## ACE Event Extraction

- The Automatic Content Extraction (ACE) evaluations have included several information extraction tasks, including *entity recognition, value recognition, time recognition, relation extraction*, and *event extraction*.
- The **2005 ACE event extraction task** included 8 general event types and 33 event subtypes, with 35 possible event roles.
- An event mention is limited in scope to one sentence. The role fillers for each event mention must be identified within the event sentence. (*This is different from template-based event extraction, which extracts event information from an entire document*.)
- Each event has **event role fillers**, which include participants, objects, and other arguments like date and time. Importantly, an event can have multiple fillers (arguments) for a single role.

1

## Event Roles vs. Semantic Roles

- **Semantic roles** (akin to thematic roles) capture arguments to a *target* word and represent the roles played in the action or concept directly expressed by the *target*.

  Example*: John was arrested by police for the murder of George.*

  TARGET = *arrested* → THEME = *John* AGENT = *police*
  TARGET = *murder* → THEME = *George*

- **Event roles** capture arguments associated with a *trigger* word (in ACE) describing an action or concept associated with a higher-level event.

  Example*: John was arrested by police for the murder of George.*

  EVENT = **DIE** →
  PERPETRATOR = *John*
  VICTIM = *George*

2

## ACE Terminology

- **Entity**: an object that belongs to a semantic category.
- **Entity mention**: a reference to an entity
- **Timex**: a time expression (e.g., day, year, date)
- **Event mention**: a phrase or sentence that describes the occurrence of an event.
- **Event trigger**: the main word that most clearly expresses an occurrence of a relevant event.
- **Event mention arguments (event role fillers)**: entity mentions that are involved in an event and their relation to the event.

3

## ACE 2005 Entity Types and Subtypes

Table 1 ACE05 Entity Types and Subtypes

| Type | Subtypes |
|---|---|
| FAC (Facility) | Airport, Building-Grounds, Path, Plant, Subarea-Facility |
| GPE (Geo-Political Entity[3]) | Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province |
| LOC (Location) | Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body |
| ORG (Organization) | Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports |
| PER (Person) | Group, Indeterminate, Individual |
| VEH (Vehicle) | Air, Land, Subarea-Vehicle, Underspecified, Water |
| WEA (Weapon) | Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified |

An **entity mention** can be a proper name, nominal, or pronoun.

4

## ACE 2005 Event Types and Subtypes

Table 7 ACE05 Event Types and Subtypes

| Types | Subtype |
|---|---|
| Life | Be-Born, Marry, Divorce, Injure, Die |
| Movement | Transport |
| Transaction | Transfer-Ownership, Transfer-Money |
| Business | Start-Org, Merge-Org, Declare-Bankruptcy, End-Org |
| Conflict | Attack, Demonstrate |
| Contact | Meet, Phone-Write |
| Personnel | Start-Position, End-Position, Nominate, Elect |
| Justice | Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon |

5

## ACE 2005 Event Roles (from Liao's dissertation)

| | | | |
|---|---|---|---|
| Person | Place | Buyer | Seller |
| Beneficiary | Price | Artifact | Origin |
| Destination | Giver | Recipient | Money |
| Org | Agent | Victim | Instrument |
| Entity | Attacker | Target | Defendant |
| Adjudicator | Prosecutor | Plaintiff | Crime |
| Position | Sentence | Vehicle | Time-After |
| Time-Before | Time-At-Beginning | Time-At-End | Time-Starting |
| Time-Ending | Time-Holds | Time-Within | |

**Table 1.2 - 35 Argument roles defined by ACE 2005**

6

## ACE Event Extraction Example

*Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment.*

Trigger = *quit (RESIGNATION)*

Arguments (Roles)

- Person: *Barry Diller*
- Organization: *Vivendi Universal Entertainment*
- Position: *chief*
- Time-within: *Wednesday*

7

## ACE Event Extraction Example

*Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment.*

Trigger = *quit* [for End-Position event]

Arguments (Roles)

- Person: *Barry Diller*
- Organization: *Vivendi Universal Entertainment*
- Position: *chief*
- Time-within: *Wednesday*

8

## Multiple Triggers Example

*Three murders occurred in France today, including the senseless slaying of Bob Cole and the assassination of Joe Westbrook.*

| Event | Trigger | Place | Victim | Time |
|---|---|---|---|---|
| DIE | *murder* | *France* | -- | *today* |
| DIE | *slay* | *France* | *Bob Cole* | *today* |
| DIE | *assassinate* | *France* | *Joe Westbrook* | *today* |

9

## Multiple Triggers Example

*Three murders occurred in France today, including the senseless slaying of Bob Cole and the assassination of Joe Westbrook.*

| Event | Trigger | Place | Victim | Time |
|---|---|---|---|---|
| DIE | *murder* | *France* | -- | *today* |
| DIE | *slay* | *France* | *Bob Cole* | *today* |
| DIE | *assassinate* | *France* | *Joe Westbrook* | *today* |

10

## Sentence-level Event Extraction System

[Grishman, Westbrook, & Meyers, 2005] developed a sentence-level event extraction system for ACE 2005.

- Using training data, event patterns are learned from sequences of constituent heads that separate a trigger word and its arguments.
- A **Trigger Labeler** uses the patterns to distinguish event mentions from non-event mentions and classify each mention by type.
- An **Argument Detector** is trained with a MaxEnt classifier to distinguish arguments from non-arguments with respect to a trigger word.
- A **Role Classifier** is trained with MaxEnt to label detected arguments with respect to event roles.
- A **Reportable-Event Classifier** is trained with MaxEnt to determine whether a potential trigger, event type, and set of arguments are describing a true event occurrence.

11

## Sentence-level Extraction Pipeline

1. Each test document is searched for instances of trigger words that occurred in the training documents.
2. For each trigger word, the *patterns* learned for that trigger are applied to identify arguments (with role labels) of the trigger.
3. The argument detector is applied to the remaining entity mentions in the sentence to look for *more* arguments.
4. If new arguments are found, the role classifier is applied to assign a role to each one.
5. Finally, the reportable-event classifier is applied to the entire context to decide whether it is truly an event.

[Ji & Grishman, 2008] and [Liao & Grishman, 2010] use this system as a component in their document-level event extraction pipelines.

12

## Sentence-Level vs. Document-Level

*Sentence-Level Event Extraction*:

Traditionally, most systems have extracted information about an event from an isolated sentence. Each sentence in a document is processed independently of the others.

*Document-Level Event Extraction*:

Recently, researchers have begun to incorporate discourse properties and information about associations across sentences in a document to improve event extraction performance.

13

## Enforcing Consistency

[Gale, Church, & Yarowsky, 1992] identified the widely recognized **One Sense Per Discourse** heuristic: within a discourse, instances of the same word have a strong tendency to share the same sense.

[Ji & Grishman] made a related observation that strong sense and event role consistency exists across related documents.

**One Trigger Sense Per Cluster:** In topically-related documents, event trigger words have a strong tendency to share the same sense.

**One Argument Role Per Cluster:** in topically-related documents, an entity has a strong tendency to participate in the same event role (argument).

14

## Motivating Observations

1. *"Within a document, there is a strong trigger consistency: if one instance of a word triggers an event, other instances of the same word will trigger events of the same type."*

   True > 99.4% of the time in the ACE corpus.

2. *"Normally one entity, if it appears as an argument of multiple events of the same type in a single document, is assigned the same role each time."*

   True > 97% of the time in the ACE corpus.

15

## Shared Trigger Sense Example

Test Sentence:

*Most US army commanders believe it is critical to pause the breakneck advance towards Baghdad to secure the supply lines and make sure weapons are operable and troops resupplied …*

Related Document:

*British and US forces report gains in the advance on Baghdad and take control of Umm Qasr, despite a fierce sandstorm which slows another flank.*

"*Advance toward*" wasn't in the training data, but "*advance on*" was. Identifying "*advance*" as a Movement_Transport event trigger in a related document suggests the same sense for in the new document.

16

## Correcting Trigger Senses

Test Sentence:

*But few at the Kremlin forum suggested that Putin's own standing among voters will be hurt by Russia's apparent diplomacy failures.*

Related Document:

*Putin boosted ties with the United States by throwing his support behind its war on terrorism after the Sept. 11 attacks, but the Iraq war has hurt the relationship.*

Originally, *hurt* was mistakenly identified as a Life_Injure event in the test sentence because it is a common trigger word for that event type.

But … *hurt* is never a trigger for Life_Injure events in the topically related documents, so that trigger label can be discarded.

17

## Shared Argument Role Example

Test Sentence:

*Vivendi earlier this week confirmed months of press speculation that it planned to shed its entertainment assets by the end of the year.*

Related Documents:

*Vivendi has been trying to sell assets to pay off huge debt, estimated at the end of last month at more than $13 billion.*

*Under the reported plans, Blackstone group would buy Vivendi's theme park division, including Universal Studios Hollywood, …*

Originally, *"Vivendi"* was not recognized as a seller in the test document.

But it was extracted as a seller in several topically related documents, which suggests it is likely to be a seller in the new documents too.

18

## Cross-Sentence and Cross-Document Extraction

[Ji & Grishman, 2008] created a system that improved event extraction output by making inferences that enforce consistency across within-document sentences as well as related documents.

A pipeline architecture gradually refines the output:

1. *sentence-level event extraction*
2. *document-level (cross-sentence) inferences*: rules enforce consistency across sentences in the same document.
3. *cross-document inferences*: an IR system retrieves related documents and rules enforce consistency with these documents.

19

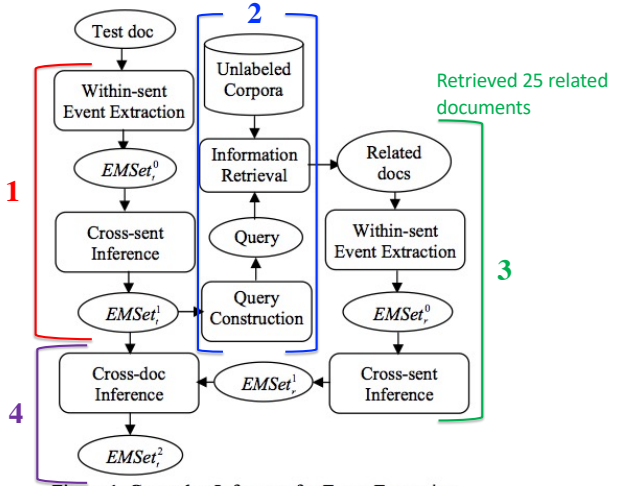


Figure 1. Cross-doc Inference for Event Extraction

20

## Empirical Evidence for Cluster Heuristics

| Candidate Triggers | | Event Type | Perc./Freq. as trigger in ACE training corpora | Perc./Freq. as trigger in test document | Perc./Freq. as trigger in test + related documents |
|---|---|---|---|---|---|
| Correct Event Triggers | *advance* | Movement_Transport | 31% of 16 | 50% of 2 | 88.9% of 27 |
| | *fire* | Personnel_End-Position | 7% of 81 | 100% of 2 | 100% of 10 |
| | *fire* | Conflict_Attack | 54% of 81 | 100% of 3 | 100% of 19 |
| | *replace* | Personnel_End-Position | 5% of 20 | 100% of 1 | 83.3% of 6 |
| | *form* | Business_Start-Org | 12% of 8 | 100% of 2 | 100% of 23 |
| | *talk* | Contact_Meet | 59% of 74 | 100% of 4 | 100% of 26 |
| Incorrect Event Triggers | *hurt* | Life_Injure | 24% of 33 | 0% of 2 | 0% of 7 |
| | *execution* | Life_Die | 12% of 8 | 0% of 4 | 4% of 24 |

21

## Rules for Within-Document Consistency

**Rule (1): Remove Triggers and Arguments with Low Local Confidence**

If *LConf(trigger, etype)* < $\delta_1$, then delete the whole event mention *EM*;
If *LConf(arg, etype)* < $\delta_2$ or *LConf(arg, etype, role)* < $\delta_3$, then delete *arg*.

**Rule (2): Adjust Trigger Classification to Achieve Document-wide Consistency**

If *XSent-Trigger-Margin(trigger)* > $\delta_4$, then propagate the most frequent *etype* to all event mentions with *trigger* in the document; and correct roles for corresponding arguments.

**Rule (3): Adjust Trigger Identification to Achieve Document-wide Consistency**

If *LConf(trigger, etype)* > $\delta_5$, then propagate *etype* to all unlabeled strings *trigger* in the document.

**Rule (4): Adjust Argument Identification to Achieve Document-wide Consistency**

If *LConf(arg, etype)* > $\delta_6$, then in the document, for each sentence containing an event mention *EM* with *etype*, add any unlabeled mention in that sentence with the same head as *arg* as an argument of *EM* with *role*.

**Rule 1**: removes questionable triggers
**Rule 2**: fixes event types and roles (hopefully)
**Rules 3 & 4**: discover more triggers and arguments

22

## Rules for Cluster-Wide Consistency

**Rule (5): Remove Triggers and Arguments with Low Cluster-wide Confidence**

If *XDoc-Trigger-Freq (trigger, etype)* < $\delta_7$, then delete *EM*;
If *XDoc-Arg-Freq(arg, etype)* < $\delta_8$ or *XDoc-Role-Freq(arg, etype, role)* < $\delta_9$, then delete *arg*.

**Rule (6): Adjust Trigger Classification to Achieve Cluster-wide Consistency**

If *XDoc-Trigger-Margin(trigger)* > $\delta_{10}$, then propagate most frequent *etype* to all event mentions with *trigger* in the cluster; and correct roles for corresponding arguments.

**Rule (7): Adjust Trigger Identification to Achieve Cluster-wide Consistency**

If *XDoc-Trigger-BestFreq (trigger)* > $\delta_{11}$, then propagate *etype* to all unlabeled strings *trigger* in the cluster, override the results of Rule (3) if conflict.

**Rule (8): Adjust Argument Classification to Achieve Cluster-wide Consistency**

If *XDoc-Role-Margin(arg)* > $\delta_{12}$, then propagate the most frequent *etype* and *role* to all arguments with the same head as *arg* in the entire cluster.

**Rule (9): Adjust Argument Identification to Achieve Cluster-wide Consistency**

If *XDoc-Role-BestFreq(arg)* > $\delta_{13}$, then in the cluster, for each sentence containing an event mention *EM* with *etype*, add any unlabeled mention in that sentence with the same head as *arg* as an argument of *EM* with *role*.

23

## Experimental Results

| Performance / System/Human | Trigger Identification +Classification | | | Argument Identification | | | Argument Classification Accuracy | Argument Identification +Classification | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | | P | R | F |
| Within-Sentence IE with Rule (1) (Baseline) | 67.6 | 53.5 | 59.7 | 47.8 | 38.3 | 42.5 | 86.0 | 41.2 | 32.9 | 36.6 |
| Cross-sentence Inference | 64.3 | 59.4 | 61.8 | 54.6 | 38.5 | 45.1 | 90.2 | 49.2 | 34.7 | 40.7 |
| Cross-sentence+ Cross-doc Inference | **60.2** | **76.4** | **67.3** | **55.7** | **39.5** | **46.2** | **92.1** | **51.3** | **36.4** | **42.6** |
| Human Annotator1 | 59.2 | 59.4 | 59.3 | 60.0 | 69.4 | 64.4 | 85.8 | 51.6 | 59.5 | 55.3 |
| Human Annotator2 | 69.2 | 75.0 | 72.0 | 62.7 | 85.4 | 72.3 | 86.3 | 54.1 | 73.7 | 62.4 |
| Inter-Annotator Agreement | 41.9 | 38.8 | 40.3 | 55.2 | 46.7 | 50.6 | 91.7 | 50.6 | 42.9 | 46.4 |

The gold standard was adjudicated by two people (H1 & H2).

The last three rows are:
H1 vs. System , H2 vs. System, H1 vs. H2

24

## Cross-Event Inference for Event Extraction

[Liao & Grishman, 2010] observed that certain types of events frequently co-occur, so they incorporated **cross-event information** into their event extraction system.

For example:

*S1: He left the company.*

*S2: He planned to go shopping before heading home.*

*left* ➡ TRANSPORT EVENT

*S2: His colleagues threw a retirement party for him.*

*left* ➡ END-POSITION EVENT

25

## Cross-Event Correlations

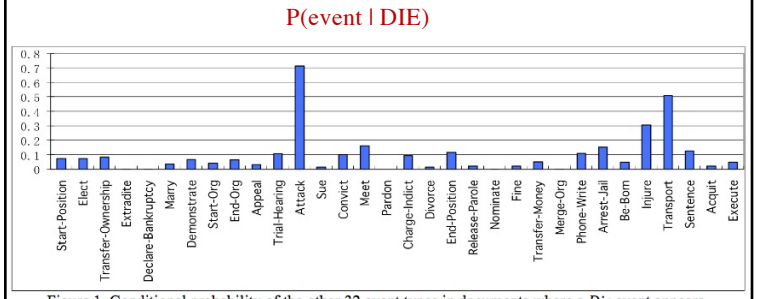



Figure 1. Conditional probability of the other 32 event types in documents where a *Die* event appears

It is common to find subevents that are highly associated with the main event, but those subevents can sometimes occur with many different types of main events.

26

## Events that Co-occur with DIE

| Event | Cond. Prob. |
|---|---|
| Attack | 0.714 |
| Transport | 0.507 |
| Injure | 0.306 |
| Meet | 0.164 |
| Arrest-Jail | 0.153 |
| Sentence | 0.126 |
| Phone-Write | 0.111 |
| End-Position | 0.116 |
| Trial-Hearing | 0.105 |
| Convict | 0.100 |

Table 3. Events co-occurring with *die* events with conditional probability > 10%

27

## Event Role Correlations

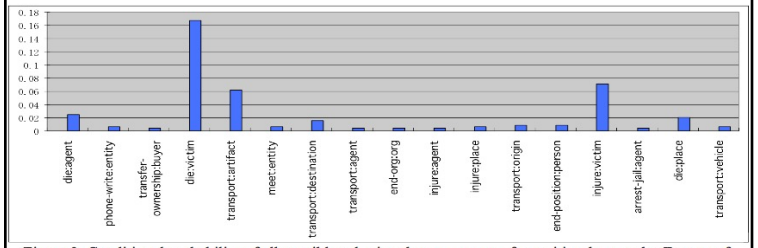



Figure 3. Conditional probability of all possible roles in other event types for entities that are the *Targets* of *Attack* events (roles with conditional probability below 0.002 are omitted)

There can be strong correlations between different types of events, and the role fillers across different types of events.

28

## Two-Pass Approach

- Liao & Grishman adopt a two-pass approach that first identifies the "easy cases" and then uses that knowledge to help identify the harder cases.

*The pro-reform director of Iran's biggest-selling daily newspaper and official organ of Tehran's municipality has **stepped** down following the **appointment** of a conservative … it was **founded** a decade ago … but a conservative city council was **elected** in the February 28 municipal polls … Mahmud Ahmadi-Nejad, reported to be a hardliner among conservatives, was **appointed** mayor on Saturday … **Founded** by former mayor Gholamhossein Karbaschi, Hamshahri …*

*British officials say they believe **Hassan** was a blindfolded woman seen being **shot** in the head by a hooded militant on a video obtained but not aired by the Arab television station Al-Jazeera. **She** would be the first foreign woman to **die** in the wave of kidnappings in Iraq … **she's** been **killed** by (men in pajamas) …*

29

## Confident Event Table

- The sentence-level event extraction system is applied to a document to identify high confidence predictions of event triggers and arguments.
- Event triggers and arguments (roles) that are labeled with high confidence are stored in a **Confident Event Table**.
- A word is assumed to be a trigger for only one type of event, and an entity is assumed to belong to just one role for an event trigger.

  If multiple labels are assigned to the same word/entity, then the highest scoring label is chosen if the difference between scores is large. If the difference between scores is small, then there is a conflict so the information is recorded in a separate **Conflict Table**.

30

**Confident table**

**Event type table**

| Trigger | Event Type |
|---|---|
| Met | Meet |
| Exploded | Attack |
| Went | Transport |
| Injured | Injure |
| Attacked | Attack |
| Died | Die |

**Argument role table**

| Entity ID | Event type | Role |
|---|---|---|
| 0004-T2 | Die | Time Within |
| 0004-6 | Die | Place |
| 0004-4 | Die | Victim |
| 0004-7 | Die | Agent |
| 0004-11 | Attack | Target |
| 0004-T3 | Attack | Time Within |
| 0004-12 | Attack | Place |
| 0004-10 | Attack | Attacker |

**Conflict table**

| Entity ID | Event type | Roles |
|---|---|---|
| 0004-8 | Attack | Victim, Agent |

Table 4. Example of document-level confident-event table (event type and argument role entries) and conflict table

31

## Document-Level Trigger Classifier

- A MaxEnt classifier is trained to predict whether a word is the trigger of an event, and if so, what type.
- The information in the confident event table is used to create features representing the other event types that have been found in the document.
- Each feature is the conjunction of:
  - the base form of the word
  - for each of the 33 event types, a binary value indicating whether this event type is present elsewhere in the document.

32

## Document-Level Argument (Role) Classifier

- A MaxEnt classifier is trained to predict whether a given mention is an argument of a given event, and if so, what role it plays.
- The information in the confident event table is used to create features for other event roles associated with this entity in the document.
- Each feature is the conjunction of:
  - the type of the given event
  - for each of the other 32 event types, the role of the given entity with respect to that event type (if one exists) or else *null*.

33

## Putting it All Together

- First, the sentence-level event extraction system is applied and high-confidence triggers and arguments are labeled.
- Next, the document-level trigger classifier is applied to all words that do not already have a label. This will often identify some additional event triggers.
- Finally, the document-level argument tagger is applied to all event triggers. Only entity mentions in the same sentence that have not already been assigned a role are considered.

  This tagger can identify arguments for the newly identified triggers as well as new arguments for the previously identified triggers.

34

## Experiments

- The 2005 ACE data set was used for evaluation: 549 training texts, 10 tuning texts, 40 test texts.
- Two baseline systems were evaluated:
  - the sentence-level event extraction component by itself
  - the [Ji & Grishman, 2008]'s approach for cross-sentence and cross-document ("within-event-type") inference rules.
- They also looked at the performance of two human annotators on 28 documents in the test set.

35

## Evaluation Results

| performance / system/human | Trigger classification | | | Argument classification | | | Role classification | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F |
| Sentence-level baseline system | 67.56 | 53.54 | 59.74 | 46.45 | 37.15 | 41.29 | 41.02 | 32.81 | 36.46 |
| Within-event-type rules | 63.03 | 59.90 | 61.43 | 48.59 | 46.16 | 47.35 | 43.33 | 41.16 | 42.21 |
| Cross-event statistical model | 68.71 | 68.87 | **68.79** | 50.85 | 49.72 | **50.28** | 45.06 | 44.05 | **44.55** |
| Human annotation1 | 59.2 | 59.4 | 59.3 | 60.0 | 69.4 | 64.4 | 51.6 | 59.5 | 55.3 |
| Human annotation2 | 69.2 | 75.0 | 72.0 | 62.7 | 85.4 | 72.3 | 54.1 | 73.7 | 62.4 |

Table 5. Overall performance on blind test data

36

## Conclusions

- Event extraction is a difficult problem – recall and precision are still only mediocre.
- But a variety of recent systems have shown that considering the entire document, as well as related documents, can be beneficial.
- These systems are still relatively shallow in their understanding of event descriptions. More explicit, richer event representations are probably needed to push performance to a higher level.

37

## Neural Models for Event Extraction

- A variety of neural models have been developed for sentence-level event extraction. As with many other tasks, NNs began with simple architectures and have gotten increasingly complex.
  - Some models improve F scores, but it's not always easy to understand why or whether they are truly better in general.
- A common theme in IE is to jointly tackle multiple tasks, for example performing entity recognition and event extraction at the same time. This idea is appealing because:
  - In a pipeline architecture, errors that occur early are propagated to later models and can't be undone.
  - Intuitively, it makes sense that different tasks can inform each other!

1

## A Joint Neural Model for Information Extraction with Global Features [Lin et al., ACL 2020]

- This paper describes a recent neural network architecture called **OneIE** that performs all of the subtasks required for sentence-level event extraction in a single framework.
- A key aspect of this model is that it creates an **information network** that captures information extracted across the entire sentence.
  - This network allows the model to consider **cross-task** and **cross-instance** interactions.
- The overall model is refreshingly straightforward: relatively simple individual task classifiers plus a decoding step at the end to find a globally optimal information network (graph) with learned features.
- This model is also language independent, so can be used for multiple natural languages.

