"What Are You Trying to Do?" Semantic Typing of Event Processes

Muhao Chen[1,2], **Hongming Zhang**[3*], **Haoyu Wang**[1], **Dan Roth**[1]
[1]Department of Computer and Information Science, UPenn
[2]Information Sciences Institute, USC
[3]Department of Computer Science and Engineering, HKUST
muhaoche@usc.edu; hzhangal@cse.ust.hk;
{why16gzi, danroth}@seas.upenn.edu

Abstract

This paper studies a new cognitively motivated semantic typing task, *multi-axis event process typing*, that, given an event process, attempts to infer free-form type labels describing (i) the type of action made by the process and (ii) the type of object the process seeks to affect. This task is inspired by computational and cognitive studies of event understanding, which suggest that understanding processes of events is often directed by recognizing the goals, plans or intentions of the protagonist(s). We develop a large dataset containing over 60k event processes, featuring ultra fine-grained typing on both the action and object type axes with very large ($10^3 \sim 10^4$) label vocabularies. We then propose a hybrid learning framework, P2GT, which addresses the challenging typing problem with indirect supervision from glosses[1] and a joint learning-to-rank framework. As our experiments indicate, P2GT supports identifying the intent of processes, as well as the fine semantic type of the affected object. It also demonstrates the capability of handling few-shot cases, and strong generalizability on out-of-domain processes.[2]

1 Introduction

Events are the fundamental building blocks of natural languages. To help machines understand events, extensive research effort has been devoted to inducing how events described in text are procedurally connected (Ning et al., 2017; Radinsky et al., 2012), and how they form *event processes*[3] (Pichotta and Mooney, 2014; Berant et al., 2014; Jindal and Roth, 2013). Consequently, such prototypical schematic sequences of events have found important use cases including storyline construction (Do et al., 2012; Radinsky and Horvitz, 2013), narrative cloze (Chaturvedi et al., 2017; Lee and Goldwasser, 2019), biological process comprehension (Berant et al., 2014) and diagnostic prediction (Zhang et al., 2020b).

Nonetheless, understanding an event process is not just about inducing temporal relations between events or inferring missing steps in an event sequence. As suggested by cognitive studies (Zacks et al., 2001; Zacks and Tversky, 2001; Kurby and Zacks, 2008), a process of events is defined more by the goals, plans, intentions, or traits of its performer, rather than by physical characteristics. For example, a series of events *digging a hole*, *putting in some seeds*, *filling with soil* and *watering the soil*, occurs in a specific sequence since these steps are directed towards the central goal of *planting a plant* by the performer. Similarly, we can tell that *making a dough*, *adding toppings*, *preheating the oven* and *baking the dough* is likely a chain of actions aimed at *cooking pizza*. Indeed, aforementioned studies show that humans understand a plausible event process by hypothesizing the objectives those co-occurring events aim for, or the ultimate consequence the process seeks to accomplish. Accordingly, we suggest that computational methods for event understanding would benefit from conceptualizing the intentions behind the processes. Moreover, inducing intentions is crucial to rich understanding of text (Rashkin et al., 2018), and could potentially support other applications such as commonsense reasoning (Sap et al., 2019), summarization (Daumé III and Marcu, 2006), reading comprehension (Berant et al., 2014) and schema induction (Huang et al., 2016).

To understand the intentions of event processes, the *first* contribution of this paper is to propose a new semantic typing task. The *event process typing* task seeks to retrieve ultra fine-grained type

[*] This work was done when the author was visiting the University of Pennsylvania.

[1] A gloss provides a sense definition for a lexeme.

[2] The contributed learning resources, software and a system demonstration are available at http://cogcomp.org/page/publication_view/915.

[3] A.k.a. event chains (Chambers and Jurafsky, 2008).

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 531–542
Online, November 19-20, 2020.
https://doi.org/10.18653/v1/P17

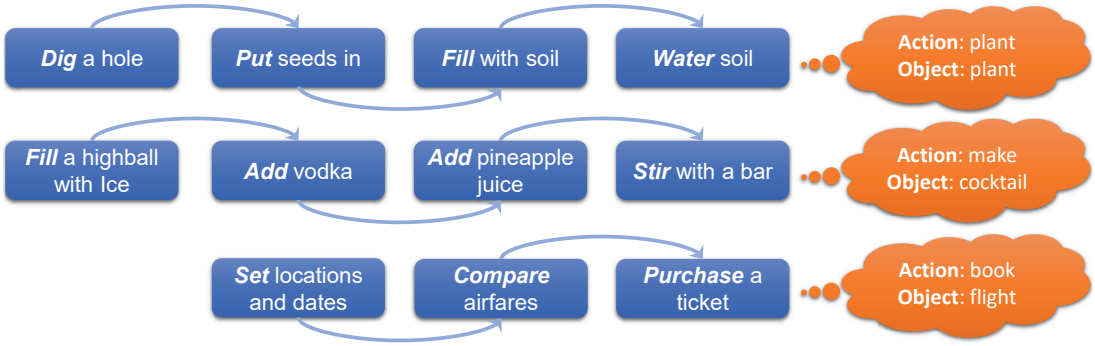

Figure 1: Examples of type inference for event processes.

information to summarize the goal and intention of the associated events. Specifically, each event process is typed along two axes: the *action type* that describes the type of action the process takes, and the *object type* that semantically types the object(s) that the process seeks to affect. Figure 1 shows several accordingly typed event processes. Motivated by recent works on entity typing (Choi et al., 2018; Zhou et al., 2018), our task employs large type vocabularies supporting diverse free-form semantic labels for both axes.

