 10 documents for feedback were selected to be documents where many of the runs submitted to Evaluation Window 1 had low confidence: For each run from

[2] https://nlp.cs.rpi.edu/kbp/2019/yago_types_of_at_least_10.json https://nlp.cs.rpi.edu/kbp/2019/yago_wordnet_definition.json https://blender04.cs.rpi.edu/ panx2/tmp/typing/

[3] http://nlp.cs.rpi.edu/kbp/2019/YAGO_AIDA_mapping.xlsx

Evaluation Window 1, NIST identified 10 “low confidence” documents, defined to be documents that had lowest mean confidence in the first 20 mentions returned for the document; the final set of feedback documents was a set of 10 documents that had a maximal number of runs that identified the document as being “low confidence”.

3 Data Annotation and Resources

The Source collection for the evaluation includes 300K documents. Participants must output entity types for all 300K documents, though only a core subset of 300 documents is evaluated. The details of the LDC data annotation for these 300 evaluation documents and AIDA related annotations are presented in a separate paper by the Linguistic Data Consortium (Getman et al., 2019). UIUC has annotated 144 additional documents following the AIDA ontology and shared them with the community.

Because ultra-fine-grained entity types bring significant challenges to both of the annotation interface and human annotation, UIUC team has created silver-standard annotation [4] derived from Wikipedia markups (Pan et al., 2017) for 16K+ YAGO entity types (571GB). One strategy for producing training data is to manually annotate data using the specific ontology for that domain; alternatively, the domain-specific ontology can be mapped to a more general, broad-coverage ontology with fine-grained types such as YAGO (Suchanek et al., 2007), and existing silver-standard annotations and other resources for YAGO can be used to develop the EDL system for the new domain. Table 1 presents some examples of the 16K+ entity types defined in YAGO.

4 Participants Overview

Table 1 summarizes the participants for KBP2019 EDL tasks. In total 8 teams submitted for the first EDL evaluation window without human feedback, 16 teams submitted runs for the second evaluation window with human feedback.

5 Evaluation Results

Table 2 summarizes the results. For public release we have anonymized the team names: each team is numbered with the rank of its best submission. Overall the human feedback has provided up to 19.1% absolute gain in F-score.

6 What’s New and What Works

6.1 Incorporating Contextual Embedding Representations

Most teams have incorporated deep contextual embedding representations in deep neural networks models. UIUC team (Lin and Ji, 2019; Lin et al., 2019) has used ELMO embeddings. IBM Team performs a two pass approach: first detect nine basic NER types (PER, ORG, LOC, FAC, GPE, LAW, CRM, WEA, VEH), followed by a second pass of refining these types into the destination fine-grained types. The first pass is a regular BERT model based name tagger, while the second pass model is a mention-centric classifier using RoBERTa and has features that comprise each mention and its context separately to specialize the mention into a fine-grained type.

6.2 Joint Fine-grained Entity Identification, Classification and Linking

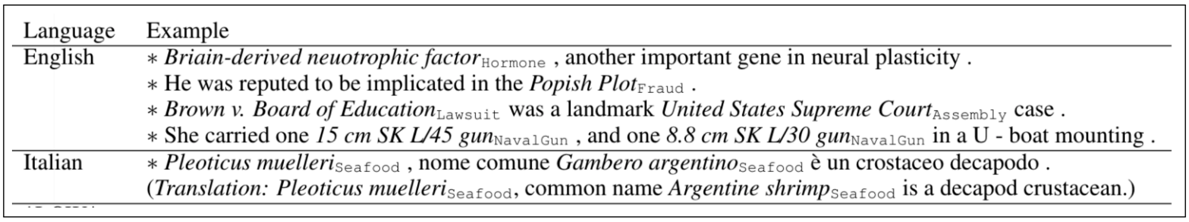
Ousia Team performs joint entity extraction along with type classification. Their methods include a pre-training and a fine-tuning phase. In the pre-train phase, they perform encode the given context with BERT along with BiLSTM. Then they compute two layers: one performing a CRF to get standard BIO tags and the other computes YAGO tags. In the fine-tune phase, these two are combined to learn a mapping to output AIDA tags.