2

## The IE Tasks

- **Entity Extraction**: identifying and labeling entity mentions, which can be names, nominals, or pronouns.

  Ex: finding all location phrases, such as *Kashmir region*
- **Relation Extraction**: assigning a relation type to an ordered pair of entity mentions.

  Ex: PART-WHOLE(*Kashmir region*, *India*)
- **Event Extraction**: identifying event triggers and their arguments

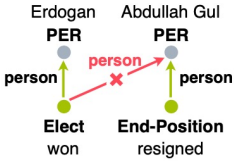
  Ex: "*assassination of a government official*" should identify *assassination* as a DIE event trigger and *government official* as the Victim of the DIE event.

3

## Motivating Example

*Prime Minister Abdullah Gul resigned earlier Tuesday to make way for Erdogan, who won a parliamentary seat in by-elections Sunday.*

The graph below shows system output with an extra person edge.




We'd like an IE system to recognize that ELECT events should not have 2 person arguments! Plus, Gul already has a role.

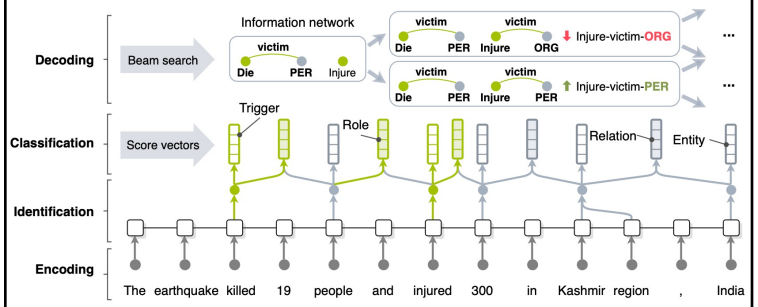
4

## The Four Stages of OneIE

1. Sentence Encoding: a contextualized embedding representation is produced for the input sentence.
2. Identification of entity mentions and event triggers.
   - these will become nodes in the information network
3. Local Classification of Nodes and Edges:
   - Labeling the nodes: entities with entity types, and the event triggers with event types.
   - Labeling edges with relations and event roles.
4. Finding the globally optimal information graph with a beam decoder.

5

## Joint Information Extraction Network (OneIE)




6

## Step 1: Sentence Encoding

- **Input**: a sentence of L words
- The sentence is passed to a pre-trained BERT encoder (transformer model), which produces a contextualized word embedding for each word in the sentence.
- BERT sometimes splits words into pieces, so they average the embedding vectors of each word's pieces.

  Ex: *Mondrian* → *Mon ##dr ##ian*

- These word embedding vectors are the input to the next stage of processing.
- They use a multilingual BERT model for the Chinese & Spanish data.

7

## Multilingual BERT

- BERT can be trained to handle multiple natural languages!
- The multilingual BERT model (on HuggingFace) is a language model that has been pre-trained with the 104 languages that have the most Wikipedia pages.
- The texts are lower-cased and tokenized. This produced a shared vocabulary size of 110,000.
- For languages like Chinese, Japanese Kanji and Korean Hanja that don't have space, a CJK Unicode block is added around every character.
- The languages with a larger Wikipedia are under-sampled and the ones with lower resources are oversampled.

8

## Step 2: Identifying Entities and Event Triggers

- **Two feed-forward neural networks (FFNs)** are trained: one for entity recognition and one for event trigger identification.
- Each FFN produces a score vector for every word over the label set.
  - They use BIO tagging for both tasks.
  - The possible labels for entities are the entity types.
  - The possible labels for event triggers are the event types.
- **A conditional random field (CRF) layer is added on top** of each FFN to handle dependencies when predicting the sequence of labels.
- The predicted types are not actually used though, to avoid propagating errors. The decoding stage at the end will ultimately decide on the entity and event trigger types.

9

## Step 3: Local Classification of Nodes & Edges

- For each identified node (entity or event trigger), if it contains multiple words then an embedding is produced for the node by averaging the vectors for each word.

1. Feed-forward neural networks are trained for each node-labeling task (entity types and event trigger types) to produce scores over each label set.
2. Feed-forward neural networks are trained for each edge-labeling task (relations and event roles).
   - An edge ($v_i$, $v_j$) is represented as the concatenation of the $v_i$ and $v_j$ span embeddings. Each edge has a start, end, and label type.

- For each task, the label with the highest scores is predicted to produce a locally best graph.

10

## Cross-Subtask Interactions

Interactions across different subtasks (recognizing entities, relations, events) can be valuable constraints! For example:

*A civilian aid worker from San Francisco was killed in an attack in Afghanistan.*


(a) Cross-subtask Interaction

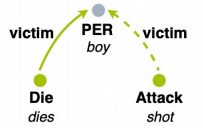
*San Francisco* may be predicted to be a **Victim** because it preceded *was killed*. But a **GPE** is unlikely to be a **Victim**.

11

## Cross-instance Interactions

Interactions across different events or relations in the same sentence can be valuable constraints.

*South Carolina **boy**, 9, dies during hunting trip after his father accidentally shot **him** on Thanksgiving Day.*




(b) Cross-instance Interactions

Recognizing that *boy* is the ATTACK Victim (trigger = *shot*) is tough because of the long distance between the words. But a DIE Victim is also likely to be the ATTACK Victim.

12

## Global Feature Templates

### Event Role Features

1. The number of entities that act as <role$_i$> and <role$_j$> arguments at the same time.
2. The number of <event_type$_i$> events with <number> <role$_j$> arguments.
3. The number of occurrences of <event_type$_i$>, <role$_j$>, and <entity_type$_k$> combination.
4. The number of events that have multiple <role$_i$> arguments.
5. The number of entities that act as a <role$_i$> argument of an <event_type$_j$> event and a <role$_k$> argument of an <event_type$_l$> event at the same time.

### Relation Features

6. The number of occurrences of <entity_type$_i$>, <entity_type$_j$>, and <relation_type$_k$> combination.
7. The number of occurrences of <entity_type$_i$> and <relation_type$_j$> combination.
8. The number of occurrences of a <relation_type$_i$> relation between a <role$_j$> argument and a <role$_k$> argument of the same event.
9. The number of entities that have a <relation_type$_i$> relation with multiple entities.
10. The number of entities involving in <relation_type$_i$> and <relation_type$_j$> relations simultaneously.

### Trigger Features

11. Whether a graph contains more than one <event_type$_i$> event.

13

## Using the Feature Templates

- Each feature template is instantiated with all possible values to generate a large feature set.
- Given a graph G, each function returns a scalar value. For example:

$$f_i(G) = \begin{cases} 1, & G \text{ has multiple ATTACK events} \\ 0, & \text{otherwise.} \end{cases}$$

- The neural network is tasked with learning a weight vector **u** for the features during training.
- The global score of G is the sum of its local score and global feature score: $s(G) = s'(G) + \boldsymbol{u}\boldsymbol{f}_G$

The local score for G is: $s'(\hat{G}) = \sum_{t \in T} \sum_{i=1}^{N^t} \max \hat{\boldsymbol{y}}_i^t$
T = tasks, $N^t$ = instances

14

## Step 4: Decoding the Globally Best Graph

- OneIE makes joint decisions for all nodes and edges to obtain a globally optimal graph.
- Ideally, we'd like to generate every possible candidate graph, calculate its global score and pick the best one. But exhaustive search is not feasible.
- SOLUTION: decoding is done via **beam search**.
  - Beam search is a greedy algorithm for heuristic search.
  - When exploring a graph, the **k** best partial solutions are expanded at each step of the search process. The value **k** is called the **beam width**.

15

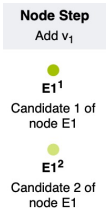
## The Decoding Process

- The beam is initialized with 1 candidate, an empty graph: B = { $K_0$ }
- At each step, each candidate in B is expanded with a node and edge.
- **Node Step**: a node $v_i$ is selected (presumably with the highest-scoring label). Given a hyper-parameter $\beta_v$ , the $\beta_v$ best labels are used. The beam is updated with a copy of each candidate graph that has $v_i$ added to it with one of the selected labels.
- **Edge Step**: edges are added between $v_i$ and previous nodes. Given a hyper-parameter $\beta_E$ , the edges ($v_j$, $v_i$) with the highest label scores are selected. The beam is updated with a copy of each candidate graph that has one of the selected edges added to it.
- If |B| > beam width k, then the global score for each candidate is computed and only the top k candidates are kept.

16

# Decoding Illustration

*He also brought a check from* ***Campbell$_{E1}$*** *to pay the fine and fees.*



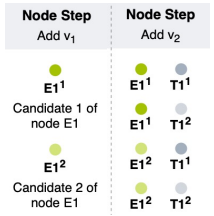

*Campbell* (E1) is selected as the first node to add, with 2 possible label candidates ($E1^1$ and $E1^2$).

No edges are added because there are no previous nodes to connect with E1.

17

# Decoding Illustration

*He also brought a check from* ***Campbell$_{E1}$*** *to pay the* ***fine$_{T1}$*** *and fees.*

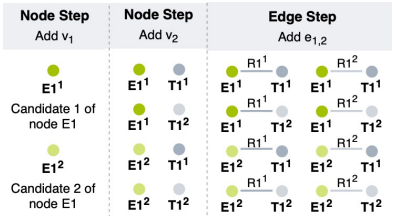



*fine* (T1) is selected as the second node to add, with 2 possible label candidates ($T1^1$ and $T1^2$)

18

# Decoding Illustration

*He also brought a check from* ***Campbell$_{E1}$*** *to pay the* ***fine$_{T1}$*** *and fees.*



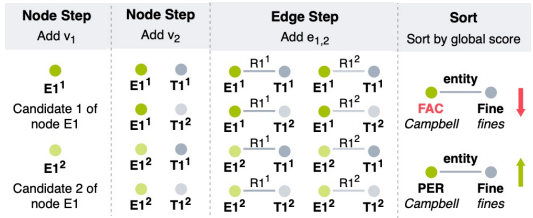

Edges are selected to link T1 with E1.

You can see here how the combinatorics can quickly create a large number of candidate graphs!

19

# Decoding Illustration

*He also brought a check from* ***Campbell$_{E1}$*** *to pay the* ***fine$_{T1}$*** *and fees.*

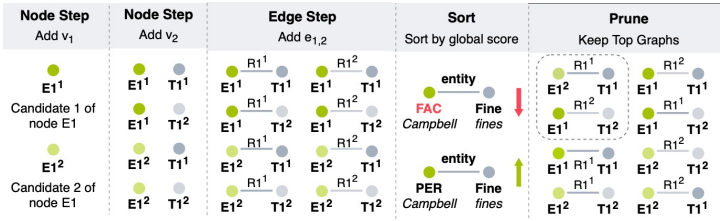



If the set of candidate graphs is greater than the beam width k, then the graphs are all scored and only the top k are kept in the beam.

20

# Decoding Illustration

*He also brought a check from* ***Campbell***$_{E1}$ *to pay the* ***fine***$_{T1}$ *and fees.*




And then the process repeats, until all of the nodes have been added.

21

# Data Sets

- ACE 2005: entities, relations, and events for English, Chinese, and Arabic.
  - 7 entity types, 6 (coarse) relation types, 33 event types, and 22 event roles.
  - ACE05-R only includes named entities and relations.
  - ACE05-E includes all entities, relations, and events.
  - ACE05-E+ includes ordered relation arguments, pronouns, and multi-token event triggers, which had been largely ignored in recent work!
- ERE-EN: Entities, Relations and Events (ERE) from a DEFT Program.
  - 7 entity types, 5 relation types, 38 event types, and 20 event roles.
  - 458 documents; 16,516 sentences.
  - They also created a Spanish version (ERE-ES)

22

# Data Set Statistics

| Dataset | Split | #Sents | #Entities | #Rels | #Events |
|---|---|---|---|---|---|
| ACE05-R | Train | 10,051 | 26,473 | 4,788 | - |
| | Dev | 2,424 | 6,362 | 1,131 | - |
| | Test | 2,050 | 5,476 | 1,151 | - |
| ACE05-E | Train | 17,172 | 29.006 | 4,664 | 4,202 |
| | Dev | 923 | 2,451 | 560 | 450 |
| | Test | 832 | 3,017 | 636 | 403 |
| ACE05-CN | Train | 6,841 | 29,657 | 7,934 | 2,926 |
| | Dev | 526 | 2,250 | 596 | 217 |
| | Test | 547 | 2,388 | 672 | 190 |
| ACE05-E$^+$ | Train | 19,240 | 47,525 | 7,152 | 4,419 |
| | Dev | 902 | 3,422 | 728 | 468 |
| | Test | 676 | 3,673 | 802 | 424 |
| ERE-EN | Train | 14,219 | 38,864 | 5,045 | 6,419 |
| | Dev | 1,162 | 3,320 | 424 | 552 |
| | Test | 1,129 | 3,291 | 477 | 559 |
| ERE-ES | Train | 7,067 | 11,839 | 1,698 | 3,272 |
| | Dev | 556 | 886 | 120 | 210 |
| | Test | 546 | 811 | 108 | 269 |

23

# Evaluations

- **Entity**: offsets and type must match the gold.
- **Relation**: offsets and type must match the gold.
- **Event Trigger**:
  - Trigger-I: offsets must match the gold (*Identification only*)
  - Trigger-C: offsets and event type must match the gold
- **Argument (Event Role)**:
  - Arg-I: offsets must match the gold (*Identification only*)
  - Arg-C: offsets and role type must match the gold

24

## Experimental Results

| Dataset | Task | DyGIE++ | Baseline | OneIE |
|---|---|---|---|---|
| ACE05-R | Entity | 88.6 | - | **88.8** |
| | Relation | 63.4 | - | **67.5** |
| ACE05-E | Entity | 89.7 | **90.2** | **90.2** |
| | Trig-I | - | 76.6 | **78.2** |
| | Trig-C | 69.7 | 73.5 | **74.7** |
| | Arg-I | 53.0 | 56.4 | **59.2** |
| | Arg-C | 48.8 | 53.9 | **56.8** |

F Scores

| Task | Entity | Trig-I | Trig-C | Arg-I | Arg-C | Relation |
|---|---|---|---|---|---|---|
| ACE05-E$^+$ | 89.6 | 75.6 | 72.8 | 57.3 | 54.8 | 58.6 |
| ERE-EN | 87.0 | 68.4 | 57.0 | 50.1 | 46.5 | 53.2 |

- DyGIE++ is a previous state-of-the-art model.
- Baseline is their system but without the globally optimal graph decoding layer.

25

## Results for Chinese and Spanish

| Task | Entity | Trig-I | Trig-C | Arg-I | Arg-C | Relation |
|---|---|---|---|---|---|---|
| ACE05-E$^+$ | 89.6 | 75.6 | 72.8 | 57.3 | 54.8 | 58.6 |
| ERE-EN | 87.0 | 68.4 | 57.0 | 50.1 | 46.5 | 53.2 |

| Dataset | Training | Entity | Relation | Trig-C | Arg-C |
|---|---|---|---|---|---|
| ACE05-CN | CN | 88.5 | 62.4 | 65.6 | 52.0 |
| | CN+EN | 89.8 | 62.9 | 67.7 | 53.2 |
| ERE-ES | ES | 81.3 | 48.1 | 56.8 | 40.3 |
| | ES+EN | 81.8 | 52.9 | 59.1 | 42.3 |

Note that adding English training data to the Chinese and Spanish models improved their performance!

26

## Positive Learned Global Features

| | Positive Feature | Weight |
|---|---|---|
| 1 | A TRANSPORT event has only one DESTINATION argument | 2.61 |
| 2 | An ATTACK event has only one PLACE argument | 2.31 |
| 3 | A TRANSPORT event has only one ORIGIN argument | 2.01 |
| 4 | An END-POSITION event has only one PERSON argument | 1.51 |
| 5 | A PER-SOC relation exists between two PER entities | 1.08 |
| 6 | A GEN-AFF relation exists between ORG and LOC entities | 0.96 |
| 7 | A BENEFICIARY argument is a PER entity | 0.93 |
| 8 | A GEN-AFF relation exists between ORG and GPE entities | 0.90 |

27

## Negative Learned Global Features

| | Negative Feature | Weight |
|---|---|---|
| 9 | An entity has an ORG-AFF relation with multiple entities | -3.21 |
| 10 | An entity has an PART-WHOLE relation with multiple entities | -2.49 |
| 11 | An event has two PLACE arguments | -2.47 |
| 12 | A TRANSPORT event has multiple DESTINATION arguments | -2.25 |
| 13 | An entity has a GEN-AFF relation with multiple entities | -2.02 |
| 14 | An ATTACK event has multiple PLACE arguments | -1.86 |
| 15 | An entity has a PHYS relation with multiple entities | -1.69 |
| 16 | An event has multiple VICTIM arguments | -1.61 |

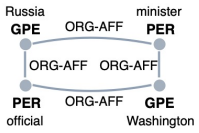
28

## Feature Analysis #1

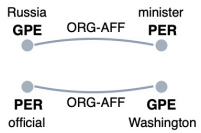
#1: **Russia**'s foreign **minister** expressed outrage at suggestions from a top **Washington official** last week that Moscow should forgive the eight billion dollars in Soviet-era debt that Baghdad owes it, as a gesture of good will.

⋆ Global feature category: 8
⋆ Analysis: It is unlikely for a person to have an ORG-AFF relation with multiple entities.

Baseline




+ Global Features



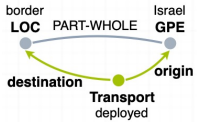

29

## Feature Analysis #2

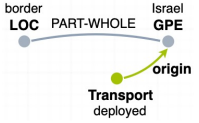
#2: They also **deployed** along the **border** with **Israel**.

⋆ Global feature category: 9
⋆ Analysis: It is uncommon that an ORIGIN argument and a DESTINATION argument have a PART-WHOLE relation.

Baseline




+ Global Features



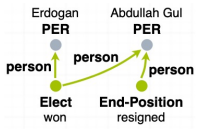

30

## Feature Analysis #3

#3: Prime Minister **Abdullah Gul resigned** earlier Tuesday to make way for **Erdogan** , who **won** a parliamentary seat in by-elections Sunday.

⋆ Global feature categories: 2 and 5
⋆ Analysis: 1. An ELECT usually has only one PERSON argument; 2. An entity is unlikely to act as a PERSON argument for END-POSITION and ELECT events at the same time.

Baseline




+ Global Features

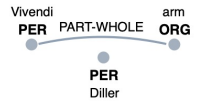

31

## Feature Analysis #4

#4: Diller will continue to play a critical role in the future of **Vivendi** 's entertainment **arm**.

⋆ Global feature category: 6
⋆ Analysis: A PART-WHOLE relation should not exist between PER and ORG entities.

Baseline




+ Global Features


32

## Feature Analysis #5

#5: He also brought a check from **Campbell** to **pay** the fines and fees.

⋆ Global feature category: 3
⋆ Analysis: As "Campbell" is likely to be an ENTITY argument of a FINE event, the model corrects its entity type from FAC to PER.

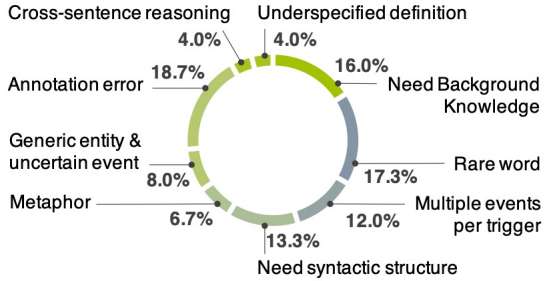
Baseline


+ Global Features


33

## Error Analysis

They manually analyzed 75 errors to better understand where the remaining challenges lie:




34

## Error Category Examples

- **Background Knowledge**: "*And Putin's media aide, Sergei Yastrzhembsky, told Kommersant Russia would not forgive the Iraqi debt*". *Kommersant* was labeled as a PERSON, but is an ORG. (It is a newspaper.)
- **Rare Words**: the long-tail problem is a well-known challenge for NLP. For example, "*caretaker*" is a PERSON. Event triggers can be multi-word expressions that are rare, and even adverbs occasionally.
- **Multiple Types per Trigger**: "*named*" can refer to both NOMINATE and START-POSITION events, and "*killed*" can refer to both ATTACK and DIE events (but usually only one is annotated).

35

## Error Category Examples

- **Complex Syntactic Structure**: "*As well as previously holding senior positions at Barclays Bank, BZW and Kleinwort Benson, McCarthy was formerly a top civil servant at the Department of Trade and Industry*." Their model missed all 3 employers: *Barclays Bank, BZW,* and *Kleinwort Benson*.
- **Uncertain Events**: future planned events can be mistaken for events that have already happened, for example: "*The statement did not give any reason for the move, but said Lahoud would begin consultations Wednesday aimed at the formation of a new government.*"
- **Metaphor**: "*Russia hints 'peace camp' alliance with Germany and France is dying… "*

36

## Summary

- The OneIE model is a framework that simultaneously learns multiple IE subtasks and incorporates global features to recognize subtask and instance interactions that are desirable or undesirable.
  - A nice combination of feature engineering with neural nets.
  - These types of features are clearly beneficial!
- However, this work is still only tackling *sentence-level event extraction*! And the results are still far from perfect.
- The challenges grow even greater when extracting event information from an entire document, where the information can be scattered about and a coherent event representation must be produced.

37

# Document-Level Event Extraction

Identifying descriptions of complex events and extracting the event role fillers associated with each incident.

| EVENT | ROLES |
|---|---|
| Terrorist act | perpetrator, victims, target |
| Natural disaster | natural force, victims, damage |
| Plane crash | vehicle, victims, cause |
| Management changes | person leaving, position, successor, organization |
| Disease outbreaks | disease, victims, symptoms, containment measures |

1

# Document-Level Event Extraction

**INPUT**: document

**OUTPUT**: one filled template for each distinct event

December 29, Pakistan - The U.S. embassy in Islamabad was damaged this morning by a car bomb. Three diplomats were injured in the explosion. Al Qaeda has claimed responsibility for the attack.

EVENT: *bombing*
TARGET: *U.S. embassy*
LOCATION: *Islamabad*
DATE: *December 29*
WEAPON: *car bomb*
VICTIMS: *three diplomats*
PERPETRATOR: *Al Qaeda*

2

# Event Template for Terrorist Acts

| | |
|---|---|
| Date | <date> |
| Location | <location> |
| Event type | <set fill> |
| Weapon | <string list> |
| Perpetrator individual | <string list> |
| Perpetrator organization | <string list> |
| Physical target | <string list> |
| Physical target effect | <set fill> |
| Human target | <string list> |
| Human target effect | <set fill> |

3

# Filled Event Template for Terrorist Acts

| | |
|---|---|
| Date | 10 January 1990 |
| Location | El Salvador: San Salvador (city) |
| Event type | BOMBING |
| Weapon | *"highpower bombs"* |
| Perpetrator individual | *"guerrilla urban commandos"* |
| Perpetrator organization | - |
| Physical target | *"car dealership"* |
| Physical target effect | some damage |
| Human target | - |
| Human target effect | no injury or death |

4

## Event Template for Disease Outbreaks

| | |
|---|---|
| Story: | <document id> |
| ID: | <template id> |
| Date: | <date> |
| Event: | OUTBREAK |
| Status: | <set fill> |
| Containment: | <set fill> |
| Country: | <set fill> |
| Victims: | <string list> |
| Disease: | <string> |

5

## Filled Event Template for Disease Outbreaks

| | |
|---|---|
| Story: | 20020714.4756 |
| ID: | 1 |
| Date: | August 14, 2002 |
| Event: | OUTBREAK |
| Status: | confirmed |
| Containment: | none |
| Country: | Switzerland |
| Victims: | *the 27 reported cases* |
| Disease: | *Creutzfeldt-Jakob Disease / [sporadic] Creutzfeldt-Jakob disease (CJD) / CJD / Sporadic CJD / hereditary dominant CJD / Swiss CJD / sporadic Creutzfeldt-Jakob disease* NOTE: disjunctive options! |

6

## Patterns/Rules vs. Sequence Tagging

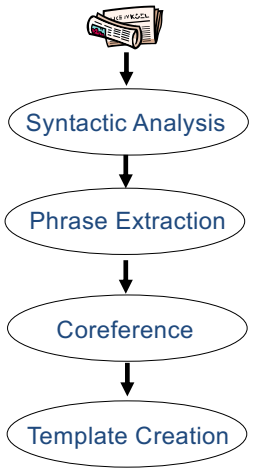
Two general approaches to event extraction:

**Pattern-based systems** use patterns or rules which identify phrases that should be extracted for each event role.

**Machine learning classifiers** label individual tokens indicating whether they should be extracted, and if so, what role they play.