To facilitate related research, we developed a large dataset extracted from wikiHow[4], as the *second* contribution of this paper. This dataset contains over 60,000 processes of primitive events, and features fine-grained action and object type labels for each process. While the dataset aims at creating rich examples of event process intentions, it is also a challenging dataset from two perspectives. First, vocabularies on both type axes are remarkably diverse, giving over 1,000 action type labels and over 10,000 object type labels. And, these fine-grained type vocabularies occur quite sparsely – around 68% of action types and 88% of object types occur fewer than 10 times. This leads to a few-shot learning scenario and, in nearly half of the cases, one-shot. Second, the free-form type labels are generally external to the lexical content of the associated events appearing in a process. Hence, this typing task could not be easily handled with an extractive method (Nenkova and McKeown, 2012).

While the task and dataset pose a non-trivial learning problem, the free-form type system allows for a practical form of indirect supervision based on gloss knowledge. As the *third* contribution, we propose a hybrid learning framework, `P2GT` (i.e., process-to-gloss based typing), to leverage such indirect supervision for event process typing. Instead of directly inferring the multi-axis type labels, we find it to be much easier to seize on the semantic relatedness between the process-gloss pair, as the gloss provides richer semantic information than the label itself. For few-shot cases, gloss definitions also represent useful side information to jump-start inducing labels that are rarely seen or completely unseen in training.

The proposed framework fine-tunes a pre-trained language model to capture the relatedness of an event process and the gloss of types with a ranking task objective. To incorporate more precise gloss information, the training process deploys a word sense disambiguation (WSD) module for both verbs and nouns. Joint learning for both action and object types is enforced to further complement scarce supervision signals. Based on extensive experimental evaluation, the proposed framework exhibits promising performance of inferring the fine-grained multi-axis type information. Specifically, it outperforms a strong RoBERTa-based baseline by 2.4-3.0 folds in *recall@1*. We also show that the incorporated gloss knowledge supports few-shot case prediction, and benefits our model's generalization to out-of-domain event processes.

2 Task and Dataset

We hereby formulate the task of multi-axis event process typing, and introduce the contributed dataset.

2.1 Task Definition

Following Chambers and Jurafsky (2008), we define a process as a sequence of primitive events $P = [e_1, e_2, ..., e_l]$ performed by one common protagonist (or performer). Since the protagonist is shared among the events, each event e_i thereof contains a *predicate* a_i mentioning an action performed

[4] `https://www.wikihow.com/`

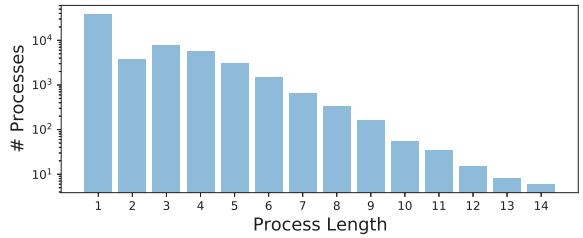

Figure 2: Distribution of process lengths.

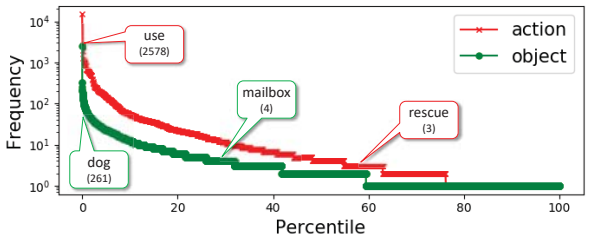

Figure 3: Distribution of action and object types. Number of frequencies are shown in the brackets.

by the protagonist, and an object o_i describing the object(s) that the action is taken upon. The goal is to conceptualize the overall intention behind the process P into two labels, i.e. A from the verb vocabulary that describes the overall action of P, and O from the noun vocabulary that describes what object(s) the process is most likely to affect. Such type inference is important to applications that require commonsense reasoning based on chains of activities, including event-based summarization, narrative prediction and open-domain QA.

2.2 Dataset

We construct a large corpus of typed event processes based on wikiHow – an online wiki-style community containing a collection of professionally edited how-to guideline articles.

Construction A set of the articles are crawled from wikiHow, where each included article describes ordered steps of activities to complete a central goal (e.g. the article "How to book a flight" describes necessary steps to complete an airline booking). Each described step of an article forms a standalone section, which provides an easy-to-consume format for obtaining event processes with clear intentions. We use AllenNLP (Gardner et al., 2018) to perform SRL on section titles of a goal-step article, and extract the `VERB` (predicate a_i) and `ARG1` (object o_i) outputs from the section titles to form the corresponding sequence of (primitive) events. Note that some articles may contain multiple step sequences for the same goal, e.g. booking a flight can be separated to two alternatives, either about booking online or via phone call. In such cases, each alternative is extracted as a separate process. Moreover, we only preserve processes where every primitive event contains both `VERB` and `ARG1`. Any `ARG0`'s are however omitted, since all events in a process share the same protagonist.