UIUC team and Diffbot team use Entity Linking (EL) to link mentions to the KB and then infer the types from the KB. The types from the KB are restricted to the AIDA types: in particular they use the parent types of the linked entity in the type ontology. Diffbot system further uses 10 representative entities for each type using a popularity metric.

6.3 Capturing Type Inter-dependency

Fine-grained entity typing is usually formulated as a multi-label classification problem. Previous approaches (Ling and Weld, 2012; Choi et al., 2018; Xin et al., 2018) typically address it with binary relevance that decomposes the problem into isolated binary classification subproblems and independently predicts each type. However, this method is commonly criticized for its label independence assumption, which is not valid for

[4] https://blender04.cs.rpi.edu/ panx2/tmp/fine-grained/



Language	Example
English	∗ *Briain-derived neuotrophic factor*$_{\texttt{Hormone}}$, another important gene in neural plasticity . ∗ He was reputed to be implicated in the *Popish Plot*$_{\texttt{Fraud}}$. ∗ *Brown v. Board of Education*$_{\texttt{Lawsuit}}$ was a landmark *United States Supreme Court*$_{\texttt{Assembly}}$ case . ∗ She carried one *15 cm SK L/45 gun*$_{\texttt{NavalGun}}$, and one *8.8 cm SK L/30 gun*$_{\texttt{NavalGun}}$ in a U - boat mounting .
Italian	∗ *Pleoticus muelleri*$_{\texttt{Seafood}}$, nome comune *Gambero argentino*$_{\texttt{Seafood}}$ è un crostaceo decapodo . (*Translation: Pleoticus muelleri*$_{\texttt{Seafood}}$, common name *Argentine shrimp*$_{\texttt{Seafood}}$ is a decapod crustacean.)

Figure 1: Some sentence examples annotated with YAGO entity types.

Team	Affiliation	Without Feedback	With Feedback
Diffbot	Diffbot	✓	✓
Dsln_nlptt	NEC Corporation & Tokyo Institute of Technology		✓
HITS_UKP	Heidelberg Institute for Theoretical Studies	✓	✓
IBM_MNLP_IE	IBM Research	✓	✓
Ousia	Studio Ousia Inc.	✓	✓
RWTH	RWTH Aachen University	✓	✓
TAI	Tencent AI Platform	✓	✓
UIUC_BLENDER	University of Illinois at Urbana-Champaign	✓	✓
ZJU_EDL	Zhejiang University	✓	✓

Table 1: Runs Submitted by KBP2019 Fine-grained Entity Extraction Participants (in alphabetical order)

Team	without feedback			with feedback		
	P	R	F_1	P	R	F_1
1	**54.0**	**51.7**	**52.8**	**60.2**	**62.7**	**61.4**
2	50.4	46.8	48.5	50.6	49.3	49.9
3	43.1	42.7	42.9	56.0	55.5	55.7
4	43.1	41.3	42.2	43.3	41.4	42.3
5	46.5	36.4	40.8	35.0	45.1	39.4
6	37.5	41.8	39.5	42.3	47.2	44.6
7	22.5	20.7	21.6	22.1	20.3	21.2
8	27.0	15.5	19.7	42.2	35.8	38.8
9				3.3	2.2	2.6

Table 2: Overall Fine-grained Entity Extraction Performance (%)

fine-grained entity typing. For example, if the model is confident at predicting the type `artist`, it should promote its parent type `person` but discourage `organization` and its descendant types. In order to capture inter-dependencies between types, UIUC team (Lin and Ji, 2019; Lin et al., 2019) has developed a hybrid model that incorporates latent type representation in addition to binary relevance. Specifically, the model learns to predict a low-dimensional vector that encodes latent type features obtained through Principle Label Space Transformation (Tai and Lin, 2012) and reconstruct the sparse and high-dimensional type vector from this latent representation.