7

## Template-Filling Pipeline




IBM fired its CEO.

IBM fired its CEO.
*Subj* *VP* *Dobj*

*Event:* FIRING
*Firer:* IBM *Employee:* its CEO

IBM fired its CEO.
John Smith was let go on Monday.

| | |
|---|---|
| Event: | FIRING |
| Firer: | IBM |
| Employee: | John Smith, CEO |
| Date: | Monday |

8

## Event Narrative Example

A bomb exploded today in a Lima restaurant, and a second device that had been placed in the same establishment was deactivated by the Peruvian national police.

There were no victims, and the explosion caused very little damage to the restaurant, which is located in the commercial area of the residential district of Miraflores.

Guerrillas of the Tupac Amaru Revolutionary Movement (MRTA) have claimed credit for the terrorist act through pamphlets they left on the premises, according to the police.

9

## Secondary Contexts in a Document

The victims were identified as David Lecky and James Donnelly.

Oqueli's body was found next to the body of Gilda Flores.

According to witnesses' reports, two 23 to 25-year-old individuals walked to Gen. Leigh's office, on the fourth floor of the building.

The destroyed UCR headquarters is in the Moreno district of Buenos Aires.

Cardenas Guerra is apparently linked to the Medellin drug cartel.

There were seven children, including four of the Vice President's children, in the home at the time.

10

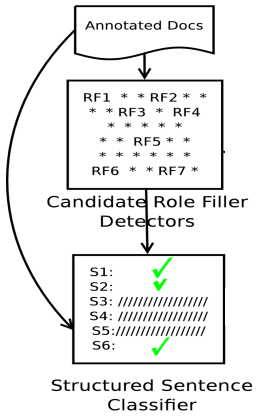
## Discourse-Guided Event Extraction

[Huang & Riloff, AAAI 2012]

- Goal: use a *relevant sentence classifier* to identify event contexts across sentences.
- The classifier can consider properties of both the current sentence and previous sentence.
- Defines several types of discourse features to capture textual cohesion across sentences.

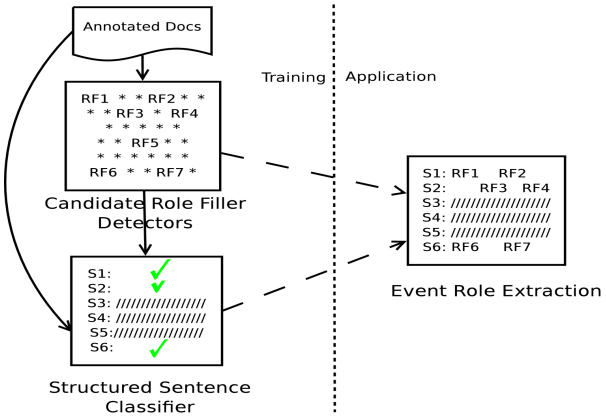
11

## Linker [Huang & Riloff,2012]




12

## Linker [Huang & Riloff,2012]




13

## Role Filler Extractor

- Train machine learning classifiers (SVMs) to label noun phrases based on a small context window around the NP. One SVM per event role.
- Each classifier contains 3 types of features:
  - lexical features (context window, head, premodifiers)
  - NER and semantic class labels
  - lexico-syntactic patterns

14

## Example of Relevant Sentence Labels

R A government building blew up in Bogota this morning.

R It was heavily damaged and three people were killed.

R Three pipe bombs were found inside the garage.

I Bogota has been ravaged by violence in recent years.

I The government has passed new laws to try and crack down on suspected terrorist organizations.

I Residents fear that kidnappings of civilians and assassinations of politicians may be on the horizon.

R The FMLN has been threatening violence and has claimed responsibility for this morning's attack.

R The three victims of the attack have been identified as Jose Flores, Jorge Garcia, and Carlos Lopez.

15

## Example of Relevant Sentence Labels

R A government building blew up in Bogota this morning.

R It was heavily damaged and three people were killed.

R Three pipe bombs were found inside the garage.

I Bogota has been ravaged by violence in recent years.

I The government has passed new laws to try and crack down on suspected terrorist organizations.

I Residents fear that kidnappings of civilians and assassinations of politicians may be on the horizon.

R The FMLN has been threatening violence and has claimed responsibility for this morning's attack.

R The three victims of the attack have been identified as Jose Flores, Jorge Garcia, and Carlos Lopez.

16

## Structured Sentence Classifier

- A sequential tagging model (CRF) is trained to label each sentence as to whether it is a relevant event context .It used unigram and bigram features as well as:
- Four types of discourse features
  - Lexical bridge features
  - Discourse bridge features
  - Discourse focus features
  - Role filler distribution features

17

## Lexical Bridge Features

Lexical Bridge features capture lexical associations between adjacent sentences.

Two types: <$verb_{i-1}$,$verb_i$>

<$noun_{i-1}$,$noun_i$>

Examples:

<explode, injure>

<bomb, building>

18

## Penn Discourse Treebank

- The ***Penn Discourse Treebank (PDTB)*** contains texts that have been manually annotated with discourse relations.
- The annotations represent the argument structure, senses and attribution of discourse connectives and their arguments.
- ***Explicit discourse connectives*** require the presence of a discourse cue phrase, such as*: if, because, so, since, but, however, as a result*
- ***Implicit discourse connectives*** indicate that most readers would infer a discourse relation between adjacent sentences.
- Discourse parsers have been developed by training with the PDTB data.

19

## Examples of Implicit Connectives

(68) Several leveraged funds don't want to cut the amount they borrow because it would slash the income they pay shareholders, fund officials said. But a few funds have taken other defensive steps. *Some have raised their cash positions to record levels.* Implicit = BECAUSE **High cash positions help buffer a fund when the market falls.** (0983)

(69) *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500.* Implicit = SO **By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs.** (0994)

20

## Discourse Bridge Features

- Discourse relations between adjacent sentences, based on the PDTB discourse parser.
- Labels explicit discourse relations based on cue phrases (e.g., *if* and *because*)
- Labels implicit discourse relations such as *cause*, *condition*, *instantiation*, and *contrast*.
- Linker captures relations both within a sentence and between the current and previous sentence.

21

## Discourse Focus Feature

- Hypothesis: two sentences are probably related if they have the same discourse focus.
- Create a feature for shared NPs (same head) in adjacent sentences that occur as a Subject, DObj, or PP(by). Feature = <NP_head, Synrole$_i$, Synrole$_{i-1}$>

(1) *A customer in the store was shot by **masked men**.*

(2) ***The two men** used 9mm semi-automatic pistols.*

⟹ <*men*, Subject, PP(by) >

22

## Role Filler Distribution Features

Purpose: capture information about the presence and types of possible role fillers in the neighborhood.

Features within a sentence:

- type and head noun of each candidate role filler
- density of candidate role fillers
- pairs of different types

Features across adjacent sentences:

- head and type of pairs across sentences
- candidates that share a discourse relation
- verb and candidate pairs across sentences

23

## Evaluation Results

| Method | Recall | Precision | F |
|---|---|---|---|
| *Local Extractor* | | | |
| Candidate RF Detectors | **75** | 30 | 42 |

24

## Evaluation Results

| Method | Recall | Precision | F |
|---|---|---|---|
| *Local Extractor* | | | |
| Candidate RF Detectors | **75** | 30 | 42 |
| with Structured Sentence Classifier | | | |
| Basic N-gram Features | 56 | 55 | 56 |
| +Extra Features | 60 | **58** | **59** |

25

## Individual Event Role Result

| System | PerpInd | PerpOrg | Target | Victim | Weapon | Average |
|---|---|---|---|---|---|---|
| Local Extraction Only | | | | | | |
| **Candidate RF Detectors** | 25/67/36 | 26/78/39 | 34/83/49 | 32/72/45 | 30/75/43 | 30/75/42 |
| with Structured Sentence Classifier | | | | | | |
| **Basic feature set** | 56/54/55 | 47/46/46 | 55/69/**61** | 61/57/59 | 58/53/56 | 55/56/56 |
| **+ Candidate RF features** | 51/57/54 | 47/47/47 | 54/69/60 | 60/58/59 | 56/60/58 | 54/59/56 |
| **+ Lexical Bridge features** | 51/57/53 | 51/50/50 | 55/69/**61** | 60/58/59 | 62/62/62 | 56/59/57 |
| **+ Discourse features** | 54/57/**56** | 55/49/**51** | 55/68/**61** | 63/59/**61** | 62/64/**63** | 58/60/**59** |

Scores are shown as Precision/Recall/F

26

## Neural Architecture for Document-level Event Extraction

- [Du & Cardie, ACL 2020] developed a neural network architecture for document-level event extraction.
- They created a **multi-granularity reader**, which explicitly captures neural representations learned from different levels of granularity (i.e., both sentences and paragraphs).
- This model was designed based on observations from earlier research that relevant information sometimes occurs near event triggers, but can also occur elsewhere in a document! It is important for models to capture both:
  1. local contexts around relevant event trigger phrases
  2. larger discourse contexts that discuss a relevant event

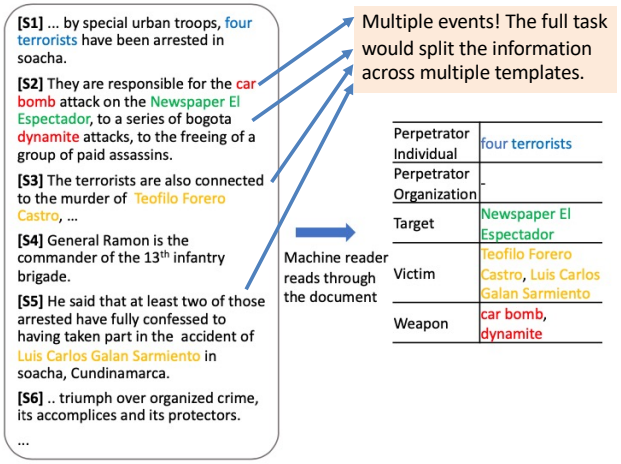
27

## Template-based Task Approximation

- As with most prior work (including Linker), Du & Cardie tackle only one piece of the document-level event extraction task: **document-level role filler extraction**.
- For complete template-based event extraction, the system must produce a set of templates for each document: one template for each distinct event.
  - If a story reports 3 bombings, 3 templates must be produced.
- Producing templates requires event tracking: determining how many events occurred and associating the extracted facts with the correct template.
- Most systems have focused on extracting all of the correct event role fillers, without mapping them into distinct templates.

28

## Example of the Task




29

## Sequence Tagging Model

- The problem was modeled as sequence labeling using BIO tags.

| ... | four | terrorists | have | been | arrested | in | soacha | ... |
|---|---|---|---|---|---|---|---|---|
| ... | B-Perpind | I-Perpind | O | O | O | O | O | ... |

| ... | are | responsible | for | the | car | bomb | attack | on | the | newspaper |
|---|---|---|---|---|---|---|---|---|---|---|
| ... | O | O | O | O | B-Weapon | I-Weapon | O | O | O | B-Target |
| el | espectador | , | to | a | series | of | bogota | dynamite | attacks | ... |
| I-Taget | I-Target | O | O | O | O | O | O | B-Weapon | O | ... |

- With gold standard templates, the answer strings are not associated with specific positions in the document. For example:

  **BOMBING**: *perpetrator* = "four terrorists", *weapon* = "car bomb"

  Solutions: All instances could be labeled, or heuristics applied (e.g., just the first instance, or the instance closest to a keyword).

30

## First Approach: k-Sentence Reader

- They create **k-sentence contexts** of different lengths for training. Starting with each sentence **i**, they concatenate **k** consecutive sentences to form overlapping sequences of length k. For example:

  $S_1\ S_2\ \ldots\ S_k$ $S_2\ S_3\ \ldots\ S_{k+1}$ $S_3\ S_4\ \ldots\ S_{k+2}$ etc.

- A sequence is **positive** if it contains ≥ 1 role filler, or **negative** otherwise. Training uses a 50/50 positive/negative sample.

- When applying the model during testing, each non-overlapping sequence of k sentences is used. For example:

  $S_1\ S_2\ \ldots\ S_k$ $S_{k+1}\ S_{k+2}\ \ldots\ S_{2k}$ etc.

- For **paragraph contexts**, k is the average paragraph length during training and the real paragraph length during testing.

31

## Embedding Layer

- A word $x_i$ in the input sequence is represented as the concatenation of its non-contextual word embedding vector and its contextual word vector.

- **Word Embedding Vector**: 100-dimensional GloVe pre-trained word embedding trained on Web crawl data.

$$\mathbf{xe}_i = \mathbf{E}(x_i)$$

- **Contextual Embedding Vector**: the contextualized embedding for $x_i$ produced by the BERT pre-trained language model given the input sequence $\{x_1, x_2, ..., x_m\}$

$$\mathbf{xb}_1, \mathbf{xb}_2, ..., \mathbf{xb}_m = \texttt{BERT}(x_1, x_2, ..., x_m)$$

$$\mathbf{x}_i = \texttt{concat}(\mathbf{xe}_i, \mathbf{xb}_i)$$

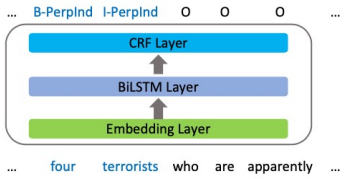
32

# Basic Architecture

- A bi-directional LSTM layer sits on top of the token representations to further learn task-specific word representations.

$$\{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_m\} = \texttt{BiLSTM}(\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\})$$

- A CRF layer sits on top of the biLSTM layer to perform sequence tagging of the BIO labels.




33

# Multi-Granularity Reader

- **Hypothesis**: recognizing event information requires richer contextual representations based on paragraph regions.
- They train both **sentence-based and paragraph-based biLSTM models** to produce improved contextual representations of the words in a paragraph.
- Given a paragraph, they apply the sentence-based biLSTM to each sentence and the paragraph-based biLSTM to the entire paragraph.

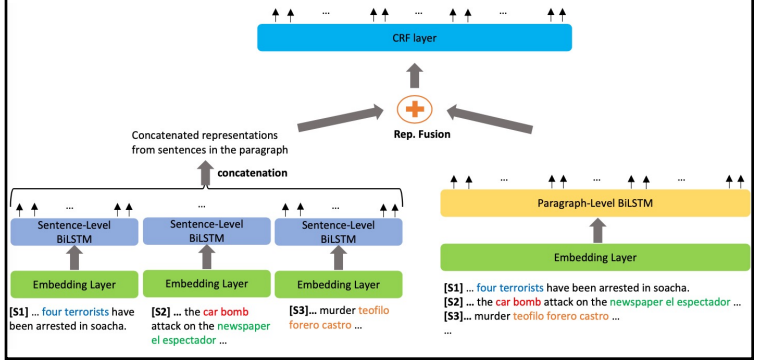
$$\{\tilde{\mathbf{p}}_1^{(1)}, \tilde{\mathbf{p}}_2^{(1)}, ..., \tilde{\mathbf{p}}_{l_1}^{(1)}\} = \texttt{BiLSTM}_{sent.}(\{\tilde{\mathbf{x}}_1^{(1)}, \tilde{\mathbf{x}}_2^{(1)}, ..., \tilde{\mathbf{x}}_{l_1}^{(1)}\})$$
$$...$$
$$\{\tilde{\mathbf{p}}_1^{(k)}, \tilde{\mathbf{p}}_2^{(k)}, ..., \tilde{\mathbf{p}}_{l_k}^{(k)}\} = \texttt{BiLSTM}_{sent.}(\{\tilde{\mathbf{x}}_1^{(k)}, \tilde{\mathbf{x}}_2^{(k)}, ..., \tilde{\mathbf{x}}_{l_k}^{(k)}\})$$

$$\{\hat{\mathbf{p}}_1^{(1)}, ..., \hat{\mathbf{p}}_{l_1}^{(1)}, ..., \hat{\mathbf{p}}_1^{(k)}, ..., \hat{\mathbf{p}}_{l_k}^{(k)}\} = \texttt{BiLSTM}_{para.}(\{\hat{\mathbf{x}}_1^{(1)}, ..., \hat{\mathbf{x}}_{l_1}^{(1)}, .., \hat{\mathbf{x}}_{l_k}^{(k)}, ..., \hat{\mathbf{x}}_{l_k}^{(k)}\})$$

34

# Fusing Representations

Finally, their model combines both sentence-level and paragraph-level representations, providing both narrow and broad views of the discourse.




35

# Fusion Operations

- They experimented with two ways of fusing the sentence-level and paragraph-level representations for a token.
- **Simple Sum**: $\mathbf{p}_i^{(j)} = \tilde{\mathbf{p}}_i^{(j)} + \hat{\mathbf{p}}_i^{(j)}$
- **Gated Fusion** (used in final model): computes a gate vector $g_i^{(j)}$ with its sentence-level token representation p˜(j) and paragraph-level token representation pˆ(j), to control how much information should be incorporated from the two representations.

$$\mathbf{g}_i^{(j)} = sigmoid(\mathbf{W}_1\tilde{\mathbf{p}}_i^{(j)} + \mathbf{W}_2\hat{\mathbf{p}}_i^{(j)} + b)$$
$$\mathbf{p}_i^{(j)} = \mathbf{g}_i^{(j)} \odot \tilde{\mathbf{p}}_i^{(j)} + (1 - \mathbf{g}_i^{(j)}) \odot \hat{\mathbf{p}}_i^{(j)}$$

$\odot$ : element-wise product

36

# Data Set & Evaluation

- MUC-4 Event Extraction Data Set (Latin American Terrorism)
  - 1700 documents with gold event templates
  - 1300 for training, 200 for development, 200 for testing
- Evaluation Metrics
  - **Head Noun Match**: two strings match if their head nouns match. Example: *men* matches *armed men* matches *several armed men.*
  - **Exact Match**: two strings must be exactly the same.
  - Duplicate extractions are conflated before evaluation, so extracting the same string multiple times counts as a single hit or miss.

37

# Evaluating Contextual Effects

- They also investigate the impact of looking at different granularities of discourse regions using their k-Sentence Reader model.
  - Single Sentence Reader (k=1)
  - Double Sentence Reader (k=2)
  - Paragraph Reader (follows paragraph breaks in text)
  - Chunk Reader (maximum # of sentences that BERT allows)
- They also investigate how much impact the contextualized word representations have.

38

# Experimental Results

| | Head Noun Match | | | Exact Match | | |
|---|---|---|---|---|---|---|
| | Prec. | Recall | F-1 | Prec. | Recall | F-1 |
| GLACIER (Patwardhan and Riloff, 2009) | 47.80 | 57.20 | 52.08 | - | - | - |
| TIER (Huang and Riloff, 2011) | 50.80 | 61.40 | 55.60 | - | - | - |
| LINKER → Cohesion Extract (Huang and Riloff, 2012) | 57.80 | 59.40 | 58.59 | - | - | - |
| *w/o contextualized embedding* | | | | | | |
| Single-Sentence Reader | 48.69 | 56.11 | 52.14 | 46.16 | 53.16 | 49.41 |
| Double-sentence Reader | 56.37 | 47.53 | 51.57 | 53.70 | 43.95 | 48.34 |
| Paragraph Reader | 53.19 | 53.16 | 53.17 | 49.45 | 49.26 | 49.35 |
| Chunk Reader | **61.76** | 37.04 | 46.31 | **56.91** | 34.92 | 43.28 |
| *w/ contextualized embedding* | | | | | | |
| Contextualized Single-Sentence Reader | 47.32 | 61.26 | 53.39 | 44.40 | **57.67** | 50.17 |
| Contextualized Double-sentence Reader | 57.17 | 53.36 | 55.20 | 53.38 | 49.22 | 51.22 |
| Contextualized Paragraph Reader | 56.78 | 52.64 | 54.64 | 53.36 | 49.65 | 51.44 |
| Contextualized Chunk Reader | 60.90 | 41.10 | 49.07 | 55.18 | 37.51 | 44.66 |
| Multi-Granularity Reader | 56.44 | **62.77** | **59.44** | 52.03 | 56.81 | **54.32** |

Note that increasing the context size too much can hurt!

39

# Breakdown by Event Role

| | PerpInd | | | PerpOrg | | | Target | | | Victim | | | Weapon | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F-1 | P | R | F-1 | P | R | F-1 | P | R | F-1 | P | R | F-1 |
| LINKER (2012) | 54 | 57 | **56** | 55 | 49 | 51 | 55 | 68 | 61 | 63 | 59 | **61** | 62 | 64 | 63 |
| Multi-Granularity Reader | 53.08 | 52.23 | 52.65 | 50.99 | 67.88 | **58.23** | 60.38 | 64.10 | **62.18** | 49.34 | 62.05 | 54.97 | 68.42 | 67.57 | **67.99** |

Note that some event roles benefit from this model much more than others!

PerpOrg and Weapon are frequently mentioned in text regions that are distant from where the event is first described.

40

# Ablation Study

They removed several components of the model (one at a time) to measure their impact on performance.

| | Head Noun Match | | | Exact Match | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F-1 | Precision | Recall | F-1 |
| Multi-granularity Reader | 56.44 | 62.77 | **59.44** | 52.03 | 56.81 | **54.32** |
| w/o gated fusion | 48.09 | 67.32 | 56.10 | 43.75 | 62.37 | 51.43 |
| w/o BERT | 59.16 | 50.80 | 54.66 | 55.48 | 46.99 | 50.88 |
| w/o CRF layer | 50.52 | 56.95 | 53.54 | 47.02 | 53.55 | 50.07 |

**Conclusion**: all 3 of these components contribute substantially to the model's performance.

41

# Qualitative Analysis

They examined some examples to better understand the strengths and weaknesses of their model.

... the announcer says president virgilio barco will tonight disclose his government's peace proposal. ..... . Near the end, the announcer adds to the initial report on the el tomate attack with a 3-minute update that adds 2 injured, 21 houses Target destroyed, and 1 bus Target burned.

Single-sentence and MG readers find the targets. Paragraph reader does not.

.... An attack came at approximately 22:30 last night. Members of the civil group and the peruvian investigative police went to the site of the explosion. The members of the republican guard antiexplosives brigade are investigating to determine the magnitude of the bomb Weapon used in this attack.

Single-sentence reader incorrectly extracts the civil group. Paragraph and MG readers do not extract it.

42

# Missed Extractions by the MG Reader

... Patriotic officer, it is time we sit down to talk, to see what we can do with our fatherland, and what are we going to do with La Tandona PerpOrg. .... To continue defending what, we ask you. ... .

La Tandona was not extracted. Humans would probably rely on world knowledge of the organization.

... said that the guerrillas are desperate and ... . The president expressed his satisfaction at the release of Santander department senator Jorge Serrano Gonzalez Target, whom he described as one of the most important people that colombian democracy has at this moment.

Gonzalez was not recognized as a kidnapping victim. Deeper understanding of the story as a whole is needed.

43

# Conclusions

- Document-level Event Extraction is one of the more difficult information extraction tasks. Performance is still not great, and we're only tackling a simplified version of the problem thus far.
- We have learned that representing discourse contexts is important for this task, and future work will probably need to focus even more on understanding how sentences relate to each other.
- There are many challenging related subproblems as well, including:
  - Coreference resolution
  - Subevent identification
  - Inference (e.g., *body of X* => *X is dead*)
  - Metaphor recognition

44

## Rehash: Exam, Question 4

**Text Corpus**

S1: George Washington was the first president of the United States .
S2: From 1789 to 1797 George Washington served as the first president of the United States .
S3: In 1789 George Washington was inaugurated as the first president of the United States and led the country .
S4: George Washington governed the United States and led the Revolutionary War .
S5: George Washington was born in the United States in 1732 .

S6: The first president of India was Rajendra Prasad .
S7: The initial president of independent India was Rajendra Prasad .
S8: Rajendra Prasad was a scholar in India and a teacher .
S9: The first constitution for India was drafted by Rajendra Prasad in 1946 .

(a) How many individual instances of <George Washington, United States>? 5
(b) How many collective instances of <George Washington, United States>? 1
(c) How many individual instances of <Rajendra Prasad, India>? 4
(d) How many collective instances of <Rajendra Prasad, India>? 1

1

(a) Show the 5-tuple pattern representation that the Snowball system would generate for each of the individual positive training instances created from <Rajendra Prasad, India>.