To obtain the type information, we first run SRL on the clause after "how to" in the article titles, from which the `VERB` term is seized as the action type label. Then on the `ARG1` output of SRL, we fetch only the lemmatized head word based on dependency parsing and lemmatization (Bird and Loper, 2004). This typically gives us the non-plural noun that represents the object type, whereas other dependents including modifiers are dropped. Consider the clause in "How to make a birthday cake", after *make* is fetched with SRL, the head word *cake* will be preserved from the `ARG1` "a birthday cake", providing an adequately abstracted label for object typing while being consistent to task definition.

Statistics The above effort obtains 62,277 clean event processes, each of which is labeled with both action and object types. Lengths of the processes are varied, for which the distribution is plotted in Figure 2. While the dataset gives a rich variety of instances for processes and intentions, it features a challenging type system for several reasons:

- *Diversity.* The fine-grained type vocabularies consist of 1,336 action types and 10,441 object types. As shown in Figure 3, both sets of labels generally form long-tail distributions.
- *Few-shot cases.* There are 68.3% of action type labels and 88.2% of object type labels occuring fewer than 10 times across all processes. This fact indicates extreme few-shot cases that are challenging to learning and inference.
- *External labels.* In around 91.2% processes, the action type labels are different from the predicates of associated events, while 84.2% of processes have object type labels that do not appear as event objects. Such generally external labels easily cause extractive or sequence-to-label prediction methods to fall short.

3 Process Typing with Gloss Knowledge

In this section, we present our method for the multi-axis event process typing task. The proposed `P2GT` framework conducts learning in three steps. A pre-trained language model is first used to produce the

representations of processes. Then, the gloss information of type vocabularies is encoded as intermediate representations for type labels using the same language model, for which WSD is performed to refine the gloss information of polysemous labels during training. Finally, the language model is fine-tuned with a ranking task objective to capture the association of process-gloss pairs. In the last step thereof, joint learning is performed for typing on both axes to complement the scarce supervision signals, where a process representation is separately projected and handled for action and object types in the latent space. Figure 4 displays the overall model architecture.

In the rest of this section, we introduce the technical details of each step for learning and inference.

3.1 Process Representation

We use the officially released *RoBERTa-base* (Liu et al., 2019) for representations of event processes. RoBERTa improves the original BERT (Devlin et al., 2019) with a modified training procedure. It is considered one of the SOTA models for semantic representation of lexical sequences.

To encode a process P, we concatenate the predicate and object (a_i and o_i) of each event (e_i). Then those contents of all primitive events in P are sequentially concatenated, while the separator token of RoBERTa `</s>` is added between the contents of every consecutive two events. The entire lexical sequence is enclosed between tokens `<s>` and `</s>` to denote the beginning and end of the sequence. Following convention (Bommasani et al., 2020), mean-pooling of hidden states produces the encoded representation of the process, denoted $\mathbf{P}$.

3.2 Label Representation

In our problem setting, directly capturing the association between a process and a free-form type label can be difficult. Hence, we propose a way of indirect supervision by using gloss knowledge as intermediate representations of type labels. The sense definitions in the glosses contain much richer semantic information of the labels themselves. Therefore, leveraging intermediate representations seeks to better characterize the semantic relatedness of processes and labels, especially when the labels are often external to the event content. Glosses also adequately provide side information to jump-start few-shot label representations.

Given a label L for either type axis, we use the same RoBERTa model (with shared parameters)

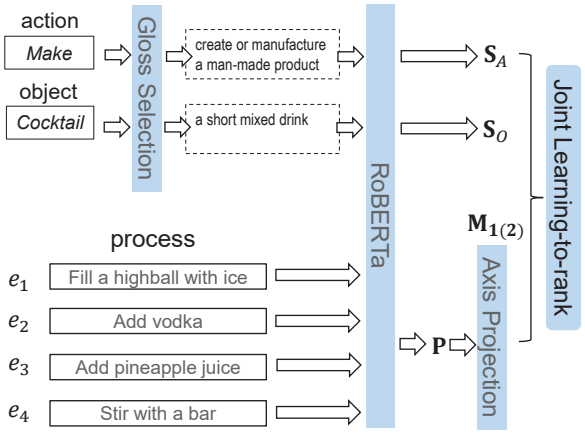

Figure 4: A gloss selection module selects the proper glosses of the training labels. Then a RoBERTa language model captures the event process, and separately generates gloss-based representations for positive and negative sampled labels. The entire learning process conducts joint learning-to-rank on both type axes.

for process representation to encode its gloss sense definition. The mean-pooling result produces the gloss-based label representation denoted $\mathbf{S}_L$. Consider that the verb and noun terms in label vocabularies can be polysemous, we employ either of the following two techniques to select the glosses in the learning phase:

- *Pre-trained WSD models*. One technique is to employ off-the-shelf WSD models that handle both verb and noun senses (Hadiwinoto et al., 2019; Huang et al., 2019). This could more precisely find the right definition for each label given the specific context of a process, and allows each (polysemous) label to have varied representations when typing different processes. During training `P2GT`, we run WSD on the concatenation of type labels $[A, O]$ to select the glosses of A and O for each training case.
- *Most frequent senses (MFS)*. Suppose a WSD model is not available, then the default way is to match a label only to its most frequent (or predominant) sense in sense-annotated corpora (Langone et al., 2004; Camacho-Collados et al., 2016). The MFS method has been a very strong baseline for unsupervised WSD (Tripodi and Navigli, 2019), as it is natural in language text that words generally express their predominant senses in most cases (McCarthy et al., 2007). Specifically for our task, the purpose is not to infer the exact sense, but rather generating a semantically rich (and allowably noisy) repre-

sentation for type labels. In practice, we find this simple technique to perform reasonably well as we type the event processes (§4.3).

Besides these two techniques, we also tried others to represent a label, including concatenating all its gloss sense definitions, or concatenating most frequent two or more senses. They however do not perform as well as the aforementioned two techniques, hypothetically due to the noise introduced to label representations. More technical details about WSD and the source inventory of glosses are to be described in Experiments (§4.1).

3.3 Learning Objective

Let (P, A, O) be a process P denoted by action and object labels A and O, our model captures the semantic associations between a RoBERTa encoded process $\mathbf{P}$ and label glosses $\mathbf{S}_A$ and $\mathbf{S}_O$ by optimizing a ranking task objective. In detail, we define the margin ranking loss for action typing as

$$L_1^P = [s(\mathbf{M}_1 \cdot \mathbf{P}, \mathbf{S}_{A'}) - s(\mathbf{M}_1 \cdot \mathbf{P}, \mathbf{S}_A) + \gamma_1]_+ ,$$

and that for object typing as

$$L_2^P = [s(\mathbf{M}_2 \cdot \mathbf{P}, \mathbf{S}_{O'}) - s(\mathbf{M}_2 \cdot \mathbf{P}, \mathbf{S}_O) + \gamma_2]_+ .$$

$[x]_+$ thereof denotes the positive part of the input x (i.e. $\max(x, 0)$). γ_1 and γ_2 are two positive constant margins. $\mathbf{M}_1$ and $\mathbf{M}_2$ correspond to two learnable linear projections dedicated to the two type axes respectively. $s(\cdot)$ is the cosine similarity measure. $A' \in V$ and $O' \in N$ are negative-sample labels. In the setting with WSD deployed in training, negative sampling randomly fetches from all glosses of labels that appear in the training data, except for the gloss(es) of the positive label. This allows chances for different glosses of a polysemous label to serve as negative samples. Otherwise, the only gloss of every negative-sample label is utilized in the MFS setting.

The eventual learning objective is to optimize the following joint loss, where D denotes the dataset:

$$L = \frac{1}{|D|} \sum_{P \in D} L_1^P + L_2^P .$$

Note that we have incorporated different margins that can trade-off between L_1^P and L_2^P, hence we do not use weight coefficients to combine these two terms of ranking losses.

3.4 Inference

The inference phase of `P2GT` performs a nearest neighbor search to type a process P. Let $\mathbf{M}$ refer to either $\mathbf{M_1}$ for the action type or $\mathbf{M_2}$ for the object type, our framework finds the gloss-based label representation that is closest to $\mathbf{M} \cdot \mathbf{P}$ from the corresponding vocabulary. Specifically for the setting with polysemous label representations, it is sufficient to consider for each label only its gloss that is embedded most closely to $\mathbf{M} \cdot \mathbf{P}$, so as to not redundantly consider candidate labels.

4 Experiments

To evaluate the proposed `P2GT` framework for event process typing, we conduct several experiments on the contributed dataset, and compare with a wide selection of baseline methods (§4.1-§4.3). A case study is also provided on typing processes from an external dataset (§4.4).

4.1 Experimental Settings

Similar to Rashkin et al. (2018), we randomly separate the 62,277 processes into a training/dev./test set using an 80/10/10% split. We report three ranking metrics, i.e. *MRR* (mean reciprocal rank), *recall@1* and *recall@10*. All metrics are preferred to be higher to indicate better performance.

We compare our framework with a number of its variants by performing the following modification: (i) Simplifying the framework by separately learning for the two type axes, instead of performing joint training; (ii) Different settings of gloss selection in training, using either WSD or FSM; (iii) Different information used to represent each primitive event e_i, e.g., only using either a_i or o_i (marked with *partial event*) according to the type axis, instead of using both. Besides, we compare with sequence-to-label (S2L) generators (Rashkin et al., 2018). A method of such is an encoder-decoder architecture trained to directly map from sequences to unigrams of the type vocabulary, which is originally used by recent work (Rashkin et al., 2018) to infer intentions from a single-clause description of a primitive event. Specifically, we employ three variants of S2L using different encoders. Besides one based on RoBERTa (marked as S2L-RoBERTa), the two others are the BiGRU encoder (S2L-BiGRU) and mean-pooling encoder (S2L-mean) with Skip-Gram word embeddings used by Rashkin et al. (2018). Note that to train S2L models, the original paper uses an cross-entropy loss