6.4 Differentiating Similar Types

Another major challenge in fine-grained entity typing is to differentiate similar types, such as `director` and `actor`, which requires the model to capture slightly different nuances in texts. Previous neural models (Shimaoka et al., 2016; Xin et al., 2018; Choi et al., 2018; Xu and Barbosa, 2018) generally extract features from pre-trained word embeddings. Instead, UIUC team (Lin and Ji, 2019; Lin et al., 2019) adopts contextualized word representations (Peters et al., 2018), which can capture context-aware word semantics and better represent out-of-vocabulary words. They further propose a two-step attention mechanism to actively extract the most relevant information from the sentence. Particularly, they calculate the attention for context words in a mention-aware manner, allowing the model to focus on different parts of the sentence for different mentions. For example, in the following sentence, "*In 2005 two federal agencies, the **US Geological Survey** and the **Fish and Wildlife Service**, began to identify fish in the **Potomac** and tributaries ...*", the model should use "federal agencies" to help classify "US Geological Survey" and "Fish and Wildlife Service" as `government_agency`, but focus on "fish" and "tributaries" to determine that "Potomac" should be a `body_of_water` (river) instead of a `city`.

7 Remaining Challenges

Coreference resolution remains a major bottleneck as always. State-of-the-art coreference resolver based on contextualized embeddings often mistakenly link pronouns. For example, in the following sentence, "*When Melnychuk's body* *was found on 22 March , police initially told local journalists he had committed suicide.*", the coreference resolver mistakenly links "*he*" to "*journalists*" because they are close, and also links "*he*" to the only single person entity "*Melnychuk*" in the sentence, and thus the UIUC system mistakenly assign a type *PER.ProfessionalPosition.Journalist* to "*Melnychuk*".

Most of the remaining challenges involve background knowledge acquisition and reasoning. For example, lexical embedding features may tag either house or garden type to "*White House Rose Garden*". We may need to acquire the types of events held in this place or visual features to infer its type not as *FAC.Building.House*.

Current language modeling based embedding features often fail to select the most indicative features and filter noise in contexts. For example, the UIUC system mistakenly assigns "*CRM.ViolentCrime.Terrorism*" to the entity mention "*terror*" without knowing it's a merely a term referring to a threat in a commentary sentence: "*According to Greg Fealy, an associate professor at the Australian National University who studies terrorism in Indonesia, the IS terror threat in Indonesia has been rising since mid-2014.*".

8 Planning or KBP2020 EDL

In KBP2020, we would like to explore the following new extensions of fine-grained EDL.

- Perform real ultra-fine-grained entity extraction for 16,000 entity types by addressing annotation challenges.
- Add entity linking and foreign languages.
- Extend to cross-media fine-grained entity extraction and grounding.
- Integrate it into new applications such as Amazon Alexa which requires new entity clustering to initiate proactive conversations.

Acknowledgments

Heng Ji's work has been supported by the U.S. DARPA KAIROS Program No. FA8750-19-2-1004, U.S. DARPA AIDA Program No. FA8750-18-2-0014, Air Force No. FA8650-17-C-7715, and U.S. ARL NS-CTA No. W911NF-09-2-0053. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*.

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie Strassel. 2019. Overview of linguistic resources for the tac kbp 2019 evaluations: Methodologies and results. In *Proc. Text Analysis Conference (TAC2019)*.

Ying Lin and Heng Ji. 2019. An attentive fine-grained entity typing model with latent type representation. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP2019)*.

Thomas Lin, Oren Etzioni, et al. 2012. No noun phrase left behind: detecting and typing unlinkable entities. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2012)*.

Ying Lin, Xiaoman Pan, Manling Li, and Heng Ji. 2019. A baseline fine-grained entity extraction system for tac-kbp2019. In *Proceedings of Text Analysis Conference (TAC2019)*.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012)*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017)*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*.

Sonse Shimaoka, Pontus Stenetorp, Kentaro Inui, and Sebastian Riedel. 2016. An attentive neural architecture for fine-grained entity type classification. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*.

Fabian Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *WWW*, pages 697–706. ACM.

Farbound Tai and Hsuan-Tien Lin. 2012. Multilabel classification with principal label space transformation. *Neural Computation*, 24(9):2508–2542.

Ji Xin, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018. Improving neural fine-grained entity typing with knowledge attention. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*.

Peng Xu and Denilson Barbosa. 2018. Neural fine-grained entity type classification with hierarchy-aware loss. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*.