*First example:*

S6:
Left = {president (1), of (1)}
Entity1 = COUNTRY
Middle = {was (1)}
Entity2 = PERSON
Right = {}

(b) Show the 5-tuple pattern representation that the Snowball system would generate for each of the collective positive training instances created from <George Washington, United States>.

Left = {to (1), 1797 (1), in (1), 1789 (1)}
Entity1 = PERSON
Middle = {was (3), the (8), first (3), president (3), of (3), served (1), as (2), inaugurated (1), governed (1), born (1), in (1)}
Entity2 = COUNTRY
Right = {and (2), led (2), in (1), 1732 (1)}

2

## Document-level Event Extraction: Real Text Example

Official sources today reported that at least eight people, including soldiers, rebels, and civilians, were killed during clashes between the army and guerrillas over the past weekend in various points of the country.

Military spokesmen for the 6th infantry brigade, headquartered in the eastern usulutan department, told acanefe that two rebels were killed and one wounded during a clash with government troops in San Agustin.

Meanwhile, the armed forces press committee (Coprefa) reported that the bodies of two guerrillas, who were presumably killed during clashes with the army, were found by soldiers in the outskirts of Santa Tecla, in the central la libertad department.

Coprefa reported that two soldiers were killed during a clash with members of the **Farabundo Marti National Liberation Front** (FMLN) in Comasagua, about 28 km to the southwest of (San) Salvador, where a rebel attack on a coffee processing plant was successfully repelled.

It reported that a civilian was killed in the crossfire and that a soldier was also killed during clashes in Zaragoza, south of San Salvador, where two guerrillas were wounded.

• • •

Salvadoran (red) cross sources today reported that a 48-year-old woman identified as Maria Luz Lopez was wounded last night when a powerful bomb, which damaged several businesses in (San) Salvador, exploded.

The bomb was planted in a heavily commercial area of downtown (San) Salvador causing heavy property loses, according to the owners who provided no specific figures.

This is the fourth dynamite attack on businesses in (San) Salvador so far in 1990.

3

## Generative Role-filler Transformers for Document-level Event Entity Extraction [Du et al., EACL 2021]

- The GRIT system tackles **role-filler entity extraction** (**REE**). which aims to extract role fillers for relevant events irrespective of template generation. (So the gold set is essentially the union of role fillers across all gold templates.)
- In contrast to sentence-level event extraction, the system should extract just ONE reference to each relevant entity. So coreference resolution is needed to prevent spurious extractions.
- GRIT adopts a generative approach to make extraction decisions. A key innovation is a ***pointer selection module*** in the decoder that can access the entire document. GRIT can also make extraction decisions across sentence boundaries.

4

# Event Extraction Example

A bomb exploded in a Pilmai alley destroying some [**water pipes**].

According to unofficial reports, the bomb contained [**125 to 150 grams of TnT**] and was placed in the back of the [**Pilmai** [**telephone company building**]].

The explosion occurred at 2350 on 16 January, causing panic but no casualties.

The explosion caused damages to the [telephone company offices]. It also destroyed a [**public telephone booth**] and [water pipes].

Witnesses reported that the bomb was planted by [[**two men**] wearing sports clothes], who escaped into the night.
…
They were later identified as [[**Shining Path**] members].

| Role | Role-filler Entities |
|---|---|
| Perpetrator Individual | **two men**, two men wearing sports clothes, Shining Path members *E1* |
| Perpetrator Organization | **Shining Path** *E2* |
| Physical Target | **water pipes**, water pipes *E3* |
| | **Pilmai telephone company building**, telephone company building, telephone company offices *E4* |
| | **public telephone booth** *E5* |
| Weapon | **125 to 150 grams of TnT** *E6* |
| Victim | - |

*Their focus is on entity-level decisions!*

5

# The GRIT Model

- GRIT is built upon the pre-trained transformer model BERT.
- Role filler extraction is treated as a *sequence-to-sequence* task, with an *encoder-decoder* framework.
  - The encoder generates a representation for the input ("source") sequence.
  - The decoder produces role filler extractions as its output ("target") sequence.
- A single pre-trained transformer model is used for both, without fine-tuning it.

6

# A Sequence-to-Sequence Task Model

- The *source sequence* consists of the tokens from the input document, with a [CLS] token prepended and a [SEP] token appended.
- The *target sequence* is the concatenation of the extractions for each event role, with a [SEP] token between each one.
- Each role filler entity is represented by the beginning (b) and end (e) tokens for *its first mention*. Multiple fillers for the same slot seem to be appended without a separator between them.
- The roles are always presented in a fixed order, and the model learns to generate them in this same order.

7

# Notation Examples

Formatting for target sequence:

<S> $e^{(1)}_{1_b}, e^{(1)}_{1_e}$, ... [SEP]
$e^{(2)}_{1_b}, e^{(2)}_{1_e}$, ... [SEP]
$e^{(3)}_{1_b}, e^{(3)}_{1_e}, e^{(3)}_{2_b}, e^{(3)}_{2_e}$, ... [SEP]

| | | |
|---|---|---|
| Perpetrator Individual | **two men**, two men wearing sports clothes, Shining Path members | two men [SEP] |
| Perpetrator Organization | **Shining Path** | Shining Path [SEP] |
| Physical Target | **water pipes**, water pipes | water pipes |
| | **Pilmai telephone company building**, telephone company building, telephone company offices | Pilmai building |
| | **public telephone booth** | public booth [SEP] |
| Weapon | **125 to 150 grams of TnT** | 125 TnT [SEP] |
| Victim | - | [SEP] |

8

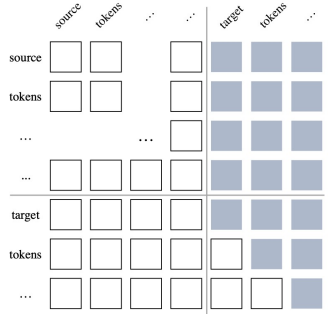
## Pointer Embeddings

- GRIT captures “pointers” that allow the decoder to refer back to places in the source document where the extractions came from.
- As the input to BERT, they use the sum of token, position, and segment embeddings (“segment” is essentially the sentence number).
- For the target representation, the position for a word corresponds to its position in the source document.
- BERT is given the source document as sequence A and the target information as sequence B.

9

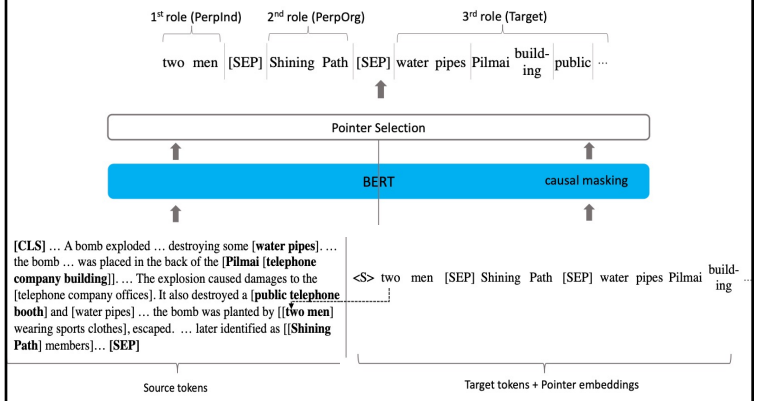
## Causal Mask for Target Embedding

- Since they use a single BERT model for both encoding and decoding, they incorporate causal masks to ensure realistic self-attention.
- When embedding the source document, the target tokens are masked.
- When embedding the target information, the mask only allows self-attention to prior tokens (to make the model autoregressive).




10

## GRIT Example




11

## Pointer Decoding

- The next layer is *pointer decoding*, which selects “pointers” to the source document rather than predicting words.
- When generating the target information at time t, they compute the dot product between the target information at time t-1 and each position in the source document.
- Softmax is then applied to get the probability of pointing to each source document term. Greedy decoding selects the source token with the highest probability.
- The predicted token is added to the target information for the next time step.
- Decoding stops when the last [SEP] token (representing the last event role) is generated!

12

## Comparative Results

*GRIT produced better F1 scores than previous models, primarily due to much higher precision.*

LINKER

Multi-granularity reader

Joint entity, relation, and event extraction

| Models | P | R | F1 |
|---|---|---|---|
| CohesionExtract (Huang and Riloff, 2012) | 58.38 | 39.53 | 47.14 |
| NST (Du and Cardie, 2020) | 56.82 | **48.92** | 52.58 |
| DyGIE++ (Wadden et al., 2019) | 57.04 | 46.77 | 51.40 |
| GRIT | **64.19**** | 47.36 | **54.50*** |

13

## A Deeper Dive into the Results

*Results are Precision/Recall/F1 using an entity-based metric that they defined, called CEAF-REE.*

| | PerpInd | PerpOrg | Target | Victim | Weapon |
|---|---|---|---|---|---|
| NST (Du and Cardie, 2020) | 48.39 / 32.61 / 38.96 | 60.00 / 43.90 / 50.70 | 54.96 / 52.94 / **53.93** | 62.50 / 63.16 / 62.83 | 61.67 / 61.67 / **61.67** |
| DyGIE++ (Wadden et al., 2019) | 59.49 / 34.06 / 43.32 | 56.00 / 34.15 / 42.42 | 53.49 / 50.74 / 52.08 | 60.00 / 66.32 / 63.00 | 57.14 / 53.33 / 55.17 |
| GRIT | 65.48 / 39.86 / **49.55** | 66.04 / 42.68 / **51.85** | 55.05 / 44.12 / 48.98 | 76.32 / 61.05 / **67.84** | 61.82 / 56.67 / 59.13 |

*They also examined subsets of documents with increasing numbers of coreferent mentions per event role.*

| | $k = 1$ | $1 < k \leq 1.25$ | $1.25 < k \leq 1.5$ | $1.5 < k \leq 1.75$ | $k > 1.75$ |
|---|---|---|---|---|---|
| NST (Du and Cardie, 2020) | 63.83 / 51.72 / 57.14 | 57.45 / 38.57 / 46.15 | 60.32 / 49.03 / 54.09 | 64.81 / 50.00 / 56.45 | 66.67 / 51.90 / 58.36 |
| DyGIE++ (Wadden et al., 2019) | **72.50** / 50.00 / 59.18 | 70.00 / 40.00 / 50.91 | 60.48 / 48.39 / 53.76 | 52.94 / 38.57 / 44.63 | 66.96 / 48.73 / 56.41 |
| GRIT | 65.85 / 46.55 / 54.55 | **74.42** / 45.71 / 56.64 | **73.20** / 45.81 / 56.35 | **67.44** / 41.43 / 51.33 | **69.75** / 52.53 / 59.93 |

14

## Performance on documents with nested role fillers across PerpInd & PerpOrg roles

| | PerpOrg (all docs) | PerpOrg (33/200) |
|---|---|---|
| | P / R / F1 | P / R / F1 |
| NST | 56.00 / 34.15 / 42.42 | 80.00 / 44.44 / 57.14 |
| DyGIE++ | 60.00 / **43.90** / 50.70 | 61.54 / 35.56 / 45.07 |
| GRIT | 66.04 / 42.68 / 51.85 | 80.77 / **46.67** / 59.15 |

*GRIT seems to help when there are dependencies across roles.*

...[[[**guerrillas**] of the [**FARC**] and the [**popular liberation army**]] (EPL)] attacked four towns in northern Colombia, leaving 17 guerrillas and 2 soldiers dead and 3 bridges partially destroyed. ...

| | PerpInd | PerpOrg |
|---|---|---|
| Gold Role-filler Entities | • guerrillas, guerrillas of FARC and popular liberation army (EPL) | • EPL, popular liberation army<br>• FARC |
| NST & DyGIE++ | • guerrillas | - |
| GRIT | • guerrillas | • FARC<br>• popular liberation army |

15

## GRIT reduces coreferent entity mentions

**[P1]**... a bomb exploded at the front door of the [home of a peruvian army general], causing damages but no casualties. ... **[P2]** The terrorist attack was ..., by ... who hurled a bomb at the [home of general enrique franco], in the San ... **[P3]** The bomb seriously damaged the [general's [vehicle]], ... and those of [neighboring [houses]].

| | Target |
|---|---|
| Gold Role-filler Entities | • home of peruvian army general, home of general enrique franco<br>• vehicle, general's vehicle<br>• houses, neighboring houses |
| NST | • home of peruvian army general<br>• home of general enrique franco |
| DyGIE++ | • home of peruvian army general<br>• home of general enrique franco<br>• houses |
| GRIT | • home of peruvian army general<br>• houses |

16

## GRIT is also more efficient

- Previous models require an additional classification layer on top of BERT.
- This means that many more parameters must be learned, which substantially increases training time.

| | additional params | training cost |
|---|---|---|
| DyGIE++ | $2H(\#\text{roles}+1)$ | ~20h |
| NST | $H(2\#\text{roles}+1)$ | ~1h |
| GRIT | 0 | <40min |

17

## Template Creation

- The complete document-level event extraction task requires the creation of one template per event, with the role fillers for each event placed in the appropriate template.
- This part of the task adds substantial additional complexity: event types must be identified, multiple events of the same type must be distinguished, and role fillers must be assigned to the appropriate event template.
- In the 1990s, the template creation process involved manual engineering that often involved discourse heuristics.
- A few early attempts were made to automate this process, but they were clunky and achieved only moderate success.

18

## Template Filling Example

Several attacks were carried out in La Paz last night, one in front of government house ...

The self-styled "**Zarate armed forces**" sent simultaneous written messages to the media, calling on the people to oppose ...

The first attack occurred at 22:30 in front of the economic ministry, just before President Paz Zamora concluded his message to ...

Roberto Barbery, has reported that dynamite sticks were hurled from a car.

The second attack occurred at 23:35, just after the cabinet members had left government house where they had listened to the presidential message.

A bomb was placed outside government house in the parking lot that is used by cabinet ministers. The police ...

As of 5:00 today, people found that an old shack on the estate was set ablaze,

| Event 1 Template | Attack |
|---|---|
| Perpetrator Indiv. | - |
| Perpetrator Org | **Zarate armed forces** |
| Physical Target | economic ministry |
| Weapon | dynamite sticks |
| Victim | - |

| Event 2 Template | Bombing |
|---|---|
| Perpetrator Indiv. | - |
| Perpetrator Org | **Zarate armed forces** |
| Physical Target | government house |
| Weapon | bomb |
| Victim | - |

| Event 3 Template | Arson |
|---|---|
| Perpetrator Indiv. | - |
| Perpetrator Org | **Zarate armed forces** |
| Physical Target | old shack |
| Weapon | - |
| Victim | - |

19

## Template Filling with Generative Transformers

[Du et al., NAACL 2021]

- Du et al. recently revisited the challenge of automatically creating and filling distinct event templates.
- Their model (GTT) tackles the problem of event recognition and role filler extraction at the same time using generative transformers.
- This model is built upon their earlier GRIT system.
  - The GRIT model was extended to also predict event types.
  - They modify the decoder to attend to the event types.

20

## Task Definition

- The task is defined as having m possible event types ($T_1$ … $T_m$).
- Each template contains k possible roles ($r_1$ … $r_k$).
- The first slot is defined as the event type, and the remaining k slots correspond to the event roles (in a fixed order).
- For each input document, the system should produce *d* templates, where $d \geq 0$.
  - Note that $d = 0$ means that no relevant events are detected in the document. So this system is essentially doing event recognition as well as role filler extraction!

21

## Casting the Task as Sequence Generation

- The *source sequence* consists of the words in the document prepended with all possible event types followed by the [SEP_T] token representing an event boundary token.
- The [CLS] token is put at the beginning and [SEP] at the end.

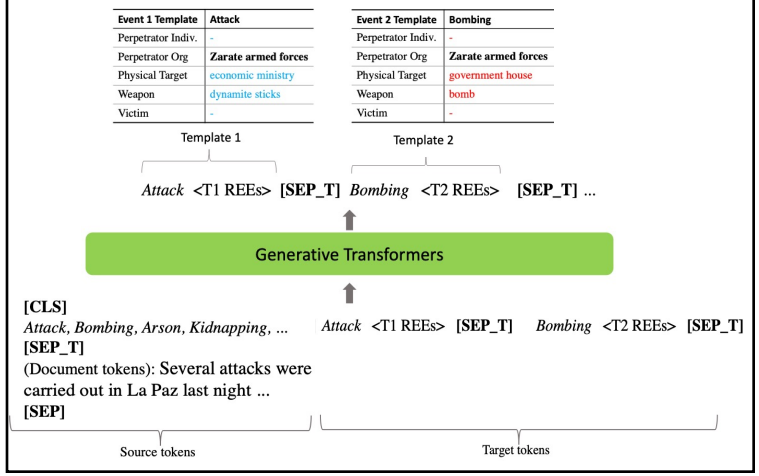
$$\begin{aligned}&[\text{CLS}]\ T_1, ..., T_m\ [\text{SEP\_T}]\\ &x_1, x_2, ..., x_n\ [\text{SEP}]\end{aligned}$$

- The *target sequence* is a concatenation of template extractions, separated by the [SEP_T] token.
- Each template has an event type followed by its role fillers.

$$\begin{aligned}&[\text{CLS}]\ T^{(1)}, < \text{Role-filler Entities} >^{(1)}\\ &[\text{SEP\_T}]\ T^{(2)}, < \text{Role-filler Entities} >^{(2)}\\ &...\\ &[\text{SEP\_T}]\ T^{(i)}, < \text{Role-filler Entities} >^{(i)}\end{aligned}$$

22

## Template Filling Architecture




23

## Model and Decoding Constraints

- The overall architecture is very similar to GRIT.
- This model uses the same encoder/decoder design and causal attention mask approach.
- The source sequence and target sequence are given to BERT to obtain their contextualized representations.
- The same pointer decoding approach is used as well. At each time step, the model generates the source token that has the highest probability.
- A new constraint is added that a document can have at most *k* events. Decoding stops after *k* templates have been generated.

24

## Experimental Results

***Complex evaluation method*:** output templates need to be aligned with gold templates. Event types MUST match. Then event fillers scored for matching templates. For missing templates, ALL role fillers are considered missing.

| Models | Event Type | PERPIND | PERPORG | TARGET | VICTIM | WEAPON |
|---|---|---|---|---|---|---|
| GRIT-PIPELINE | 62.28 | 38.40 | 35.36 | 36.30 | **54.97** | 53.45 |
| DYGIE++ (Wadden et al., 2019) | 61.95 | 32.44 | 25.73 | **45.04** | 49.48 | 51.60 |
| SEQTAGGING (Du and Cardie, 2020) | 60.22 | 30.59 | 26.79 | 36.60 | 43.62 | 51.70 |
| GTT | **67.44** | **44.04** | **41.79** | 32.39 | 54.12 | **59.71** |

*Alternative approach for comparison:*

GRIT-PIPELINE: (1) GRIT is used for role-filler entity extraction. (2) Event types are assigned to each entity as a multi-label classification task.

25

## Digression …

**Multi-class classification:** k different classes. Each instance can belong to just one class.

**Multi-label classification:** k different classes. Each instance can belong to multiple classes.

*Trick*: Multi-label classification can be transformed into multi-class classification. How? Define a new set of classes representing the power set of the original k classes.

If k = 3, then define $2^3 = 8$ new classes representing all possible combinations of the 3 original class labels:

{} {a} {b} {c} {ab} {ac} {bc} {abc}

26

## Micro-average Results on Test Set

| Models | P | R | F1 |
|---|---|---|---|
| GRIT-PIPELINE | **63.88** | 37.56 | 47.31 |
| DYGIE++ (Wadden et al., 2019) | 61.90 | 36.33 | 45.79 |
| SEQTAGGING (Du and Cardie, 2020) | 46.80 | 38.30 | 42.13 |
| GTT | 61.69 | **42.36** | **50.23*** |

*GTT achieves recall gain, but at cost of some precision.*

27

## Documents with Multiple Events

| Models | P | R | F1 | Δ |
|---|---|---|---|---|
| GRIT-PIPELINE | 65.17 | 26.05 | 37.22 | -21.33% |
| DYGIE++ | 69.90 | 27.05 | 39.01 | -14.81% |
| SEQTAGGING | 51.00 | 29.06 | 37.02 | -12.13% |
| GTT | 56.76 | 38.08 | 45.58 | -9.26% |

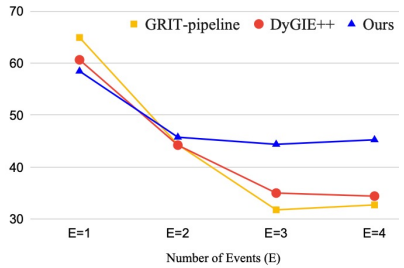
Table 3: Performance on the subset of documents which contain more than one gold event. Δ: relative change of F1, as compared to the Full Test setting.

Note that F1 scores are lower for these documents because the task is harder in these cases.

But performance degrades much less for GTT than the other systems.

28

## Performance vs. Number of Events




GTTs performance is similar to other systems for documents with just 1-2 events but much better when there are more events.

29

## Conclusions

- Performance on document-level event extraction had hit a plateau but has improved recently.
- However, template generation is still far from solved. And role filler extraction performance has substantial room for improvement.
- My take: we need better discourse modeling to understand the relationships between information in different sentences.
  - Most of the NLP tasks that we focus on are either sentence-based or they depend relatively little on full document understanding.

30

## Subjectivity in Language

- Subjective language is the expression of ***private states***: opinions, sentiments, emotions, evaluations, beliefs, speculations, stances.
- A *private state* is not open to objective observation or verification. [Quirk et al., 1985]
- ***Subjectivity analysis*** is the general task of identifying private states mentioned in text.
- ***Subjectivity classification*** determines whether text is subjective or objective.

1

## Types of Subjectivity

**Emotions:** emotional state of someone
*"I am angry/happy/excited/sad."*

**Evaluations:** emotion or judgement toward something
*"Great product!", "What an idiot."*
*"This movie is action-packed and thrilling"*

**Opinions***:* a view formed about something:
*"The minimum wage should be increased by $2/hr."*

**Stances:** a position taken by an entity
*"The University of Utah is against the new policy"*

**Beliefs:** a personal belief
*"I think that UFOs are real."*

**Speculations:** speculation, uncertainty, allegations, hedging
*"I suspect that the butler did it."*

2

## Sentiment & Opinion Analysis

- ***Sentiment Analysis*** is often used an umbrella term for recognizing many types of subjective language. But most commonly it focuses just on identifying positive and negative emotions or evaluations.
- ***Opinion Analysis*** is more broad and also includes opinions that may not be emotions or evaluations.
- Classifiers typically assign ***polarity*** (or ***orientation***) labels: ***positive***, ***negative***, *or* ***neutral***.
- NLP systems can operate at different levels of granularity:
  - document or sentence classification
  - extracting sentiments & aspects
  - extracting opinion expressions, targets, & sources/holders

3

## Sentiment Lexicons

Many sentiment lexicons and lists have been created by hand, semi-automatically, or automatically. A few are:

- General (Harvard) Inquirer [Stone et al., 1966]
- Liu et al's opinion lexicon [Liu et al., 2005]
- OpinionFinder MPQA Subjectivity lexicon [Wilson et al., 2005]
- SentiWordNet [Esuli and Sebastiani, 2006]
- Micro-WNOp [Cerini et al. 2007]
- AFINN, designed for microblogs [Nielson, 2011]
- NRC Word-Emotion Association Lexicon [Mohammad & Turney, 2013]
- NRC Valence, Arousal, and Dominance (VAD) Lexicon [Mohammad, 2016]

4

## Semi-Supervised Induction of Sentiment Lexicons

- Sentiment lexicons are often produced with semi-supervised learning methods that begin with "seed" words and then apply heuristics to learn the polarity of new words.
- One of the earliest works by Hatzivassiloglou and McKeown (1997) exploited the observation that conjoined adjectives usually have the same polarity.
  - Ex: "*happy and excited*" , "*sad and depressed*"
- A general class of methods often used for this problem are **label propagation** algorithms. Terms are encoded in a graph structure with edges capturing their similarity and sentiment values are iteratively refined.