Type axes	Action			Object		
Metrics	*MRR*	*recall@1*	*recall@10*	*MRR*	*recall@1*	*recall@10*
S2L-mean-pool	3.72	1.96	5.95	1.01	0.80	1.66
S2L-BiGRU	7.94	4.40	12.71	4.20	2.72	6.19
S2L-RoBERTa	8.36	5.31	14.69	4.88	3.24	8.10
Single P2GT-MFS (partial event)	18.03	14.36	17.16	10.36	6.37	17.64
Single P2GT-WSD (partial event)	18.07	14.05	17.82	10.72	6.68	18.03
Single P2GT-MFS	24.10	19.67	32.40	13.71	8.86	23.09
Single P2GT-WSD	25.83	19.93	37.50	14.19	9.32	24.84
Joint P2GT-MFS	28.57	20.63	**43.14**	15.26	10.62	25.01
Joint P2GT-WSD	**29.11**	**21.21**	42.84	**15.70**	**11.07**	**25.51**

Table 1: Results (in percentage) for multi-axis event process typing. S2L methods with different encoding techniques are original or adopted from Event2Mind (Rashkin et al., 2018). *partial event* marks the cases where only a_i (or o_i) is encoded for each event e_i in the process to infer the action (or object) type. *Joint* or *Single* denotes whether to use joint training for both type axes or not. MFS and WSD marks ways of gloss selection in training.

Event processes	**Predictions**
Position yourself ⇒ Trim your eyebrows ⇒ Use the eyebrow pencil	*A*: strop, **highlight**, **thread**, blunt, *sharpen* *O*: **unibrow**, **eyebrow**, straightener, **eyelash**, razor
Learn how to strum ⇒ Use a metronome ⇒ Play to recorded songs ⇒ Grow skills	*A*: **play**, **practice**, *strum*, *tune*, box *O*: cymbal, **mandolin**, **guitar**, **dulcimer**, flute
Get a referral ⇒ Verify the specialist 's qualifications ⇒ Ask questions ⇒ Assess whether treatment is working	*A*: **find**, **choose**, use, apply, drink *O*: **therapist**, *physician*, **specialist**, *surgeon*, *psychiatrist*
Go to DMV ⇒ Take photos ⇒ Take vision test ⇒ Take permit test ⇒ Take road test	*A*: **obtain**, *verify*, explore, drive, polish *O*: **license**, check, visa, carfax, toll
Create your clan ⇒ Maintain your clan ⇒ Add another clan ⇒ Defend the borders ⇒ Do the hunting	*A*: **adopt**, **create**, **spawn**, *homestead*, become *O*: **clan**, **warrior**, *headhunter*, skirmish, *necrons*
Prepare the jack ⇒ Locate the filler hole ⇒ Fill the oil ⇒ Close the filler hole	*A*: *bleed*, *grease*, **add**, **fill**, inflate *O*: **oil**, pump, *biodiesel*, blowing, choke

Table 2: Top 5 predictions on examples of test cases by Joint-P2GT-WSD. Ground truths are underscored, reasonably correct labels are boldfaced, and close ones are italic. Few-shot labels appearing ≤10 times are in blue.

to model the distribution of unigrams. We instead train the process encoder to directly fit the embeddings of label surface forms similar to a reverse dictionary (Hill et al., 2016; Chen et al., 2019), which offers notably better performance.

4.2 Model Configuration

We use sense definitions from WordNet (Miller, 1995) to define the labels. While such glosses cover all verbs in the action type vocabularies, there are 7.92% of processes where object type labels do not find WordNet senses. For each such case, we select from WordNet the lexeme that is embedded most closely to the label, and use the predominant sense of that lexeme to generate the label representation. For the training setting with WSD, we use the BERT-NN model (Hadiwinoto et al., 2019), which is one of the SOTA WSD methods that is trained on the SemCor corpus (Langone et al., 2004). In fact, despite the ones that are dedicated to nouns (Scarlini et al., 2020; Pasini and Navigli, 2017), other SOTA methods for WSD (Huang et al., 2019; Maru et al., 2019; Tripodi and Navigli, 2019) may also apply to our framework, for which we leave the investigation to future work.

We use AMSGrad (Reddi et al., 2018) to optimize the learning objective, with the learning rate set to 0.0001. The batch size is set to 64 to fit the memory of one Titan RTX 6000 GPU. Training is limited to 50 epochs that is enough for all models to converge. Margins are chosen from 0.0 to 0.4 with a step of 0.1, based on *recall@1* performance on the dev. set. Accordingly, $\gamma_1 = 0.2$ and $\gamma_2 = 0.1$ are selected for Single P2GT methods, while both margins are set to 0.1 for the joint-learning P2GT.