5

## Sentiment Propagation Algorithm

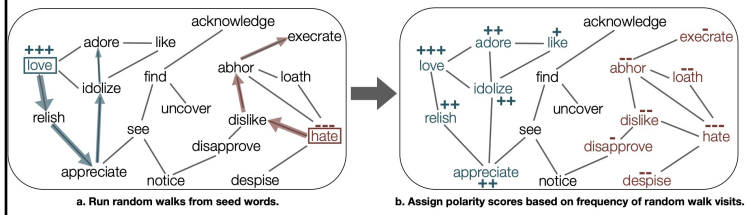
[Hamiton et al., 2016]

1. Build a graph with vertices representing the words in your corpus.
2. For each vertex v, add an edge between v and each of its k nearest neighbors, based on cosine similarity of the word embedding vectors. The cosine similarity score is also used as the edge weight.
3. Perform "random walks" starting with the seed word nodes. The next node w is chosen with probability proportional to the edge weights from v to w.
4. Positive and negative scores can then be computed for each node based on the likelihood of reaching it from a positive or negative seed word.
5. To be more robust over different seed sets, confidence estimates were produced by repeated experiments with different subsets of the seed words.

6

## Examples

| Domain | Positive seed words | Negative seed words |
|---|---|---|
| Standard English | good, lovely, excellent, fortunate, pleasant, delightful, perfect, loved, love, happy | bad, horrible, poor, unfortunate, unpleasant, disgusting, evil, hated, hate, unhappy |
| Finance | successful, excellent, profit, beneficial, improving, improved, success, gains, positive | negligent, loss, volatile, wrong, losses, damages, bad, litigation, failure, down, negative |
| Twitter | love, loved, loves, awesome, nice, amazing, best, fantastic, correct, happy | hate, hated, hates, terrible, nasty, awful, worst, horrible, wrong, sad |

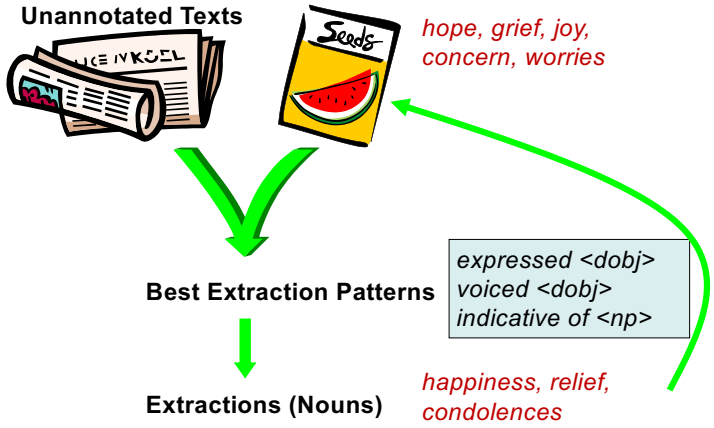



a. Run random walks from seed words.

b. Assign polarity scores based on frequency of random walk visits.

7

## Bootstrapped Learning of Subjective Nouns and Patterns [Riloff, Wiebe, Wilson, 2003]




8

## Learning Subjective Expressions

| | |
|---|---|
| expressed <dobj> | condolences, hope, grief, views, worries |
| indicative of <np> | compromise, desire, thinking |
| inject <dobj> | vitality, hatred |
| reaffirmed <dobj> | resolve, position, commitment |
| voiced <dobj> | outrage, support, skepticism, opposition, gratitude, indignation |
| show of <np> | support, strength, goodwill, solidarity |
| <subj> was shared | anxiety, view, niceties, feeling |

9

## Examples of Strong Subjective Nouns

| | | |
|---|---|---|
| anguish | exploitation | pariah |
| antagonism | evil | repudiation |
| apologist | fallacies | revenge |
| atrocities | genius | rogue |
| barbarian | goodwill | sanctimonious |
| belligerence | humiliation | scum |
| bully | ill-treatment | smokescreen |
| condemnation | injustice | sympathy |
| denunciation | innuendo | tyranny |
| devil | insinuation | venom |
| diatribe | liar | |
| exaggeration | mockery | |

10

## Examples of Weak Subjective Nouns

| | | |
|---|---|---|
| aberration | eyebrows | resistant |
| allusion | failures | risk |
| apprehensions | inclination | sincerity |
| assault | intrigue | slump |
| beneficiary | liability | spirit |
| benefit | likelihood | success |
| blood | peaceful | tolerance |
| controversy | persistent | trick |
| credence | plague | trust |
| distortion | pressure | unity |
| drama | promise | |
| eternity | rejection | |

11

## Contextual Polarity

- Sentiment lexicons capture the **prior polarity** of words and phrases.
- However, the polarity of a word often depends on context due to polysemy, negation, polarity shifters, scoping, expressions, etc.

Example from [Wilson, Wiebe, & Hoffmann 2005]:

*Philip Clapp, president of the National Environment **Trust**, sums up **well** the general thrust of the reaction of environmental movements: "There is no **reason** at all to believe that the **polluters** are suddenly going to become **reasonable**.*

12

## Negation and Polarity Shifters

- People often try simple approaches to sentiment classification that simply count the number of positive vs. negative words found in a lexicon.
- But this approach is often inaccurate because it ignores negation and polarity shifters, which TOGGLE polarity.

There was a lot of damage. → NEG
There was no damage. → POS
There was little damage.→ POS

She showed much empathy. → POS
She never showed empathy. → NEG
She showed a lack of empathy. → NEG

13

## Opinion Extraction

**Opinion extraction** systems typically aim to recognize three components of opinions:

Opinion Expression*:* phrase that describes an attitude toward or evaluation of something

Opinion Holder (Source)*:* the entity whose opinion is being expressed

Opinion Target*:* the entity, object, or concept that the opinion is about (toward)

*According to UN officials, the human rights record in Syria is horrendous.*

14

## Opinion Holders and Targets

- Opinion holders are typically either the *speaker/writer* (implicitly) or a Person or Organization entity (but could also be something like a report).
- Opinion targets can be almost anything! So identifying the boundaries is difficult. For example:
  - I dislike John.
  - I dislike the uniforms of the Utah Jazz.
  - I dislike the recent actions of the government that raised tariffs on tens of thousands of German products, which resulted in ….

15

## Extracting Opinion Propositions and Holders

[Bethard et al., 2004] developed one of the earliest systems to identify propositional opinions and the opinion holders (sources).

- **Opinion**: answer to the question "How does X feel about Y"
- **Propositional Opinion**: an opinion localized in an argument of a verb, generally a sentential complement.
- **Opinion Holder**: the entity who holds the opinion

For example:

- *She believes [you have to use the system to change it].*
- *Still, Vista officials realize [they're relatively fortunate].*
- *["I'd be destroying myself"] replies Mr. Korotich.*

16

## Extracting Opinion Propositions and Holders

[Bethard et al., 2004] developed one of the earliest systems to identify propositional opinions and the opinion holders (sources).

- **Opinion**: answer to the question "How does X feel about Y"
- **Propositional Opinion**: an opinion localized in an argument of a verb, generally a sentential complement.
- **Opinion Holder**: the entity who holds the opinion

For example:

- *She believes [you have to use the system to change it].*
- *Still, Vista officials realize [they're relatively fortunate].*
- *["I'd be destroying myself"] replies Mr. Korotich.*

17

## Data

The first step is to label sentences with respect to 3 categories:

**NON-OPINION**, **OPINION-PROPOSITION**, or **OPINION-SENTENCE**

An OPINION-SENTENCE contains an opinion that extends beyond the scope of a verb argument.

Examples:

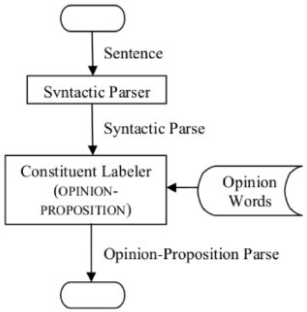
OPINION-PROPOSITION: "*It makes the system more flexible argues a Japanese businessman.*"

OPINION-SENTENCE: "*It might be imagined by those who are not themselves Anglican that the habit of going to confession is limited only to markedly High churches but that is not necessarily the case.*"

18

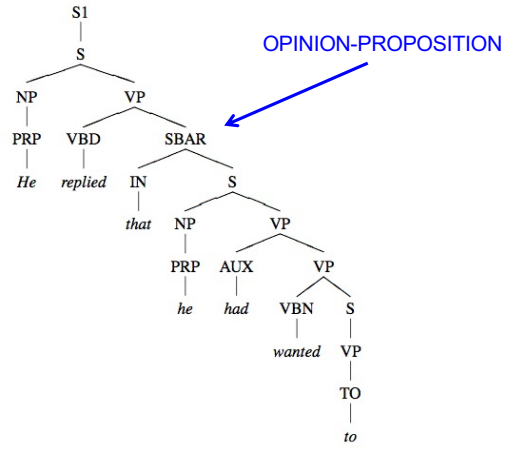
## One-Tiered Architecture

The first system is an SVM classifier that labels syntactic constituents in a parse tree as either OPINION-PROPOSITION or NULL.




19

## Example




20

## Opinion-Proposition Classifier

They followed the same design as a semantic role labeling classifier by [Pradhan et al., 2003] with 8 syntactic features:

1. the verb
2. verb's cluster
3. subcategorization type of the verb
4. syntactic phrase type of the potential argument
5. head word of the potential argument
6. before/after position of the argument relative to the verb
7. parse tree path between verb and potential argument
8. voice (active/passive) of the verb

This feature set was later augmented with features derived from an acquired set of opinion words.

21

## Opinion Word Features

Given a constituent to classify, the following features captured opinion word information:

**Counts**: the number of opinion words in the constituent.

**Score Sum**: the sum of the opinion scores for each opinion word in the constituent, sometimes with a minimum score threshold.

**ADJP**: a binary feature indicating whether the constituent contains a complex adjective phrase. (Simple adjectives produce many false hits.) For example:

*excessively affluent*
*more bureaucratic*

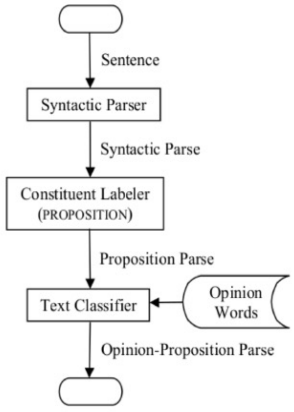
[Note: I've observed "ADV ADJ" to be a useful pattern too!]

22

## Two-Tiered Architecture

The second system performs two steps:

1. An SVM classifier is trained to label constituents only for the PROPOSITION role.
2. A second classifier determines whether the proposition is an OPINION-PROPOSITION, using a sentence-level approach.




23

## Labeling Propositions as Opinions

Three Naïve Bayes classifiers were trained to determine whether a proposition is an OPINION-PROPOSITION.

1. The first model is trained using approximate sentence labels from fact-heavy vs. opinion-heavy texts.

   Sentences in editorials and letters to the editor are assumed to contain opinions.

   Sentences in news and business articles are assumed to be factual.

   The sentence containing each proposition is classified and the proposition is assigned the label of its sentence.

2. The second model is trained at the sentence-level but predictions are based only on the text of the proposition.

3. For the third model, both training and testing use only the text of the propositions (with the same approximate labeling during training).

24

## Results for Two-Tiered Architecture

The first component that labels PROPOSITION constituents achieved 62% recall with 82% precision. (This was a 10% precision gain over the more general semantic role classifier.)

The results for the 3 models to determine which PROPOSITION constituents are opinions are shown below:

| Train on | Predict on | Measure | Features | | | | |
|---|---|---|---|---|---|---|---|
| | | | Words | Bigrams | Trigrams | POS | Orientation |
| Sentence | Sentence | Recall | 33.38% | 29.69% | 30.09% | 30.05% | 43.72% |
| | | Precision | 67.84% | 63.13% | 62.50% | 65.55% | 67.97% |
| Sentence | Proposition | Recall | 37.48% | 37.32% | 37.79% | 36.03% | 28.81% |
| | | Precision | 53.95% | 59.00% | 59.83% | 55.00% | 68.41% |
| Proposition | Proposition | Recall | 42.77% | 38.07% | 37.84% | 35.01% | 25.75% |
| | | Precision | 59.56% | 61.63% | 60.43% | 58.77% | 61.66% |

Table 5: Two-tiered Approach Results for Opinion Propositions.

25

## Results and Conclusions

- This work focused on one type of opinion recognition, **propositional opinions**, and identified the opinion holders.
- This approach is very syntactically-oriented, requiring an alignment between the propositions/holders and syntactic constituents.
  - This approach cannot identify cases where a proposition spans multiple sentences, or the holder is in a different sentence than the proposition.
- The two architectures exhibited a recall/precision trade-off:
  51% R with 58% P for 1 Tiered
  43% R with 68% P for 2 Tiered

26

## Opinion Role Extraction

- [Wiegand & Ruppenhofer, CoNLL 2015] observed that *opinion roles* (holders and targets) of *opinion verbs* are often aligned with the verb's semantic roles, but the mapping depends on the verb.

*Peter criticized Mary*
Holder Target

*Peter disappoints Mary*
Target Holder

- Also, many opinion verbs implicitly express the sentiment of the speaker, so the opinion holder does not appear. For example:

*At my work, they are constantly gossiping.*

27

## Verb Categorization for Opinion Roles

- [Wiegand & Ruppenhofer, CoNLL 2015] propose that verbs can be grouped into 3 categories that share ***opinion role subcategorization*** frames.
- Thus, categorizing a verb into one of these 3 classes essentially dictates how to extract its opinion roles!
- There are often multiple ways to interpret an opinion, partly due to polysemy, but for the sake of simplicity they only seek to learn the most common interpretation for a verb.

28

## Verbs with Agent View (AG)

Verbs with an Agent View convey the opinion of its agent, so the agent is the opinion holder and the "patient" is the opinion target.

[Here, "patient" refers to a semantic role capturing the object of the verb's action, such as what was criticized or argued. Similar to its theme.]

*Peter criticized Mary.*
*They like the idea.*
*The guests complained about the noise.*
*They argue that this plan is infeasible.*

Note that the patient of the verb can appear in different types of syntactic constituents, including NPs, PPs, and infinitive or complement phrases!

29

## Verbs with Patient View (PT)

Verbs with a Patient View are the opposite, and have the agent as the opinion target and the patient as the opinion holder.

*Peter disappoints Mary.*
*The noise irritated the guests.*
*The gift pleased her very much.*
*The policy raising tariffs on German products upset Tom.*

30

## Verbs with Speaker View (SP)

Verbs with a Speaker View express an opinion that comes from the speaker/writer (implicitly), rather than from an entitiy involved in the action. The agent is usually the target.

*At my work, they are constantly gossiping.*
*They cheated in the exam.*
*The young baseball player has improved.*
*He besmirched the King's name.*

Occasionally, the patient of Speaker View verbs can <u>also</u> be a target, as for "besmirched" above. But since this is less common, they only assign the agent to be a target.

31

## Opinion Verbs & Gold Standard

- This work aims to classify an existing set of opinion verbs to extract the holders and targets from their arguments.
- They use the 1,175 opinion verb lemmas in the Subjectivity Lexicon [Wilson et al., 2005].
- They annotated their semantic roles to be consistent with PropBank and assigned holder, target, and speaker labels.
- They measured inter-annotator agreement on 400 verbs:
  opinion holders: $\kappa = 60.8$
  opinion targets: $\kappa = 62.3$
  speaker views: $\kappa = 59.9$

32

## Distribution of Verb Types in Gold Standard Data

| Agent (AG) | | Patient (PT) | | Speaker (SP) | |
|---|---|---|---|---|---|
| Freq | Percent | Freq | Percent | Freq | Percent |
| 450 | 38.3 | 188 | 16.0 | 537 | 45.7 |

33

## Pattern-based Seed Initialization

- **To induce seed AG verbs**: identify verbs that co-occur with ***prototypical opinion holders*** (e.g., *opponents* and *critics*). For example:

  *Opponents claim these arguments miss the point.*
  *Critics argue that the proposed limits were unconstitutional.*

- **To induce seed PT verbs**: identify **morphologically related adjectives**, which then reveal a PT verb. Specifically, they identify verbs in past participle form that are identical to a predicate adjective. For example:

  *He has upset me.* *I am upset.*

- **To induce seed SP verbs**: extract verbs using 3 patterns: (1) *accused of* $X_{VBG}$, (2) *blamed for* $X_{VBG}$, and (3) *help to* $X_{VB}$. For example:

  *He was accused of falsifying the documents.*
  *The UN was blamed for misinterpreting climate data.*

34

## Extracted Verb Seeds

They used the North American News Text Corpus for seed extraction and comparison of verb similarities.

The top 12 extracted verb seeds for each category were:

| | |
|---|---|
| **AG** | argue, contend, speculate, fear, doubt, complain, consider, praise, recommend, view, acknowledge, hope |
| **PT** | interest, surprise, please, excite, disappoint, delight, impress, shock, trouble, embarrass, annoy, distress |
| **SP** | murder, plot, incite, blaspheme, bewitch, bungle, despoil, plagiarize, prevaricate, instigate, molest, conspire |

Ultimately they used the top 40 AG verbs, 30 PT verbs, and 50 SP verbs.

35

## Similarity Metrics

1. **Word Embeddings**: get embedding vectors for opinion words using word2vec, and compare with cosine similarity.

2. **WordNet**: used a similarity metric based on WordNet's graph structure (WordNet::Similarity).

3. **Coordination**: coordination typically preserves semantic coherence, so they apply a dependency parser to detect verb coordination (e.g., *They criticize and hate him*).

   The similarity score is just the frequency of two verbs appearing in a conjunction.

36

## Digression: Pointwise Mutual Information

**Pointwise mutual information (PMI)** measures the degree to which two words are statistically dependent.

$$PMI(w_1, w_2) = \log_2 \left[ \frac{P(w_1 \,\&\, w_2)}{P(w_1) * P(w_2)} \right]$$

Mutual information can be used to measure the strength of association between a relation and its slot fillers:

$$MI(a, Slot, b) = \log_2 \left[ \frac{|\, a, Slot, b \,|\; X \;|\, *, Slot, * \,|}{|\, a, Slot, * \,|\; X \;|\, *, Slot, b \,|} \right]$$

37

## MI Example for Relation & Filler

MI (*will buy*, SlotX, *Smith*) =

$$\log_2 \left[ \frac{|\, will\ buy, SlotX, Smith \,|\; X \;|\, *, SlotX, * \,|}{|\, will\ buy, SlotX, * \,|\; X \;|\, *, SlotX, Smith \,|} \right]$$

where:

|*will buy*, *SlotX*, *Smith*| = # times *Smith* fills SlotX for path *will buy*

|*, *SlotX*, *|* = total # fillers for SlotX in all paths

|*will buy*, *SlotX*, *|* = # fillers in SlotX for path *will buy*

|*, *SlotX*, *Smith*|= # times *Smith* fills SlotX in all paths

38

## Dependency-based Similarity

**4. Distributional similarity based on dependency relations:**
*(x, r, y)* where *x* and *y* are words and *r* is a relation.
Example: (argue-V, nsubj, critics-N)

$$sim(v1, v2) = \frac{\sum_{(r,w) \in T(v1) \cap T(v2)} MI(v1, r, w) + MI(v2, r, w)}{\sum_{(r,w) \in T(v1)} MI(v1, r, w) + \sum_{(r,w) \in T(v2)} MI(v2, r, w)}$$

where T(x) is the set of pairs (r, y) such that

$$\log_2 \left[ \frac{|\, x, r, y \,|\; X \;|\, *, r, * \,|}{|\, x, r, * \,|\; X \;|\, *, r, y \,|} \right] > 0$$

From DIRT = [Lin & Pantel, 2001]

39

## Semi-Supervised Learning: Propagation Methods

They tried 2 methods to propagate labels from the seed verbs to new verbs.

1. **K nearest neighbor classification**: kNN methods identify the k closest (most similar/related) neighbors to input X and typically assign the predominant label to X.

2. **Label Propagation**: label propagation is a class of semi-supervised learning methods that iteratively propagate labels across nodes in a graph based on edges with weights that capture similarity/relatedness.

40

## Experimental Results

| | | Acc | Prec | Rec | F1 |
|---|---|---|---|---|---|
| **Baselines** | **Majority Class** | 45.7 | 14.2 | 33.3 | 20.9 |
| | **Only Seeds** | 8.9 | 87.0 | 9.8 | 17.6 |
| **Coordination** | **kNN** | 45.2 | 61.5 | 47.3 | 53.4 |
| | **graph** | 42.7 | 68.7 | 39.7 | 50.4 |
| **WordNet** | **kNN** | 52.8 | 51.5 | 50.7 | 51.1 |
| | **graph** | 51.1 | 51.9 | 51.5 | 51.5 |
| **Embedding** | **kNN** | 59.3 | 58.4 | 61.0 | 59.7 |
| | **graph** | 64.0 | 70.5 | 59.4 | 64.5 |
| **Dependency** | **kNN** | 65.7 | 63.8 | 65.4 | 64.5 |
| | **graph** | **70.3** | **72.0** | **68.0** | **70.6** |

For kNN models, k=3.

41

## Inspecting the Similarity Metrics

Inspecting the most similar words identified by each metric reveals that distributional similarity based on dependency relations indeed looked the best.

The 12 verbs most similar to *outrage*, *which is PT. The underlined words are not PT verbs.*

| | |
|---|---|
| **Coordin.** | appear, believe, refuse, vow, want, offend, shock, help, exhilarate, challenge, support, distort |
| **WordNet** | appal, scandalize, anger, rage, sicken, temper, hate, fear, love, alarm, dread, tingle |
| **Embedd.** | anger, dismay, disgust, protest, alarm, enrage, shock, regret, concern, horrify, appal, sorrow |
| **Depend.** | anger, infuriate, alarm, shock, stun, enrage, incense, dismay, upset, appal, offend, disappoint |

Coordination was probably poor due to sparsity.

42

## Multilingual Results

They also replicated their system to handle German! This illustrates the generality of their approach.

Overall, the same conclusions held.

| | | Coordin. | | WordNet | | Embedd. | | Depend. | |
|---|---|---|---|---|---|---|---|---|---|
| | **Major.** | **kNN** | **graph** | **kNN** | **graph** | **kNN** | **graph** | **kNN** | **graph** |
| English | 20.9 | 53.4 | 50.4 | 51.1 | 51.5 | 59.7 | 64.5 | 64.5 | **70.6** |
| German | 22.9 | 43.8 | 48.9 | 53.2 | 59.9 | 54.3 | 60.9 | 58.3 | **63.1** |

43

## In-Context Classification

- Next, they built a classifier to use their induced verb knowledge for extracting opinion information in sentence contexts.
- They sampled ~1100 sentences from the N. American News Corpus in which their opinion verbs occurred.
- They manually annotated the opinion holders, targets, and targets evoked by speaker views. The data has: 753 holders, 745 targets, and 499 speaker view targets.
- They trained 3 SVM classifiers, one for each type of opinion information.

44

## Feature Set for Contextual Classifier

| Features | Description |
|---|---|
| cand_lemma | head lemma of candidate (phrase) |
| cand_pos | part-of-speech tag of head of candidate phrase |
| cand_phrase | phrase label of candidate |
| cand_person | is candidate a person |
| verb_lemma | verb lemmatized |
| verb_pos | part-of-speech tag of verb |
| word | bag of words: all words within the sentence |
| pos | part-of-speech sequence between cand. and verb |
| distance | token distance between candidate and verb |
| const | path from constituency parse tree from cand. to verb |
| subcat | subcategorization frame of verb |
| $\text{srl}_{propbank}$/dep | semantic role/dependency path between cand. and verb (semantic roles based on PropBank) |
| brown | Brown-clusters of cand_word/verb_word/word |
| $\text{srl}_{framenet}$ | frame element name assigned to candidate and the frame name (to which frame element belongs) |
| fine-grain_lex | is candidate holder/target/$\text{target}_{speaker}$ according to the fine-grained lexicon |
| coarse-grain_lex | is candidate holder/target/$\text{target}_{speaker}$ according to the coarse-grained lexicon |
| $\text{induc}_{graph}$ | is candidate holder/target/$\text{target}_{speaker}$ according to the coarse-grained lexicon automatically induced with graph clustering (and induced seeds (§4.1)) |

45

## Contextual Classification Results

Evaluation results using 10-fold cross-validation:

| Features | Holder | Target | $\text{Target}_{Speaker}$ |
|---|---|---|---|
| standard | 63.59 | 54.18 | 40.06 |
| +$\text{srl}_{framenet}$ | 65.44* | 55.70* | 42.14 |
| +$\text{induc}_{graph}$ | 68.06*° | 59.61*° | 46.66*° |
| +$\text{srl}_{framenet}$+$\text{induc}_{graph}$ | 69.70*° | 60.47*° | 47.33*° |
| +coarse-grain_lex | 68.56*° | 59.89*° | 54.31*°† |
| +$\text{srl}_{framenet}$+coarse-grain_lex | 69.70*° | 60.68*° | 54.06*°† |
| +fine-grain_lex | 69.83*°† | 62.89*°† | **56.71***°† |
| +$\text{srl}_{framenet}$+fine-grain_lex | **70.80***°† | **63.72***°† | 56.64*°† |

statistical significance testing (paired t-test, significance level $p < 0.05$) *: better than *standard*; °: better than +$srl_{framenet}$; †: better than +$induc_{graph}$

46

## Conclusions

- Opinion extraction can be decomposed into subproblems: identifying opinion expressions, and their holders & targets.
- Aligning opinion targets of verbs with semantic roles makes sense linguistically and constrains the boundaries. This work offered a new insight that different classes of verbs behave differently in terms of their opinion roles.
- However, the results are still far from perfect. Opinions vary greatly in scope.
- And not all opinions are expressed as verbs! For other cases, opinion target boundaries can be elusive.

47

## Opinion Roles and Semantic Roles

- Prior research observed that opinion holders and opinion targets often align with the arguments of verbs (or verb phrases) that express an opinion. For example:

(1) Australia said $[\text{it}]_H$ $[\textbf{feared}]_{O_{neg}}$ $[\text{violence}]_T$ if voters thought the election had been stolen.

- Nouns can also take syntactic arguments, and opinion role fillers can also be found in their arguments too.

*The **disgust** of citizens toward the policy was apparent at the rally.*

holder target

- Research question: can semantic role labeling help with opinion extraction?

1

## PropBank-style Semantic Roles

- The **PropBank** project produced semantic role annotations on the Wall Street Journal portion of the Penn Treebank (which already had parse tree annotations).
- PropBank defines predicate-argument structures for verbs with semantic role assignments for each verb's arguments.
- The predicate is labeled as REL (for relation) and is either a verb or a verb + particle (e.g., "*keep up*").
- PropBank's semantic role arguments are not named, but indicated as Arg0, Arg1, Arg2, etc. The meaning is specific to one verb sense! They do not have the same meaning for different verbs or different senses of the same verb.

2

## PropBank Definitions

PropBank provides *Frames files* which defines a set of roles (*roleset*) for verb senses from VerbNet. There are two types of roles: numbered arguments and adjuncts.

**Numbered Arguments**: A0-A5

- Arg0 *usually* refers to the verb's agent.
- Arg1 *usually* refers to the verb's patient/theme (if it has one)
- All other arguments vary from verb to verb.

**Adjuncts**: optional, general arguments that any verb can take

| | |
|---|---|
| AM-ADV : general-purpose | AM-MOD : modal verb |
| AM-CAU : cause | AM-NEG : negation marker |
| AM-DIR : direction | AM-PNC : purpose |
| AM-DIS : discourse marker | AM-PRD : predication |
| AM-EXT : extent | AM-REC : reciprocal |
| AM-LOC : location | AM-TMP : temporal |
| AM-MNR : manner | |

3

## PropBank Frame File Example

*Frame File for the verb 'expect':*

Roles:
Arg0: expecter
Arg1: thing expected

Example: Transitive, active:
*Portfolio managers expect further declines in interest rates.*

| | |
|---|---|
| Arg0: | Portfolio managers |
| REL: | expect |
| Arg1: | further declines in interest rates |

4

## PropBank Framesets

Different senses of a verb may have different semantic roles. In this case, framesets are used to define the semantic roles for each verb sense.

Example for the verb "left":

*Frameset leave.01 "move away from":*
Arg0: entity leaving
Arg1: place left

*Frameset leave.02 "give":*
Arg0: giver
Arg1: thing given
Arg2: beneficiary

5

## PropBank Examples

- [John]$_{ARG0}$ **broke** [the window]$_{ARG1}$.
- [The window]$_{ARG1}$ **broke**.
- [John]$_{ARG0}$ **opened** [the door]$_{ARG1}$ [with his foot]$_{ARG2}$.
- [John]$_{ARG0}$ tried to **kick** [the football]$_{ARG1}$.
- [He]$_{ARG0}$ **expects** [Ford to meet the deadline easily]$_{ARG1}$.

6

## Opinion Role Labeling (ORL) with Semantic Role Labeling (SRL)

- Marasovic & Frank [NAACL 2018] hypothesized that *multi-task learning for opinion role labeling and semantic role labeling* could be effective.
- **Multi-task learning systems** learn a shared representation to accomplish multiple tasks. The tasks should be related, so that both tasks will benefit from a shared representation.
- Earlier work [Katiyar & Cardie, 2016] found that neural models (LSTMs) did not outperform an earlier CRF-based model. Marasovic & Frank hypothesized that the small size of the MPQA dataset used to train these models was a factor.
- Much bigger datasets are available for SRL, so they tried leveraging this data for multi-task learning to see if it could be used to improve opinion role labeling.

7

## SRL Example

Output from a PropBank SRL system on a MPQA sentence:

| | Australia | said | it | feared | violence | if | voters | thought |
|---|---|---|---|---|---|---|---|---|
| say.01 | A0 | - | A1 | A1 | A1 | A1 | A1 | A1 |
| fear.01 | - | - | **A0** | - | **A1** | AM-ADV | AM-ADV | AM-ADV |
| think.01 | - | - | - | - | - | - | A0 | - |
| steal.01 | - | - | - | - | - | - | - | - |

| | the | election | had | been | stolen | . |
|---|---|---|---|---|---|---|
| say.01 | A1 | A1 | A1 | A1 | A1 | - |
| fear.01 | AM-ADV | AM-ADV | AM-ADV | AM-ADV | AM-ADV | - |
| think.01 | A1 | A1 | A1 | A1 | A1 | - |
| steal.01 | A1 | A1 | - | - | - | - |

NOTE: "AM-" refers to adjuncts, which are general arguments that any verb can optionally have). "ADV" is a general-purpose adjunct.

8

## Issues with SRL / ORL Alignment

Note that A0 and A1 roles can be reversed for different verbs:

(1) Australia said [it]$_H$ [**feared**]$_{O_{neg}}$ [violence]$_T$ if voters thought the election had been stolen.

(2) [I]$_H^{A1}$ am very [**pleased**]$_{O_{pos}}$ that [the Council has now approved the Kyoto Protocol thus enabling the EU to proceed with its ratification]$_T^{A0}$.

Note different scopes for the Targets of concerned:

(4) Rice told us [the administration]$_H$ was [**concerned**]$_{O_{neg}}$ that [Iraq]$_T$ would take advantage of the 9/11 attacks.

(5) [The Chinese government]$_H$ is deeply [**concerned**]$_{O_{neg}}$ about [the sudden deterioration in the Middle East situation]$_T$, Tang said.

9

## SRL Model

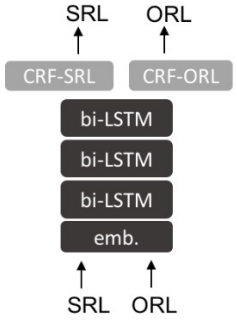
- The neural architecture is an SRL model from [Zhou & Hu, 2015] that performs "end-to-end" SRL without syntactic analysis.
- The model is a stack of bi-directional LSTMs with a CRF layer on top that assigns the semantic roles with IOB tagging.
- Input = the token embeddings plus 3 additional features: the predicate, the predicate's context (+3 words), and an indicator for whether the a token is in the predicate context. For example:

| time | argu | pred | ctx-p | $m_r$ | label |
|---|---|---|---|---|---|
| 1 | A | set | been set . | 0 | B-A1 |
| 2 | record | set | been set . | 0 | I-A1 |
| 3 | date | set | been set . | 0 | I-A1 |
| 4 | has | set | been set . | 0 | O |
| 5 | n't | set | been set . | 0 | B-AM-NEG |
| 6 | been | set | been set . | 1 | O |
| 7 | set | set | been set . | 1 | B-V |
| 8 | . | set | been set . | 1 | O |

10

## Fully-shared (FS) Multi-Task Model

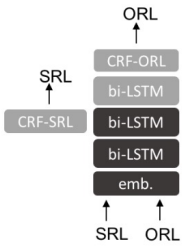
The first model shares all the parameters when learning both tasks, except that each task has its own output layer.




11

## Hierarchical (H-MTL) Multi-Task Model

- Hierarchical models are used when one task is thought to be higher-level than the other but could benefit from its representations. For example, syntactic parsing should benefit from POS tagging representations.
- In their H-MTL model, a model is trained for SRL and then the ORL task is trained with the representations produced by the last LSTM layer for SRL.




12

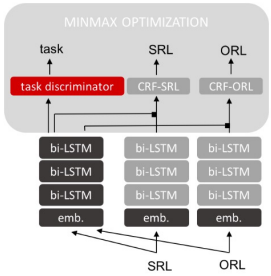
## Shared-Private (SP) Multi-Task Model

- In addition to the stack of shared LSTMs, the model also includes a stack of task-specific LSTMs for each task.
- Representations produced by the outermost shared LSTM and the task-specific LSTM are passed on to the task-specific CRF layer.
- This design allows the system to potentially benefit both from shared representations as well as representations learned for each individual task.

13

## Adversarial Shared-Private (ASP) Model

- The ASP Model aims to prevent the shared model from learning too many task-specific features (which would be redundant with the task-specific LSTMs).
- A task discriminator is added to predict which task the current batch of data belongs to based only on the shared LSTMs. This challenges the shared LSTMs it capture features across both tasks.




14

## Data Sets

- SRL: news articles from the CoNLL-2005 shared task, annotated with PropBank predicate-argument structures.
- ORL: MPQA (Multi-Perspective Question Answering) corpus of news articles with annotations for opinion expressions and opinion roles.

| | task | train size | dev size | test size | $\|\mathcal{Y}\|$ |
|---|---|---|---|---|---|
| CoNLL'05 | SRL | 90750 | 3248 | 6071 | 106 |
| MPQA (4-CV) | ORL | 3141.25 | 1055 | 1036.75 | 7 |
| MPQA (10-CV) | ORL | 3516.3 | 1326 | 349.3 | 7 |

15

## Evaluation Details

1. Set aside 132 documents for development and used remaining 350 documents for 10-fold cross-validation.
2. Set aside 100 documents for development and used the remaining documents for 4-fold cross-validation to have bigger test sets with less variance. Two cross-validation runs were done with random seeds.

- Mean F1 scores are reported, averaged over the folds. The standard deviations are shown as$_{\text{subscripts}}$.
- Binary F1 scores are based on a binary overlap match. Proportional F1 scores are based on proportional overlap match.
- For comparison, they trained the SRL model directly for opinion role extraction (Z&X-STL in the next tables). This is single-task learning.

16

# Opinion Role Extraction Results

*10-fold cross-validation*

| | test (MPQA) | | | |
|---|---|---|---|---|
| | holder | | target | |
| | binary F1 | prop. F1 | binary F1 | prop. F1 |
| Z&X-STL | $80.24_{2.91}$ | $77.98_{2.90}$ | $76.30_{2.55}$ | $71.18_{2.55}$ |
| FS-MTL | $83.47^{\bullet}_{2.26}$ | $81.80^{\bullet}_{2.26}$ | $\mathbf{77.60}_{2.52}$ | $\mathbf{73.77}_{2.28}$ |
| H-MTL | $\mathbf{84.03}^{\bullet}_{2.65}$ | $\mathbf{82.34}^{\bullet}_{2.51}$ | $77.41_{2.14}$ | $73.10_{1.96}$ |
| SP-MTL | $82.19^{\bullet}_{2.49}$ | $80.11^{\bullet\diamond}_{2.36}$ | $76.01_{3.03}$ | $71.51_{3.34}$ |
| ASP-MTL | $83.15^{\bullet}_{2.92}$ | $81.12^{\bullet}_{2.66}$ | $75.89_{2.66}$ | $71.21_{2.78}$ |

*repeated 4-fold cross-validation*

| | test (MPQA) | | | |
|---|---|---|---|---|
| | holder | | target | |
| | binary F1 | prop. F1 | binary F1 | prop. F1 |
| Z&X-STL | $80.42_{1.92}$ | $77.48_{2.06}$ | $73.84_{1.17}$ | $67.03_{2.13}$ |
| FS-MTL | $\mathbf{83.67}^{\bullet}_{1.52}$ | $\mathbf{81.59}^{\bullet}_{1.50}$ | $77.04^{\bullet}_{1.45}$ | $73.01^{\bullet}_{2.53}$ |
| H-MTL | $82.80^{\bullet}_{1.87}$ | $80.40^{\bullet}_{1.91}$ | $\mathbf{77.12}^{\bullet}_{1.34}$ | $\mathbf{73.16}^{\bullet}_{1.78}$ |
| SP-MTL | $82.51^{\bullet}_{2.17}$ | $80.03^{\bullet}_{2.00}$ | $74.61^{\diamond}_{1.32}$ | $68.70^{\diamond}_{2.32}$ |
| ASP-MTL | $81.77^{\diamond}_{1.74}$ | $79.32^{\diamond}_{1.62}$ | $74.92^{\bullet\diamond}_{0.84}$ | $69.89^{\bullet}_{1.80}$ |

17

# Analysis

- They did a detailed analysis comparing the results of the single-task model (Z&X-STL) and the FS-MTL model using the 4-fold evaluation results. That produced 8 different systems (4 for each of the 2 random seeds).
  - They considered a model to be *correct* if it predicted a role overlapping with the correct role in 6 of 8 systems.
  - They considered a model to be *incorrect* if it predicted a role overlapping with the correct role in at most 2 systems.
  - Other cases were considered to be *inconsistent* across sthe models.
- They categorized an instance as *Easy* if both models were correct. They categorized an instance as *Hard* if neither model got it correct.

18

# Analysis Statistics

| HOLDER | easy | hard |
|---|---|---|
| % opinions that are predicates | 91.32 | 93.33 |
| % holders that are subjects | 77.84 | 38.79 |
| % holders that are A0 roles | 74.10 | 33.33 |
| avg. distance between holders & opinions | 1.54 | 7.56 |

| TARGET | easy | hard |
|---|---|---|
| % opinions that are predicates | 92.58 | 89.20 |
| % target's heads that are objects | 22.12 | 14.77 |
| % targets that are A1 roles | 70.62 | 42.61 |
| % targets that are A2 roles | 9.00 | 0.57 |
| avg. distance between targets & opinions | 2.29 | 8.46 |

19

# Correct Opinion Holders

1 *Malinga* $_{FS,ZX}$ **said** according to the guidelines in the booklet, the election had been legitimate .

2 movie um-hum that 's interesting so that was a good movie too well do *you* $_{FS,ZX}$ **think** we've covered baseball i think so okay well have a good night

3 *The nation* $_{FS,ZX}$ should certainly **be concerned** about the plans to build a rocket launch pad , work on the infrastructure for which is due to start in 2002 , with launches beginning from 2004 .

4 Bam on Sunday said *she* $_{FS,ZX}$ **believed** Zimbabwe's election was not free and fair , adding they were not in line with international standards as well as those of her organisation .

5 The majority report , endorsed only by the ANC , said *the observer mission* $_{FS,ZX}$ **had noted** that over three million Zimbabweans had cast their votes and this substantially represented the will of the people .

# Correct Opinion Targets

1 Indonesia **has come under pressure** from several quarters to *take tougher action against alleged terrorist leaders* but has played down the threat $_{ZXFS}$ .

2 Mugabe even talked about his **desire** to *keep safeguarding Zimbabwe 's sovereignty and land* $_{ZX}$ *in spirit* $_{FS}$ *when he dies* , a dream which the veteran leader said forced him to sacrifice a bright teaching career in the 1950s to lead [...].

3 Under his blueprint , the government **hopes** to *stabilize the economy* through curtailing state expenditure , reforming public enterprises and expanding agriculture $_{FS,ZX}$ .

4 He said those who thought the election process would be rigged were supporters of the MDC party , adding that they were prejudging and **wanted** to *direct the process* $_{FS,ZX}$ .

5 People in the rural areas support the ruling party because our party has been genuine on **its policy** on *land reform* $_{FS,ZX}$.

20

## Incorrect Opinion Holders

| | |
|---|---|
| 1 | It would be entirely improper if , in *its* **defense of** Israel $_{FS}$ , the United States continues to exert pressure on [...] . |
| 2 | Indonesia $_{FS,ZX}$ **has come under pressure** from *several quarters* to take tougher action against alleged terrorist leaders but has played down the threat . |
| 3 | Australia should adhere to the *Cardinal Principle of International Law* , which **states** that all nations in the world must first respect and promote the humanitarian interests and progress of all humankind . |
| 4 | *The department* said that it will cost $ 600 for an HIV/AIDS patient per year at this time , and the following years this cost is **expected** to stand at just $ 400/year for one patient as the production of such drugs becomes stable . |
| 5 | The Organisation of African Unity OAU $_{ZX}$ also backed Zimbabwean President Robert Mugabe 's re-election , with its observer team $_{FS,ZX}$ **describing** the poll as " transparent , credible , free and fair " . |
| 6 | Regarding the American proposed Anti-Missile Defense System too , neither Russia , China , Japan , nor even the European Union , had shown any enthusiasm ; rather *they* $_{FS}$ had all $_{FS,ZX}$ **expressed** their reserves on the project . |
| 7 | The president renewed his pledge to thwart terrorist groups $_{FS,ZX}$ *who* **want** to " mate up " with regimes hoping to acquire weapons of mass destruction and said " nations will come with us " if the US-led war on terrorism is extended . |

The correct answer is *italicized*.
Blue = FS-MTL Yellow = Z&X-STL Green = Both

21

## Incorrect Opinion Targets

| | |
|---|---|
| 1 | *State-sanctioned land invasions* , several times **declared** illegal by Zimbabwe 's courts , as well as a drought have disrupted Zimbabwe 's food production and famine is already looming in much of the country . |
| 2 | But he **told** the nation $_{FS,ZX}$ that in spite of *stiff opposition to the agrarian reforms from powerful Western countries , especially the country 's former colonial power of Britain* , he would press ahead to seize farms from whites and [...] . |
| 3 | If the Europeans wish to influence Israel in the political arena – in *a direction* that many in Israel **would support wholeheartedly** – they will not be able to promote their positions in such a manner . |
| 4 | They $_{FS,ZX}$ are fully aware that these are dangerous *individuals* , he **said** during a press conference [...] . |
| 5 | And her little girl **just complained** , " I don't want to *wash the dishes* " . |
| 6 | During *President Bush's speech* , I **thought** of heckling $_{ZX}$ ; ' What are you going to do with the Kyoto Protocol ? $_{FS}$ ' |
| 7 | At first I **didn't want** to apply for it $_{FS,ZX}$ , but the principal called me during the summer months and said , " Sandra the time is running out , you need to apply ". |

The correct answer is *italicized*.
Blue = FS-MTL Yellow = Z&X-STL Green = Both

22

## Holders: FS-MTL is Correct, Z&X-STL Wrong

| | |
|---|---|
| 1 | Yoshihisa Murasawa , a management consultant for Booz-Allen & Hamilton Japan Inc. , said *his firm* $_{FS,ZX}$ will likely be recommending acquisitions of Japanese companies more $_{ZX}$ often to foreign clients in the future . |
| 2 | *The source* $_{FS}$ , interviewed by Interfax in Grozny , **expressed confidence** that that the command of the Russian forces in Chechnya would soon " be able to obtain documentary confirmation " that Khattab was dead . |
| 3 | *The Commonwealth team earlier* this week $_{FS}$ **said** that " the conditions in Zimbabwe did not adequately allow the free and fair expression of will by the electorate ". |
| 4 | Publishing such biased reports will only create **mistrust** among nations $_{FS}$ regarding the objectives and independence of the UN Commission on Human Rights . |
| 5 | The Inkatha Freedom Party , Democratic Alliance , New National Party , African Christian Democratic Party , the Pan Africanist Congress and the United Christian Democratic Party $_{ZX}$ had disagreed with the ANC $_{FS}$ **conclusion** . |
| 6 | The Nigerian leader , President Olusegun Obasanjo $_{ZX}$ , had urged the *minister* $_{FS,ZX}$ not to attack Blair frontally over Britain 's negative position regarding Zimbabwe , but to deal [...] . |
| 7 | US diplomats $_{ZX}$ say *Bush* $_{FS,ZX}$ **will seek to support** Kim 's Nobel Prize winning policy by offering new talks with the North , while remaining firm about North Korea 's missile sales and its feared chemical and biological weapons programmes. |

The correct answer is *italicized*.
Blue = FS-MTL Yellow = Z&X-STL Green = Both

23

## Targets: FS-MTL is Correct, Z&X-STL Wrong

| | |
|---|---|
| 1 | In most cases he **described** *the legal punishments* $_{FS}$ *like floggings and executions of murderers and major drug traffickers that are applied based on the Shria , or Islamic law as human rights violations* . |
| 2 | In another **verbal attack** Kharazi accused *the United States* $_{FS}$ of wanting to exercise " world dictatorship " since the " horrible attacks " of September 11 . |
| 3 | He said those who thought the election process would be rigged were supporters of the MDC party , adding that they **were prejudging** and wanted to direct the process $_{ZX}$ . |
| 4 | However , the fact that certain countries have a more balanced view of the conflict $_{ZX}$ is not the only reason to doubt that *anti-Israeli decisions* $_{FS}$ will , in fact , **be adopted** . |
| 5 | But *his tough stand on P'yongyang* $_{FS}$ **has provoked concern** in Seoul $_{ZX}$ , where President Kim Tae-chung , who is in the last year of his five-year term , has been trying to prise the hermit state out of isolation . |

The correct answer is *italicized*.
Blue = FS-MTL Yellow = Z&X-STL Green = Both

24

## Conclusions

- Multi-task learning can be an effective strategy for improving performance for a task when small amounts of task-specific gold data are available but there exists a related task with substantially more gold data.
- Their analysis found that long-distance dependencies remain challenging for opinion extraction.
- Reminder: these models were *given* an opinion expression and then extracted its role fillers. But finding the opinion expressions is also a key part of the task!
- Opinion extraction is a challenging IE task. Boundaries are tough (for humans & NLP systems!), and people bring a lot of knowledge to bear when identifying opinions.

25

## Temporal Information Extraction

- Time expressions and temporal relations are important to identify for many (most!) applications and domains.
- Temporal IE is especially relevant for event-related applications, such as event detection, event extraction, and event timeline construction.
- Most current work focuses on events in the past. Identifying future references poses additional types of challenges.
- Temporal recognition is also a basic element of language understanding!

1

## Temporal Understanding Examples

**Question**: *When did airlines as a group last raise fares?*

***Last week**, Delta boosted thousands of fares by $10 per round trip, and most big network rivals immediately matched the increase.*

*A fare increase initiated last week by United Airlines was matched by competitors over **the weekend**, marking the second fare increase in two weeks.*

*On Monday, American Airlines raised fares on all domestic flights in the United States. Other airlines were slow to follow suit, but **two weeks later** also raised their prices.*

*On Monday, American Airlines raised fares on all domestic flights in the United States. Other airlines have not yet matched the fare increases.*

2

## Types of Temporal Expressions

| ***Absolute*** | ***Relative*** | ***Durations*** |
|---|---|---|
| 11/02/17 | yesterday | 5 days |
| 2020 | tonight | 2 months |
| May 4 | in the morning | 10 minutes |
| May | last month | 10:00-11:00 |
| 11:00pm | 2 days ago | until midnight |
| 11 o'clock | next quarter | a few years |
| 23:59 MDT | in an hour | for an hour |
| Noon | before noon | in the winter |
| Christmas | my birthday night | dawn to dusk |
| Christmas eve | when Xmas is over | January-April |
| Pioneer Day | after the election | before 2020 |
| 9/11 | soon | this semester |

3

## Lexical Triggers for Time Expressions

**Nominal Nouns:** *morning, noon, night, winter, dusk, dawn, eve, hour, minute, sunrise, sunset*

**Proper Nouns:** *January, Monday, New Year's Eve, Labor Day, Easter, Passover, Ramadan*

**Adjectives/Prefixes:** *recent, last, next, annual, early, late, mid*

**Adverbs:** *hourly, daily, monthly, yearly, annually*

Most time expression recognizers identify trigger words and look at context around them.

Numeric "shape" patterns are needed too (e.g., 3/14/2017)

4

## General Approaches

**Rule-based Systems:** finite-state automata can be built to recognize patterns for temporal expressions. Cascaded finite-state machines typically start with simple expressions and build patterns of greater complexity.