4.3 Results

We report the results of event process typing on both axes in Table 1, whereof the results for typing actions are generally better than those for the object

Event processes	Predictions
Make explosive materials ⇒ Obtain a container ⇒ Obtain shrapnel ⇒ Install a trigger	*A*: **detonate**, **assemble**, blacken *O*: **grenade**, **blaster**, **mine**
Ignore order ⇒ Enter area ⇒ Enforce blockade ⇒ Force to retreat from area	*A*: **conquer**, *disarm*, **invade** *O*: *barrier*, **soldier**, **fortress**
Capture two opposition posts ⇒ Kill many fighters ⇒ Destroy three armed trucks ⇒ Confiscate artillery guns	*A*: *kill*, **demolish**, **fight** *O*: *melee*, **conflict**, **stronghold**
Cooperate with the counsel investigation ⇒ Open his remarks ⇒ Apologize many times ⇒ Try to restore public trust	*A*: *respond*, disagree, **accept** *O*: *apology*, *disagreement*, **slander**
Travel in a presidential motorcade ⇒ Be shot once in the back ⇒ Be taken to hospital ⇒ Be pronounced dead	*A*: *survive*, **die**, tackle *O*: **assassin**, crash, *roadkill*
Give advance notice ⇒ Give notice ⇒ Issue dividends	*A*: **honor**, **pay**, **reward** *O*: *finance*, **equity**, **subsidy**
Target quotes ⇒ Target shares quotes ⇒ Ask to clarify offer ⇒ Challenge to merge agreement ⇒ Challenge to merge businesses	*A*: **compare**, **maximize**, **negotiate** *O*: *prospectus*, **quote**, **settlement**
Clean windows ⇒ Buy plants ⇒ Hang pictures ⇒ Paint walls ⇒ Carpet floors	*A*: **redecorate**, **decorate**, *refurbish* *O*: **room**, **bedroom**, *makeover*

Table 3: Case study for typing event processes in the news domain. The predictions are given by Joint P2GT-WSD trained on our full dataset. Each case is given top 3 predictions on both axes, whereof reasonably correct ones are boldfaced, and relevant ones are italic. Few-shot labels appearing up to 10 times in our dataset are in blue.

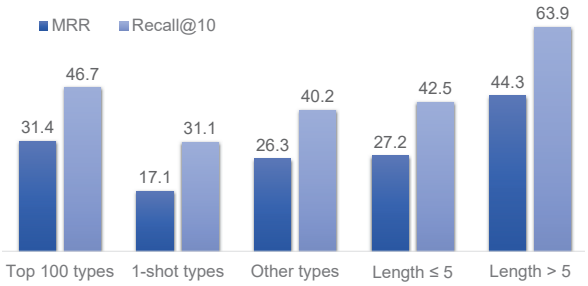

Figure 5: Comparison of action typing on different portions of the test set. We compare results by Joint P2GT-MFS for top 100 frequent types, one-shot types and the rest, as well as results on processes of different lengths. 5 is the median length of processes in the dataset.

axis due to different sizes of candidate spaces.

The results by the S2L baseline methods show that incorporating pre-trained RoBERTa offers noticeably better performance than other encoding techniques. However, it is drastically superseded by the Single P2GT setting without WSD-based gloss selection. When typing the action, with the same representation of event processes, P2GT supercedes S2L-RoBERTa with an absolute increase of *MRR* by 15.47% (ca. 1.88× relative increase), and that of *recall@1* by 14.36% (ca. 2.7× relative increase). For object typing, the absolute increments are 8.83% in *MRR* (ca. 1.82× relatively) and 5.62% in *recall@1* (ca. 1.72× relatively). This also indicates that incorporating gloss knowledge for label representations brings along the most substantial improvement to the task, inasmuch as glosses attain rich semantic information to jump-start type labels and realistically support sequence-level learning-to-rank.

On the other hand, incorporating WSD for gloss selection in training slightly causes absolute increment of up to 1.73% in *MRR* for action typing and 0.48% in *MRR* for object typing. This is partly due to that the predominant sense definitions can generally seize precise or close definitions to represent the labels in most cases. Hence, sense selection provides lesser improvement, especially when the candidate space is large. It is noteworthy that, partially giving the predicate or object information of associated events is not enough to infer the type information. In fact, the performance drop is in accordance with human cognition, as giving a chain of predicates only or objects only is not enough to predict the intentions of the event process. Consider the first example in Figure 1, observing only a chain of the protagonist's actions *dig*, *put*, *fill* and *water*, or only the participating objects *hole*, *seed* and *soil* are clearly not enough to infer the overall action and the objective that directs the entire process. Accordingly, the partial event representation causes significant performance drop of 3.35-7.76% in terms of *MRR*, and 2.49-5.86% in terms of *recall@1*. Lastly, joint learning brings along performance gain by 1.51-4.47% in *MRR* and 0.96-1.76% in *recall@1*, indicating the effectiveness of lever-

aging complementary supervision signals. Note that the evaluation strictly enforces *exact match* in large candidate spaces, thus underestimating the system performance. While it is difficult for the model to always rank the ground-truth labels on the top, it can often infer reasonably close labels as top predictions, for which a couple of examples are shown in Table 2.

To understand how differently our method performs on processes of different characteristics, we additionally perform an error analysis. In Figure 5, we compare action prediction by P2GT with joint learning and MFS-based gloss selection on different proportions of the test set. It is expected that the performance on more frequent labels are better than on infrequent ones due to ampler training cases. Nonetheless, on the extremely challenging one-shot cases, our method still performs reasonably well, and drastically excels the overall results by baseline methods. Additionally, we observe that typing longer event processes is easier, as they provide more contextual information of associated events to help inferring the central goal. In contrast, as short processes are less informative, *MRR* scores for those sized 2 and 3 are 24.17% and 25.41%.

4.4 Case Study

We conduct a case study using a subset of the NYT narrative cloze dataset provided by Lee and Goldwasser (2019). This dataset includes a series of event processes extracted from news reports, and we use those processes to showcase the prediction of P2GT on out-of-domain processes. According to Table 3, although the content and concepts of processes in military and political news are mostly irrelevant to the intentional goals in our dataset, P2GT is able to infer reasonably correct type information on both axes. Particularly, many of the top predictions give few-shot labels. This further exhibits that gloss knowledge is effective to improve the generalization of the typing model, both in terms of handling domain shifting and few-shot cases. Specifically, the case study also points out the direction of our further study on how well gloss-based label representations can generally benefit domain adaptation and few-shot learning in natural language understanding tasks.

5 Related Work

Prediction tasks on event processes have attracted much attention recently, while many existing works focus on extraction and completion of event processes. For example, Radinsky et al. (2012; 2013) mine sequences of frequently co-ocurring events from multiple temporally connected documents, and use the sequence knowledge to predict the future event(s) of a process. Berant et al. (2014) propose to extract biological processes with SRL, and help machine reading comprehension for biological articles. A series of other works learn for sequential event prediction using language models (Chaturvedi et al., 2017; Peng et al., 2019) or association rules (Letham et al., 2013), and further cope with downstream tasks such as narrative cloze tests. On the contrary, fewer efforts have been made for inferring the intentions or central goals behind a composite of events. A recent work by Rashkin et al. (2018) is particularly relevant to this topic, which learns a sequence-to-label generator to predict the intention of one primitive event based on a single-clause description. This is however essentially different from our focus on processes of multiple events.

Semantic typing has been investigated for language components other than events, such as entities and word senses. Due to the large body of work in this line of research, we can only provide a highly selected summary for most recent outcomes. For entity typing, recent research has coped with highly challenging problem settings. Those include few-shot or zero-shot typing with contextual distant supervision (Zhou et al., 2018) and description-based label embeddings (Obeidat et al., 2019). Others realize ultra-fine type systems with the help of head-word supervision (Choi et al., 2018), hierarchical learning-to-rank (Chen et al., 2020) and structured label representations (Xiong et al., 2019; Hao et al., 2019). Several aforementioned techniques are also employed to supersense typing (Levine et al., 2020; Peters et al., 2019) and POS tagging (Owoputi et al., 2013). In terms of type labeling, our work is inspired by Choi et al. (2018)'s way of leveraging free-form lexemes for ultra-fine entity types. Nevertheless, besides typing on a different modality, our work is also distinguished in the multi-axis typing system, and the way of leveraging gloss-based indirect supervision.

Representation learning of gloss knowledge has been incorporated in various tasks. A number of works encode gloss definitions for monolingual (Hill et al., 2016; Noraset et al., 2017; Pilehvar, 2019; Hedderich et al., 2019) and cross-lingual

(Chen et al., 2019; Zhang et al., 2020a) reverse dictionary prediction, as well as out-of-vocabulary lexical representation (Kumar et al., 2019; Prokhorov et al., 2019; Bahdanau et al., 2017). Definitions have also been leveraged to generate zero-shot entity representations in knowledge graphs (Kartsaklis et al., 2018; Chen et al., 2018; Long et al., 2017). Some other works inject gloss representations to improve WSD (Huang et al., 2019; Luo et al., 2018; Blevins and Zettlemoyer, 2020). GlossBERT (Huang et al., 2019) thereof formalizes the WSD problem as classifying context-gloss pairs. Our learning approach on process-gloss pairs is connected to that approach, whereas we handle a learning-to-rank objective, and make inference in a much larger candidate space than the sense space of a single word.

6 Conclusion

We propose a new task of event process understanding, by semantically typing the intended action of an event process and the object(s) it seeks to affect. To facilitate research in this direction, we develop a new dataset, gathering over 60 thousand event processes with ultra fine-grained type vocabularies. We further propose a hybrid learning framework, which leverages indirect supervision from gloss knowledge. The proposed `P2GT` framework fine-tunes RoBERTa to capture the association of process-gloss pairs. Label gloss selection mechanisms and joint training are incorporated to further improve the performance. Experiments show that `P2GT` offers promising performance on inferring the fine-grained type information, and exhibits satisfactory generalizability on out-of-domain event processes.

For future work, we are interested in identifying salient events in processes, i.e., those that most significantly define the central goals. Incorporating process typing into downstream tasks such as summarization and commonsense QA is also an important direction.

Acknowledgement

We appreciate the anonymous reviewers for their insightful comments. Also, we would like thank Jennifer Sheffield and other members of the UPenn Cognitive Computation Group for giving suggestions that improved the manuscript.