**Sequential Classifiers:** given annotated training data, sequential taggers can label time expressions using BIO tags.

**Constituent-based Classifiers:** given annotated training data, a constituent-based classifier can learn to label syntactic constituents (e.g., chunks or parse tree nodes).

Pro: *boundary issues are separate syntactic problem*
Con: *time expressions must align with syntactic constituents*.

5

## Common Classifier Features for Time Expressions

- Lexical token
- Tokens in local context window
- Part-of-speech tags for target and window tokens
- Syntactic chunk/phrase type for target and window tokens
- Temporal keywords ("lexical triggers"), including days of the week, months, holidays, and general time words (e.g., "*morning*")
- Character-based "shape" features, for example:

| | |
|---|---|
| ##/##/#### | 03/14/2017 |
| ##-##-#### | 03-14-2017 |
| ##:## | 11:50 |
| ##:##(am,pm) | 11:50am |
| ##:##(am,pm)(time-zone) | 11:50am (EDT) |

6

## False Hits

Even strong time keywords can produce false hits for a variety of reasons, such as:

- Film, TV, book, and song titles, such as:
  - *Any Given Sunday* (Oliver Stone film), *48 Hours* (TV series), *1984* (Orwell), *Monday Monday* (Mamas & Papas)
- Organization names, such as:
  - *Black September* (terrorist group), *Tuesday Morning* (discount store)
- People names, such as:
  - *April, May, June* are common female names
- Word sense (and part-of-speech) ambiguity
  - *March, May*

7


8

## Temporal Normalization

- For real applications, time expressions need to be anchored to real dates (e.g., *Tuesday* → 3/29/2022) and mapped into a standardized format.
- Often, this requires extracting the dateline of a document (e.g., news article) or identifying its publication date.
- Th normalization process can be complex and typically relies on hand-coded rules. Some examples:
  - "*last Tuesday*"
  - "*this weekend*"
  - "*3 weeks ago*"
  - "*in a week*"
  - "*on Thanksgiving*" (note: the date changes each year!)

9

## Time Normalization Format Examples

Some normalization formats for fully specified dates:

| Unit | Pattern | Sample Value |
|---|---|---|
| Fully specified dates | YYYY-MM-DD | 1991-09-28 |
| Weeks | YYYY-Wnn | 2007-W27 |
| Weekends | PnWE | P1WE |
| 24-hour clock times | HH:MM:SS | 11:13:45 |
| Dates and times | YYYY-MM-DDTHH:MM:SS | 1991-09-28T11:00:00 |
| Financial quarters | Qn | 1999-Q3 |

And … dates are not always fully specified, so standards need to include partially specified dates too. For example:

- partial date *1995*
- open-ended ranges, such as before/after a specific date
  - *- 08 JAN 90* *08 JAN 90 -*

10

## Temporal Relations

For many event-related applications, we want to know the temporal relationship between events (i.e., the ordering of events relative to each other).

Time expressions are usually present for only some of the events. Recognizing temporal relations between events allows for a relative timeline to be constructed.

The **TimeBank** corpus is a widely used resource that contains annotated temporal expressions, event mentions, and *Temporal Links (TLINKs)* between events and temporal expressions.

**TimeML** is a mark-up language for temporal information.

11

## TimeBank 1.2

- 183 news articles annotated with time expressions, event mentions, and temporal links (and some other things too)
- Thirteen types of temporal links between events:
  - **after/before**
  - **includes/is_included**
  - **during/simultaneous** (for events/states that persist)
  - **iafter/ibefore** (*immediately* after or before)
  - **identity** (coreference)
  - **begins/ends** (one event is beginning/end of another)
  - **begun_by/ended_by** (inverse relation to begins/ends)

12

## (Simplified) TimeBank Annotation Example

*The Russian airline Aeroflot has been <E1: **hit**> with a writ for loss and damages, <E2: **filed**> in Hong Kong by the families of seven passengers <E3: **killed**> in an air <E4: **crash**>.*

*All 75 people <E7-state: **on board**> the Aeroflot Airbus <E5: **died**> when it <E6: **ploughed**> into a Siberian mountain in <T1: **March 1994**>.*

TLINKS:
(**is_included** E6:**ploughed** T1:**March 1994**)
(**before** E2:**filed** E1:**hit**)
(**before** E3:**killed** E2:**filed**)
(**is_included** E3:**killed** E4:**crash**)
(**includes** E7:**on board** E5:**died**)
(**iafter** E5:**died** E6:**ploughed**)
(**identity** E6:**ploughed** E4:**crash**)

13

## TIPSem Temporal IE System

- TIPSem [Llorens et al., 2010] is a set of temporal IE systems that performed well in TempEval-2 tasks. (TIPSem = **T**emporal **I**nformation **P**rocessing based on **Sem**antic information)
- They used similar sequential tagging (CRF) models across six different temporal tasks.
- An unusual emphasis of their work is the incorporation of semantic information.
- They created systems for both English and Spanish, demonstrating the generality of their approach.

14

## Task A: Recognizing Temporal Expressions

Task A is to recognize time expressions as defined by the TimeML TIMEX3 tag, as well as "type" and "value" attributes.

types = {DATE, TIME, DURATION, SET}
The possible values depend on the type.

Examples:

<TIMEX3 tid="t1" type="DATE" value="1999-SU">
***the summer of 1999*** </TIMEX3>

<TIMEX3 tid="t2" type="TIME" value="T24:00">
***twelve o'clock midnight*** </TIMEX3>

<TIMEX3 tid="t3" type="DURATION" value="P2D">
***two entire days*** </TIMEX3>

<TIMEX3 tid="t4" type="SET" value="XXXX-10">
***every October*** </TIMEX3>

15

## Task B: Recognizing Events

- Task B is to recognize and classify events, as defined by the TimeML EVENT tag.
- Events can be expressed as verbs, nominalizations, adjectives, predicative clauses, or PPs (but only the heads are annotated). Events also have 4 types of information:

**Polarity** captures negation (e.g., for events that did not happen)

**Tense** captures temporal verb forms (e.g., past, present and future)

**Aspect** captures verbal information about how events & states extend over time. In English: neutral, progressive, perfect, progressive perfect, and (in past tense) habitual. For example: "*I lose*" vs. "*I am losing*" vs. "*I have been losing*".

**Modality** captures ability, possibility, permission or obligation as indicated by modal verbs (e.g., *could, should, must*).

16

## Example of Event Annotations

*Five days after he <EVENT eid="e1" class="OCCURRENCE"> **came** </EVENT> back …*

*A major <EVENT eid=e2" classes="OCCURRENCE"> **earthquake** </EVENT> in Indonesia ...*

*After many months of renewed <EVENT eid="e3" class=STATE"> **hostility** </EVENT> ...*

*The <EVENT eid="e4" class="OCCURRENCE"> **attack** </EVENT> was not <EVENT eid="e5" classes="I_STATE"> **expected** </EVENT> at all …*

The full annotations would include polarity, tense, aspect, and modality as well.

17

## Tasks C-F: Temporal Relation Links

**Task C**: Determine the temporal relation between an event and a time expression in the same sentence, where the event syntactically dominates the time expression or they occur in the same NP.

**Task D**: Determine the temporal relation between an event and the ***document creation time (DCT).***

**Task E**: Determine the temporal relation between two main events in consecutive sentences.

**Task F**: Determine the temporal relation between two events where one event syntactically dominates the other event. For example: "*she heard an explosion*" or "*he said they postponed the meeting*".

18

## Types of Relation Links

Temporal relations have 6 relation types: **Before, After, Overlap, Before-or-Overlap, Overlap-or-After**, or **Vague**.

**Examples**:

*Mary taught*$_{e1}$ *on Tuesday*$_{t1}$. → Overlap(e1, t1)

*The country defaulted*$_{e2}$ *on debts for that entire year* → Before(e2, DCT)

*The students heard*$_{e1}$ *a fire alarm*$_{e2}$ → Overlap(e1,e2)

*He said*$_{e1}$ *they had postponed*$_{e2}$ *the meeting* → After(e1,e2)

19

## Classification Models

- TIPSem systems are CRF models for sequential tagging trained with supervised learning.
- For Tasks A and B, the input is a word sequence and the output is BIO labels on the words.
- For Tasks C-F, the input is instances of the classes (e.g., TIMEX3 and EVENT instances) and the output is relation links.
- In addition to traditional features, they use information from a semantic role labeling (SRL) system. For example:

*John*$_{AGENT}$ *sold*$_{VERB}$ *Mary*$_{RECIPIENT}$ *a car*$_{THEME}$

20

## TipSem Architecture

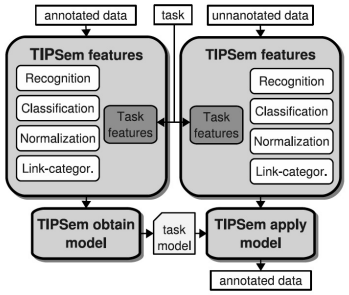
They grouped the tasks into 4 categories, which each use different types of feature sets:

**Recognition**: TIMEX3 and EVENT bounding

**Classification**: TIMEX3 types and EVENT classes

**Normalization**: TIMEX3 values

**Link Categorization**: Tasks C to F




21

## General Features

- **Morphological**: lemmas and POS tags, for context windows of size +/- 2.
- **Syntactic**: Phrase level syntactic information from parse trees produced by constituency parsers.
- **Polarity, tense, and aspect**: hand-crafted rules that use POS tags

For example: will + VERB → FUTURE tense

22

## Semantic Features

- **Role:** the semantic role based on the verb it depends on in the parse. The CCG SRL tool was used for English, and AnCora for Spanish.
- **Governing Verb:** the verb that the token depends on in the parse.
- **Role+Verb combination:** governing verb paired with role.
- **Role configuration:** for verbs that head a sentence or nested sentence, the set of roles that depend on the verb are captured.
- **Lexical semantics:** the top 4 semantic classes for a word, based on WordNet for English and EuroWordNet for Spanish.

23

## Link Categorization Features

**Head preposition**, if TIMEX or EVENT in a PP.

**Syntactic relation** of the TIMEX and EVENT: same sentence, subsentence, or subphrase.

**Time position:** if EVENT is not directly linked to a TIMEX, then is it before, after, or overlapping with the TIMEX.

**Interval**, if there is an interval indicator near the TIMEX.

**TIMEX type.**

**Semantic roles** if the EVENT or TIMEX are labeled with a temporal role. For example: *After$_{TEMP}$ he left$_{e1}$ home …*

(NOTE: Different tasks used different combinations of the features.)

24

## Evaluation

- The evaluation used 17K words for English and 10K words for Spanish: half for Tasks A & B and half for Tasks C–F.
- Results are based on tokens for Tasks A & B, and links for Tasks C-F. *EN* = English, *ES* = Spanish.
- TipSem-B versions *exclude* the semantic features.

TIMEX3 detection

| System *lang* | Prec. | Rec. | $F_{\beta=1}$ | type | value |
|---|---|---|---|---|---|
| TIPSem *EN* | **0.92** | 0.80 | 0.85 | 0.92 | 0.65 |
| TIPSem *ES* | 0.95 | **0.87** | **0.91** | **0.91** | 0.78 |
| TIPSem-B *EN* | 0.88 | 0.60 | 0.71 | 0.88 | 0.59 |
| TIPSem-B *ES* | **0.97** | 0.81 | 0.88 | 0.99 | 0.75 |

Table 1: Task A - English and Spanish

25

## Results for Task B

Task B is recognizing and classifying event words.

| System *lang* | Prec. | Recall | $F_{\beta=1}$ | class |
|---|---|---|---|---|
| TIPSem *EN* | 0.81 | 0.86 | **0.83** | **0.79** |
| TIPSem *ES* | 0.90 | **0.86** | **0.88** | **0.66** |
| TIPSem-B *EN* | **0.83** | 0.81 | 0.82 | 0.79 |
| TIPSem-B *ES* | **0.92** | 0.85 | 0.88 | 0.66 |

Table 2: Task B - English and Spanish

26

## Results for Tasks C & D

*Same sentence task*

| System *lang* | $F_{\beta=1}$ |
|---|---|
| TIPSem *EN* | 0.55 |
| TIPSem *ES* | **0.81** |
| TIPSem-B *EN* | 0.54 |
| TIPSem-B *ES* | 0.81 |

Table 3: Task C - English and Spanish

*DCT task*

| System *lang* | $F_{\beta=1}$ |
|---|---|
| TIPSem *EN* | **0.82** |
| TIPSem *ES* | **0.59** |
| TIPSem-B *EN* | 0.81 |
| TIPSem-B *ES* | 0.59 |

Table 4: Task D - English and Spanish

27

## Results for Tasks E & F

Events in consecutive sentences

| System *lang* | $F_{\beta=1}$ |
|---|---|
| TIPSem *EN* | 0.55 |
| TIPSem-B *EN* | 0.55 |

Table 5: Task E - English

Events in syntactic relation

| System *lang* | $F_{\beta=1}$ |
|---|---|
| TIPSem *EN* | 0.59 |
| TIPSem-B *EN* | 0.60 |

Table 6: Task F - English

28

## On-Going and Future Events

From [Huang et al., EMNLP 2016]:

(1) *The metro workers' strike in Bucharest has entered the fifth day.* (On-Going)

(2) *BBC unions demand more talks amid threat of new strikes.* (Future)

(3) *Pro-reform groups have called for nationwide protests on polling day.* (Future)

Events often are captured by nouns, in which case future tense markers do not apply!

29

## On-going and Future Events

**On-going:** An event that has started and is still in progress or likely to resume[2] in the immediate future. There should be no reason to believe that it has ended.

**Future Planned:** An event that has not yet started, but a person or group has planned for or explicitly committed to an instance of the event in the future. There should be near certainty it will happen.

**Future Alert:** An event that has not yet started, but a person or group has been threatening, warning, or advocating for a future instance of the event.

**Future Possible:** An event that has not yet started, but the context suggests that its occurrence is a live possibility (e.g., it is anticipated, feared, hinted at, or is mentioned conditionally).

30

## Examples

**On-going**

[EN] Negotiations continue with no end in sight for the 2 week old *strike*.
Yesterday's rallies have caused police to fear more today.

**Future Planned**

[EN] 77 percent of German steelworkers voted to *strike* to raise their wages.
Peace groups have already started organizing mass *protests* in Sydney.

**Future Alert**

[EN] Farmers have threatened to hold *demonstrations* on Monday.
Nurses are warning they intend to walkout if conditions don't improve.

**Future Possible**

[EN] Residents fear *riots* if the policeman who killed the boy is acquitted.
The military is preparing for possible protests at the G8 summit.

31

## There are also many types of future-oriented verbs, such as:

| |
|---|
| **Commitment:** threaten, vow, promise, pledge, commit, declare, claim, volunteer, anticipate |
| **Coming_to_be:** enter, emerge, plunge, kick, mount reach, edge, soar, promote, increase, climb, double |
| **Purpose:** plan, intend, project, aim, object, target |
| **Permitting:** allow, permit, approve, subpoena |
| **Experiencer_subj:** fear, scare, hate |
| **Waiting:** expect, wait |
| **Scheduling:** arrange, schedule |
| **Deciding:** decide, opt, elect, pick, select, settle |
| **Request:** ask, urge, order, encourage, demand, appeal, request, summon, implore, advise, invite |
| **Evoking:** raise, press, back, recall, pressure, force, rush, pull, drag, respond |

32

## Summary

- Extracting temporal information is extremely important for many applications, but can be deceptively challenging.
- Recognizing temporal expressions is not trivial, but IE systems can achieve high recall and precision.
- However, linking times to events, and understanding the relative ordering of events, is much more complex.
- TimeBank is just one proposed representation -- conceptual challenges remain as to the best way to represent both time and events.
- Parallel challenges include identifying events and states in text, subevents, and event coreference resolution.

33

## Open Information Extraction

- Traditional *entity* and *relation extraction* systems learn how to identify instances of a specific entity class or relation, usually given labeled examples.
- In contrast, the goal of ***open information extraction (OpenIE)*** systems is to extract instances of any semantic class or relation (i.e., an "open" set of classes/relations).

  Essentially, extract everything you can find!
- Open IE systems typically learn from the Web, benefitting from the Web's vast amount of text.

  → need shallow methods, for robustness and speed

1

## Open Information Extraction Paradigm

[Etzioni et al, 2011] explained the Open Information Extraction (Open IE) paradigm as:

*[Open IE] eschews hand-labeled training examples, and avoids domain-specific verbs and nouns, to develop* ***unlexicalized, domain-independent extractors*** *that scale to the Web corpus.*

*Open IE systems avoid specific nouns and verbs at all costs. The extractors are unlexicalized – formulated only in terms of syntactic tokens (e.g., part-of-speech tags) and closed-word classes (e.g., of, in, such as). Thus, Open IE extractors focus on generic ways in which relationships are expressed in English – naturally generalizing across domains.*

2

## Traditional Relation Extraction vs. Open IE

- Since no specific relation is targeted, the syntactic/superficial forms of relational phrases must be identified.
- The relation phrases then need to be clustered or normalized to determine which instances represent the same relation. For example:

  {*is headquartered in*, *is based in*} → Headquarters Relation

Very general feature sets are needed to:

1. cover an unlimited and unknown set of relations. Even anchoring on Named Entities is not ok as it would be a problem for many relations.
2. robustly and efficiently process large amounts of Web text.

3

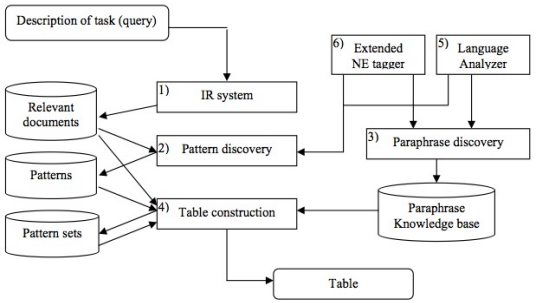
## Open IE Research

- Several research efforts have focused on developing Open IE systems to automatically acquire knowledge from the Web. Extracting information from a set of texts is sometimes called *Machine Reading (or Macro Reading).*
- The goals of Open IE are to create systems that can:
  - robustly process and extract knowledge from massive amounts of Web text.
  - populate and organize the extracted information in large knowledge bases.
  - develop methods that can continually harvest new knowledge, both to acquire new facts that emerge and to enable new types of relations to be identified.

4

## Prior Work: On-Demand Information Extraction

[Sekine 2006] proposed **"on-demand" information extraction**, where a user would provide a query for a desired relation and the system would automatically learn paraphrases and build a table of extracted information.




5

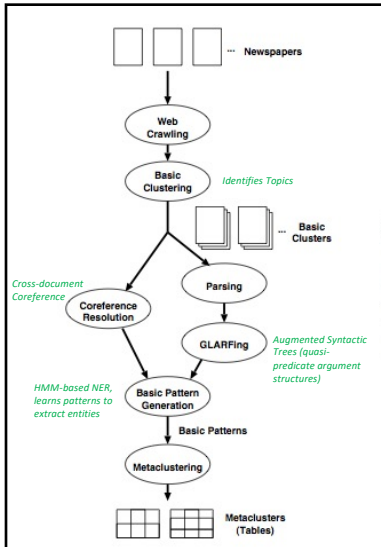
## Preemptive Information Extraction

- [Shinyama and Sekine, 2006] explored the idea of **preemptive information extraction** and proposed:

  "*a technique called Unrestricted Relation Discovery that discovers all possible relations from texts and presents them as tables.*"

- Their system used clustering, pattern learning, and meta-clustering to build a set of tables filled with information extracted for different relations, without training data.
- This work was among the earliest research similar to Open IE, and a preliminary system was built. But the effort did not continue on a large scale. However, similar efforts were undertaken by other research groups on a much larger scale…

6




Output Results

| | |
|---|---|
| Source articles | 28,009 |
| Basic clusters | 5,543 |
| Basic patterns (token) | 643,767 |
| Basic patterns (type) | 7,990 |
| Metaclusters | 302 |
| Metaclusters (rows ≥ 3) | 101 |

Nominations Table

| Article | 1:confirm | 2:be-confirmed |
|---|---|---|
| 2005-09-21 | Senate | Roberts |
| 2005-10-03 | Supreme Court | Miers |
| 2005-10-20 | Senate | Bush |
| 2005-10-26 | Senate | Sauerbrey |
| 2005-10-31 | Senate | Mr. Alito |
| 2005-11-04 | Senate | Alito |
| 2005-11-17 | Fed | Bernanke |

Hurricanes Table

| Hurricane | Date (Affected Place) | Articles |
|---|---|---|
| Philippe | Sep 17-20 (?) | 6 |
| * Rita | Sep 17-26 (Louisiana, Texas, etc.) | 566 |
| * Stan | Oct 1-5 (Mexico, Nicaragua, etc.) | 83 |
| * Tammy | Oct 5-? (Georgia, Alabama) | 18 |
| Vince | Oct 8-11 (Portugal, Spain) | 12 |
| * Wilma | Oct 15-25 (Cuba, Honduras, etc.) | 368 |
| Alpha | Oct 22-24 (Haiti, Dominican Rep.) | 80 |
| * Beta | Oct 26-31 (Nicaragua, Honduras) | 55 |
| * Gamma | Nov 13-20 (Belize, etc.) | 36 |

7

## KnowItAll

- A research group at the University of Washington began an OpenIE research project called **KnowItAll**, which produced a steady stream of research results related to open information extraction for many years.
- The emphasis of this project has been massive Web-scale IE, with an emphasis on speed and extracting large volumes of information.
- Consequently, many of the methods use very shallow pattern matching and few NLP tools.
- The original KnowItAll system used Hearst's hyponym patterns to identify relation instances in an iterative learning framework.

**Hyponym Patterns**: <x> such as <y>, <x> including <y>, etc.

8

## UW's Open IE Research Efforts

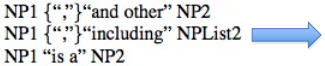
Following KnowItAll, UW embarked on a long-term research effort focused on Open IE from the Web and it continued to evolve.

- **ReVerb**: identifies and extracts unspecified binary relations.
- **RESOLVER**: a probabilistic relational model for determining whether two relation expressions are "synonymous".
- **TextRunner** generates labeled examples using heuristics and trains a classifier for unrestricted relation extraction. RESOLVER is incorporated to identify synonymous relation phrases.
- **SHERLOCK**: learns first-order Horn Clauses as inference rules. For example: Contains(Food, Chemical) :- IsMadeFrom(Food, Ingredient) ∧ Contains(Ingredient, Chemical);

9

## KnowItAll Rule Examples [Etzioni et al., 2004]

NP1 {","} "such as" NPList2
NP1 {","}"and other" NP2
NP1 {","}"including" NPList2
NP1 "is a" NP2
NP1 "is the" NP2 "of" NP3
"the" NP1 "of" NP2 "is" NP3



```
Extraction Rule:

  NP1 "such as" NPList2
  & head(NP1)="countries"
  & properNoun(head(each(NPList2)))
  =>
  instanceOf(Country,head(each(NPList2)))
  keywords: "countries such as"
```

```
Extraction Rule for a Binary Relation:

  NP1 "plays for" NP2
  & properNoun(head(NP1))
  & head(NP2)="Seattle Mariners"
  =>
  instanceOf(Athlete,head(NP1))
  & instanceOf(SportsTeam,head(NP2))
  & playsFor(head(NP1),head(NP2))
  keywords: "plays for", "Seattle Mariners"
```

10

## Tradeoffs Between Open and Traditional Relation Extraction [Banko & Etzioni, ACL 2008]

- "Open IE is a *relation-independent* extraction paradigm that is tailored to massive and heterogeneous corpora such as the Web."
- "An Open IE system extracts a diverse set of relational tuples from text without *any* relation-specific input."
- In contrast, traditional relation extraction systems begin with training examples for a specific type of relation and learn to identify instances of that type of relation.
  - Each relation requires its own training process.

11

## Why is Open IE more difficult?

1. Open IE has to identify both the set of entities that participate in a relation as well as the textual clues that reveal the relation.
2. A relation-independent process means that relation-specific features cannot be exploited.
3. Many relations can have a wide variety of argument types, so anchoring the relation with named entities is not sufficient. Even the general types of arguments are not constrained in any way.