This research is supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the BETTER Program, and by Contract FA8750-19-2-1004 with the US Defense Advanced Research Projects Agency (DARPA). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Dzmitry Bahdanau, Tom Bosc, Stanislaw Jastrzebski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. 2017. Learning to compute word embeddings on the fly. *CoRR*, abs/1706.00286.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story Comprehension for Predicting What Happens Next. In *In proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. 2018. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3998–4004.

Muhao Chen, Yingtao Tian, Haochen Chen, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. 2019. Learning to represent bilingual dictionaries. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 152–162, Hong Kong, China. Association for Computational Linguistics.

Tongfei Chen, Yunmo Chen, and Benjamin Van Durme. 2020. Hierarchical entity typing via multi-level learning to rank. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8465–8475, Online. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia. Association for Computational Linguistics.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Sydney, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint Inference for Event Timeline Construction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Junheng Hao, Muhao Chen, Wenchao Yu, Yizhou Sun, and Wei Wang. 2019. Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1709–1719.

Michael A. Hedderich, Andrew Yates, Dietrich Klakow, and Gerard de Melo. 2019. Using multi-sense vector embeddings for reverse dictionaries. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 247–258, Gothenburg, Sweden. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R. Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–268, Berlin, Germany. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Prateek Jindal and Dan Roth. 2013. Extraction of Events and Temporal Expressions from Clinical Narratives. *Journal of Biomedical Informatics (JBI)*.

Dimitri Kartsaklis, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Mapping text to knowledge graph entities using multi-sense LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1959–1970, Brussels, Belgium. Association for Computational Linguistics.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Christopher A Kurby and Jeffrey M Zacks. 2008. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79.

Helen Langone, Benjamin R. Haskell, and George A. Miller. 2004. Annotating WordNet. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 63–69, Boston, Massachusetts, USA. Association for Computational Linguistics.

I-Ta Lee and Dan Goldwasser. 2019. Multi-relational script learning for discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4214–4226, Florence, Italy. Association for Computational Linguistics.

Benjamin Letham, Cynthia Rudin, and David Madigan. 2013. Sequential event prediction. *Machine learning*, 93(2-3):357–380.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Teng Long, Emmanuel Bengio, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2017. World knowledge for reading comprehension: Rare entity prediction with hierarchical LSTMs using external descriptions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 825–834, Copenhagen, Denmark. Association for Computational Linguistics.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411, Brussels, Belgium. Association for Computational Linguistics.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3534–3540, Hong Kong, China. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark. Association for Computational Linguistics.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Rasha Obeidat, Xiaoli Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Description-based zero-shot fine-grained entity typing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 807–814, Minneapolis, Minnesota. Association for Computational Linguistics.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.

Tommaso Pasini and Roberto Navigli. 2017. Train-o-Matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88, Copenhagen, Denmark. Association for Computational Linguistics.

Haoruo Peng, Qiang Ning, and Dan Roth. 2019. KnowSemLM: A knowledge infused semantic language model. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 550–562, Hong Kong, China. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Karl Pichotta and Raymond Mooney. 2014. Statistical script learning with multi-argument events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 220–229, Gothenburg, Sweden. Association for Computational Linguistics.

Mohammad Taher Pilehvar. 2019. On the importance of distinguishing word meaning representations: A case study on reverse dictionary mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2151–2156, Minneapolis, Minnesota. Association for Computational Linguistics.

Victor Prokhorov, Mohammad Taher Pilehvar, Dimitri Kartsaklis, Pietro Lio, and Nigel Collier. 2019. Unseen word representation by aligning heterogeneous lexical semantic spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6900–6907.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918. ACM.

Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, (AAAI)*, pages 8758–8765. AAAI Press.

Rocco Tripodi and Roberto Navigli. 2019. Game theory meets embeddings: a unified framework for word sense disambiguation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 88–99, Hong Kong, China. Association for Computational Linguistics.

Wenhan Xiong, Jiawei Wu, Deren Lei, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Imposing label-relational inductive bias for extremely fine-grained entity typing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 773–784, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeffrey M Zacks, Todd S Braver, Margaret A Sheridan, David I Donaldson, Abraham Z Snyder, John M Ollinger, Randy L Buckner, and Marcus E Raichle. 2001. Human brain activity time-locked to perceptual event boundaries. *Nature neuroscience*, 4(6):651–655.

Jeffrey M Zacks and Barbara Tversky. 2001. Event structure in perception and conception. *Psychological bulletin*, 127(1):3.

Lei Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020a. Multi-channel reverse dictionary model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 312–319.

Tianran Zhang, Muhao Chen, and Alex Bui. 2020b. Diagnostic prediction with sequence-of-sets representation learning for clinical event. In *Proceedings of the 18th International Conference on Artificial Intelligence in Medicine (AIME)*.

Ben Zhou, Daniel Khashabi, Chen-Tse Tsai, and Dan Roth. 2018. Zero-shot open entity typing as type-compatible grounding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2065–2076, Brussels, Belgium. Association for Computational Linguistics.