12

## Common Syntactic Patterns

- 500 randomly sampled sentences were reviewed to manually identify the types of constructions that captured a relation expression.
- 95% of the identified patterns could be grouped into 8 lexico-syntactic categories.
- While these patterns are not sufficient to identify a relation, these results suggest that most relation expressions can be captured by a small set of pattern templates.

13

## Common Lexico-syntactic Patterns

95% of the 500 sampled sentences had relation expressions matching one of these patterns.

| Relative Frequency | Category | Simplified Lexico-Syntactic Pattern |
|---|---|---|
| 37.8 | Verb | $E_1$ Verb $E_2$<br>*X established Y* |
| 22.8 | Noun+Prep | $E_1$ NP Prep $E_2$<br>*X settlement with Y* |
| 16.0 | Verb+Prep | $E_1$ Verb Prep $E_2$<br>*X moved to Y* |
| 9.4 | Infinitive | $E_1$ to Verb $E_2$<br>*X plans to acquire Y* |
| 5.2 | Modifier | $E_1$ Verb $E_2$ Noun<br>*X is Y winner* |
| 1.8 | $Coordinate_n$ | $E_1$ (and\|,\|-\|:) $E_2$ NP<br>*X-Y deal* |
| 1.0 | $Coordinate_v$ | $E_1$ (and\|,) $E_2$ Verb<br>*X , Y merge* |
| 0.8 | Appositive | $E_1$ NP (:\|,)? $E_2$<br>*X hometown : Y* |

14

## Seed Labeled Data

Relation-independent heuristics were applied to the Penn Treebank to obtain labeled relation instances, which were designed to approximate syntactic dependencies and semantic roles.

NO parsing or semantic analysis! Just NP chunking and POS tags.

For example:

**Class**: + **Heuristic**: Subject,Verb,Object (SVO) Triple

**Example**: "*<Einstein> received <the Nobel Prize>*"

**Class**: - **Heuristic**: ADVP crossing

**Example**: "*He studied <Einstein's work> when visiting <Germany>.*"

15

## O-CRF

- Labeled instances for training are generated heuristically.
- A sequential tagging model (CRF) is trained to label tokens that express a binary relation using IOB tags.
- A noun phrase chunker is applied and all NPs pairs within a certain distance from each other are candidates for a relation instance.
- The feature set includes POS tags, regular expressions to detect things like capitalization and punctuation, context words, and conjunctions of features for adjacent positions in a context window of size +/- 6 words.
- Context words are only captured for closed class words and not for open class words! Presumably for improved generality.

16

## Relation Extraction as Sequence Labeling

Each relation must be anchored by two noun phrases, which are called "entities" (ENT).

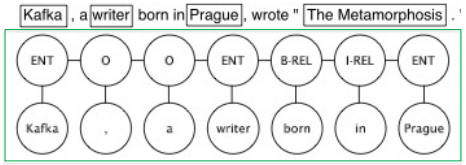



Figure 1: Relation Extraction as Sequence Labeling: A CRF is used to identify the relationship, *born in*, between *Kafka* and *Prague*

17

## O-CRF's Limitations

- Relations can only be identified if they are explicitly mentioned in a text.
- Relations can only be identified through lexical context. Document style features are not considered.
- Relations can only be identified between NPs within the same sentence.
- O-CRF does not cluster/normalize relations.
  - Relation "synonyms" (paraphrases) were identified by a different system called RESOLVER [Yates and Etzioni, 2007].

18

## Evaluation of Open Relation Extraction

The first evaluation compares the performance of O-CRF with the TextRunner Open IE system (O-NB). TextRunner had extracted 7.5 million tuples from 9 million Web pages.

Both systems were tested on 500 sentences .

| | O-CRF | | | O-NB | | |
|---|---|---|---|---|---|---|
| **Category** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| Verb | 93.9 | 65.1 | 76.9 | 100 | 38.6 | 55.7 |
| Noun+Prep | 89.1 | 36.0 | 51.3 | 100 | 9.7 | 55.7 |
| Verb+Prep | 95.2 | 50.0 | 65.6 | 95.2 | 25.3 | 40.0 |
| Infinitive | 95.7 | 46.8 | 62.9 | 100 | 25.5 | 40.6 |
| Other | 0 | 0 | 0 | 0 | 0 | 0 |
| **All** | **88.3** | **45.2** | **59.8** | 86.6 | 23.2 | 36.6 |

19

## Relation-Specific Extraction

- For comparison, a traditional relation extraction system was trained with a CRF model, which they called R1-CRF.
- R1-CRF is identical to O-CRF except:
  - R1-CRF was trained from manually labeled positive and negative instances of a specific relation R.
  - R1-CRF used both closed-class and open-class words as features. (O-CRF could only use closed-class words.)
  - No additional steps are needed to identify the relation type, since it is trained to identify only instances of relation R.

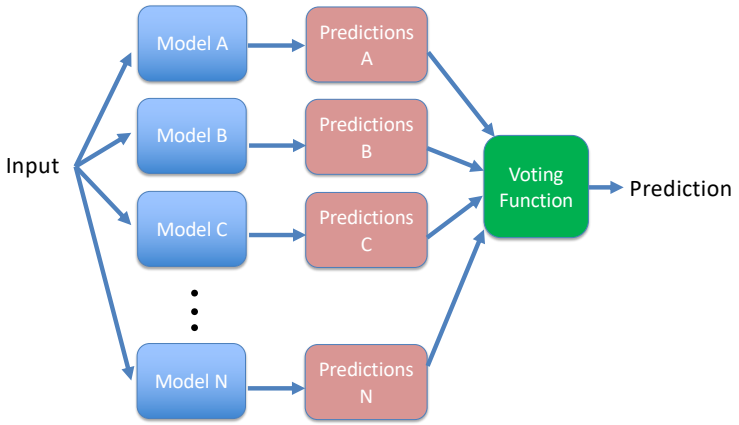
20

## Ensembles

- **Ensemble methods** are widely used in NLP and typically yield better performance than individual systems.
- An **ensemble** is a set of different systems (models) that perform the same task. Ensemble methods consider the output of all the models to make a final decision.
- Typically, **Voting** methods decide on a class label based on the set of labels produced by the models, usually majority vote or a minimum number of votes received.
- Other functions are possible too, such as ranking labels based on confidence values.

21

## Ensemble Architecture



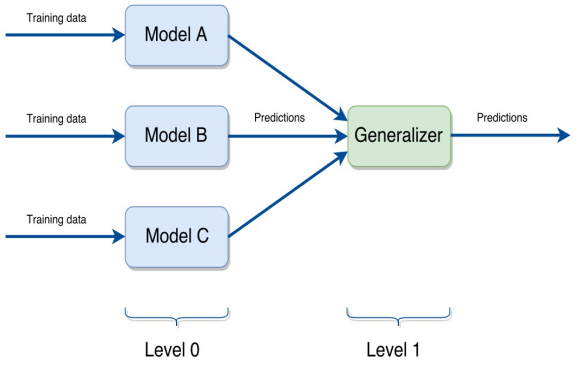

22

## Stacking Learning

- **Stacked learning** trains a "meta-classifier" that learns how to weight or combine the output values of multiple individual systems to make better decisions.
- With voting-based ensembles, determining a minimum number of votes can be tricky. And as the number of components grows, the threshold may need to change.
- Stacked learning automatically learns how to best use the outputs of the component classifiers, so can simply be retrained if new components are added.
- Many types of features can be easily incorporated in the meta-classifier.

23

## Stacked Ensemble Architecture




From [Divina et al., 2018]

24

## A Stacked Relation Extractor (H-CRF)

- A **hybrid relation extraction system (H-CRF)** is created using a stacking framework.
- The O-CRF and R1-CRF classifiers are the individual components.
- The H-CRF meta-classifier also uses a CRF sequential tagging model for learning.
- The H-CRF's feature set includes:
  - probability estimates for the O-CRF and R-CRF's labels
  - an edit distance measure between the predicted relations
  - a feature indicating whether either model returned No Relation
  - lexical and POS terms between the two candidate NPs

25

## Comparison for Known Relation Extraction

- The performance of O-CRF and R1-CRF was then compared for specific relations.
- Labeled data was acquired for 4 relations: corporate acquisitions, birthplaces, product inventions, and award winners. The data was divided into training and test sets.
- For each relation, R1-CRF was trained using the labeled training set. Both models were then evaluated on the test set.
- Recall and Precision were measured on the relation tuples that were generated by each system.

26

## Evaluation Results for Known Relations

| | O-CRF | | R1-CRF | | |
|---|---|---|---|---|---|
| **Relation** | **P** | **R** | **P** | **R** | **Train Ex** |
| Acquisition | 75.6 | 19.5 | 67.6 | 69.2 | 3042 |
| Birthplace | 90.6 | 31.1 | 92.3 | 64.4 | 1853 |
| InventorOf | 88.0 | 17.5 | 81.3 | 50.8 | 682 |
| WonAward | 62.5 | 15.3 | 73.6 | 52.8 | 354 |
| **All** | **75.0** | 18.4 | 73.9 | **58.4** | 5930 |

The two systems achieve comparable levels of precision.
But recall is **much** higher for R1-CRF!

However … R1-CRF reqiures labeled training data for the relation, while O-CRF was not trained specifically for this relation.

27

## How Much Training Data is Needed?

So they looked at learning curves to determine how much labeled training data was necessary to achieve roughly the same precision.

| | O-CRF | | R1-CRF | | |
|---|---|---|---|---|---|
| **Relation** | **P** | **R** | **P** | **R** | **Train Ex** |
| Acquisition | 75.6 | 19.5 | 67.6 | 69.2 | 3042* |
| Birthplace | 90.6 | 31.1 | 92.3 | 53.3 | 600 |
| InventorOf | 88.0 | 17.5 | 81.3 | 50.8 | 682* |
| WonAward | 62.5 | 15.3 | 65.4 | 61.1 | 50 |
| **All** | **75.0** | 18.4 | 70.17 | **60.7** | >4374 |

- For the WonAward relation, 50 training examples were needed.
- For the BirthPlace relation, 600 training examples were needed.
- For the Acquisition and InventorOf relations, R1-CRF never achieved comparable precision, even with substantial training data.

28

## Analysis of Results

- R1-CRF benefits a lot from the lexical features.

  Example: "*Yahoo to Acquire Inktomi*"

  *Acquire* is mistagged as a proper noun, so O-CRF is confused. But R1-CRF still recognizes "*acquire*" as a relation trigger.

- O-CRF also failed to recognize synonyms for the relation.

  R1-CRF identified 16.25 synonyms per relation, on average. With RESOLVER, O-CRF found only 6.5 synonyms per relation.

**Conclusions**:

Open IE provides good precision without relation training data.

But when higher recall is needed and manually labeling data is possible, traditional RE is desirable.

29

## Evaluating the Hybrid Extractor

| | R1-CRF | | | Hybrid | | |
|---|---|---|---|---|---|---|
| **Relation** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| Acquisition | 67.6 | 69.2 | 68.4 | 76.0 | 67.5 | 71.5 |
| Birthplace | 93.6 | 64.4 | 76.3 | 96.5 | 62.2 | 75.6 |
| InventorOf | 81.3 | 50.8 | 62.5 | 87.5 | 52.5 | 65.6 |
| WonAward | 73.6 | 52.8 | 61.5 | 75.0 | 50.0 | 60.0 |
| **All** | 73.9 | **58.4** | 65.2 | **79.2** | 56.9 | **66.2** |

- Using both O-CRF and RI-CRF in the stacked ensemble framework produces better precision (79%) than either one alone.
- Recall does not improve, but is nearly as good as the R1-CRF.
- The hybrid approach requires labeled training data for the relation, so the trade-off is manual effort for higher precision.

30

## Never-Ending Language Learning (NELL)

- The NELL effort at Carnegie Mellon University uses semi-supervised learning methods to automatically extract large amounts of knowledge from the Web.
- The NELL project aims to continually acquire knowledge and improve its performance. From [Carlson et al., AAAI 2010]:

*"By a "never- ending language learner" we mean a computer system that runs 24 hours per day, 7 days per week, forever, performing two tasks each day:*

*1. **Reading task**: extract information from web text to further populate a growing knowledge base of structured facts and knowledge.*

*2. **Learning task**: learn to read better each day than the day before, as evidenced by its ability to go back to yesterday's text sources and extract more information more accurately."*

31

## Read the Web

Research Project at Carnegie Mellon University

Home | Project Overview | Resources & Data | Publications | People

NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:


- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

So far, NELL has accumulated over 50 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 2,810,379 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.

32

# Never-Ending Language Learning (NELL)

- NELL consists of an ensemble of extraction methods that can learn:
  - *semantic categories*, such as cities, companies, and teams
  - *relations*, such as HasOfficesIn(Organization, Location)
- NELL includes natural language pattern learners, extractors for semi-structured text (e.g., tables and lists), morphological similarity learners, probabilistic inference rule learning, etc.
- [Carlson et al., 2010] presented a method that trains extractors category and relation extractors using small amounts of labeled data, and applies them to the Web. "Coupling constraints" are defined across extractors to improve accuracy.

33

# Examples of Different Types of Learning

CPL: Semantic Class Learning with Contextual Patterns

| Predicate | Pattern |
|---|---|
| emotion | hearts full of $X$ |
| beverage | cup of aromatic $X$ |
| newspaper | op-ed page of $X$ |
| teamPlaysInLeague | $X$ ranks second in $Y$ |
| bookAuthor | $Y$ classic $X$ |

CSEAL: Web Page Wrapper Induction

| Predicate | Web URL | Extraction Template |
|---|---|---|
| academicField | http://scholendow.ais.msu.edu/student/ScholSearch.Asp | ` [X] -` |
| athlete | http://www.quotes-search.com/d_occupation.aspx?o=+athlete | `<a href='d_author.aspx?a=[X]'>-` |
| bird | http://www.michaelforsberg.com/stock.html | `<option>[X]</option>` |
| bookAuthor | http://lifebehindthecurve.com/ | `</li> <li>[X] by [Y] &#8211;` |

RL: Horn Clauses induced by Rule Learner

| Probability | Consequent | Antecedents |
|---|---|---|
| 0.95 | athletePlaysSport($X$, basketball) | ⇐ athleteInLeague($X$, NBA) |
| 0.91 | teamPlaysInLeague($X$, NHL) | ⇐ teamWonTrophy($X$, Stanley Cup) |
| 0.90 | athleteInLeague($X$, $Y$) | ⇐ athletePlaysForTeam($X$, $Z$), teamPlaysInLeague($Z$, $Y$) |
| 0.88 | cityInState($X$, $Y$) | ⇐ cityCapitalOfState($X$, $Y$), cityInCountry($X$, USA) |
| † 0.62 | newspaperInCity($X$, New York) | ⇐ companyEconomicSector($X$, media), generalizations($X$, blog) |

34

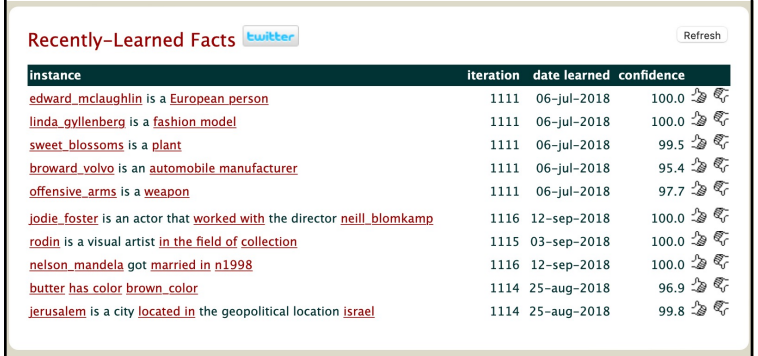
# Examples of Learned Knowledge

| Predicate | Instance | Source(s) |
|---|---|---|
| ethnicGroup | Cubans | CSEAL |
| arthropod | spruce beetles | CPL, CSEAL |
| female | Kate Mara | CPL, CMC |
| sport | BMX bicycling | CSEAL, CMC |
| profession | legal assistants | CPL |
| magazine | Thrasher | CPL |
| bird | Buff-throated Warbler | CSEAL |
| river | Fording River | CPL, CMC |
| mediaType | chemistry books | CPL, CMC |
| cityInState | (troy, Michigan) | CSEAL |
| musicArtistGenre | (Nirvana, Grunge) | CPL |
| tvStationInCity | (WLS-TV, Chicago) | CPL, CSEAL |
| sportUsesEquip | (soccer, balls) | CPL |
| athleteInLeague | (Dan Fouts, NFL) | RL |
| starredIn | (Will Smith, Seven Pounds) | CPL |
| productType | (Acrobat Reader, FILE) | CPL |
| athletePlaysSport | (scott shields, baseball) | RL |
| cityInCountry | (Dublin Airport, Ireland) | CPL |

Table 1: Example beliefs promoted by NELL.

35

# Snapshot from 3/30/2022



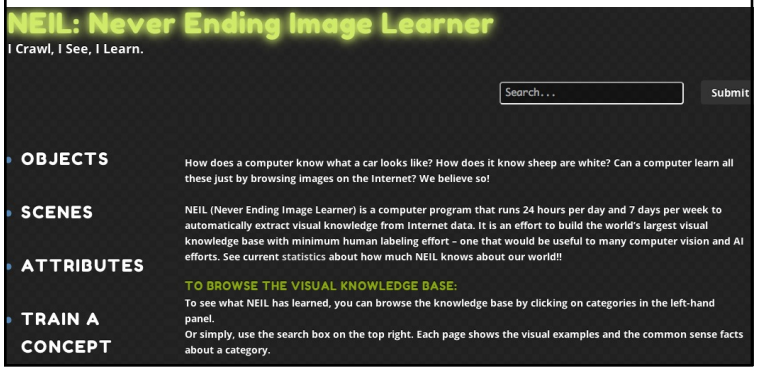

I guess the learning ended in 2018… ;)

36


37

## Never-Ending Image Learner (NEIL)

Interesting Aside: NELL also inspired a follow-on effort at CMU called NEIL to continually extract visual knowledge!




38

## Conclusions

- Open Information Extraction holds great promise for automatically constructing large and rich knowledge bases.
- These efforts have advanced the state-of-the-art for robustly and efficiently extracting large volumes of diverse knowledge from unstructured, often unwieldly Web text.
- However, there is ample room for improvement in the accuracy, organization, and richness of the learned knowledge.
  - *organizing the learned knowledge is a key challenge!*
- Open IE learners tend to learn the most prevalent facts and relations, and are less able to learn less common knowledge or acquire specialized concepts with domain-specific idiosyncracies.

39

# Common Sense Knowledge for NLP

- People rely on a great deal of common sense knowledge when understanding language.
- Most NLP systems are trained for a specific task using the words in the input and (sometimes) additional features. The features typically represent relatively shallow or general information (e.g., proximity, orthography, syntax, or general semantic knowledge).
- Our NLP models need to know the same common sense knowledge that people do in order to:
  - make more intelligent decisions
  - make the same inferences that people do

1

# The Winograd Challenge (Coreference Resolution)

*The city councilmen refused the demonstrators a permit because*
***they*** *feared violence.* → **they** = councilmen
***they*** *advocated violence.* → **they** = demonstrators

*The lawyer asked the witness a question, but* ***he*** *was reluctant to*
*answer it.*
*repeat it.*

*The man couldn't lift his son because* ***he*** *was so* *weak.*
*heavy.*

*I poured water from the bottle into the cup until* ***it*** *was* *full.*
*empty.*

2

# Machine Translation

Machine translation systems often do quite well with word sense disambiguation by relying on the words in the surrounding context.

For example:
*The electrician is working.* (*working* → Labor)
*The telephone is working.* (*working* → Functional)

But these systems often get confused when words associated with different senses are interspersed. For example:

| English original | Google translation |
|---|---|
| The electrician is working. | Der Elektriker **arbeitet**. |
| The electrician that came to fix the telephone is working. | Der Elektriker, die auf das Telefon zu beheben kam **funktioniert**. |
| The telephone is working. | Das Telefon **funktioniert**. |
| The telephone on the desk is working. | Das Telefon auf dem Schreibtisch **arbeitet**. |

3

# Temporal Knowledge

The duration of events is a form of temporal knowledge that relies on common sense knowledge.

*Julie dropped her iPhone on the floor.* (seconds)

*Julie made a sandwich for lunch.* (minutes)

*Julie watched a movie.* (hours)

*Julie went on vacation to Seattle.* (days/weeks)

*Julie took a class on Natural Language Processing.* (months)

*Julie got her computer science degree at the Univ. of Utah.* (years)

4

## Story Understanding

- *Mary went to a restaurant.*
- *George went to the dentist.*
- *Julie finished the watermelon.*
- *Julie finished the book.*
- *Tom ordered a pizza.*
- *Tom ordered a taxi.*
- *The boy had a bone stuck in his throat.*
- *Max needed money to get his car fixed. He called his sister.*
- *The lion spotted an antelope on the hill.*

5

## A Short Story

John got up one morning and discovered his power was out.

Unable to shave, he called his next door neighbor and asked if he could come over to borrow the bathroom. But everyone on the street was out. So John called FG&E and drove to work hoping no one would see him before he found a bathroom with hot water. Unfortunately, he ran into his boss on the elevator. He explained his predicament, but did not feel reassured by Mr. Carver's silence. John stumbled through the rest of the week half-expecting to find a pink slip in his mailbox.

We make **numerous** commonsense inferences when we read!

6

## Common Sense for Visual Understanding

What is this?

What is this?

We use our world knowledge to make inferences about images too!

7

## Plausible Reasoning

- When we communicate, we make **inferences** about things that are very likely to be true, but not guaranteed (***defeasible inferences***).
  - We make assumptions about the most common or typical situations.
  - We usually assume that we are being told the truth and that we are being given complete information (*Gricean maxims*).
  - We often reason based on similar situations that we know about ("case-based reasoning").
- Generating a knowledge base of common sense information can capture the types of default ("prior") knowledge that we assume to be true *in the absence of information to the contrary.*

8

## Transitivity across Relations?

- Commonsense knowledge is often captured in ISA (hypernym/hyponym) hierarchies, and we often assume transitivity for hierarchical relations. For example:

  (dog ISA mammal) & (mammal ISA animal) → (dog ISA animal)

- But this is not *always* true … the world is complicated.

chair IS furniture
car seat IS chair
car seat is NOT furniture!

I fit inside my clothes
My clothes fit inside a drawer.
I do NOT fit inside a drawer.

9

## CYC

- CYC was an ambitious project started in 1984 aimed at manually compiling a massive repository of common sense knowledge. It was originally envisioned as a 10-yr project, but still exists today!
- As of 2012, the public version OpenCyc 4.0 contained 239,000 concepts and over 2 million facts, organized in a taxonomy.
- A larger ResearchCyc is available with a license, and contains 500,000 concepts and 5 million facts.
- Some people have reported using CYC for Web query expansion, question answering, and intelligence analysis.
- As with many large, manually curated KBs, people have complained that is organized poorly and unevenly. How to represent knowledge and organize knowledge is a critical but often overlooked problem!

10

## ConceptNet

**ConceptNet** [Speer et al., AAAI 2017] is a large knowledge graph that connects words with phrases that capture a variety of relations.

ConceptNet was created from a mix of hand-crafted resources, crowd-sourcing, and "games with a purpose". (It is a derivative from the Open Mind Common Sense project.)

- Facts acquired from Open Mind Common Sense (OMCS) (Singh 2002) and sister projects in other languages (Anacleto et al. 2006)
- Information extracted from parsing Wiktionary, in multiple languages, with a custom parser ("Wikiparsec")
- "Games with a purpose" designed to collect common knowledge (von Ahn, Kedia, and Blum 2006) (Nakahara and Yamada 2011) (Kuo et al. 2009)
- Open Multilingual WordNet (Bond and Foster 2013), a linked-data representation of WordNet (Miller et al. 1998) and its parallel projects in multiple languages
- JMDict (Breen 2004), a Japanese-multilingual dictionary
- OpenCyc, a hierarchy of hypernyms provided by Cyc (Lenat and Guha 1989), a system that represents common sense knowledge in predicate logic
- A subset of DBPedia (Auer et al. 2007), a network of facts extracted from Wikipedia infoboxes

11

## ConceptNet's Content

- ConceptNet contains over 21 million edges and 8 million nodes. Its English vocabulary contains 1.5 million nodes.
- ConceptNet includes 36 core relations.
  - **Symmetric relations**: *Antonym*, *DistinctFrom*, *EtymologicallyRelatedTo*, *LocatedNear*, *RelatedTo*, *SimilarTo*, and *Synonym*
  - **Asymmetric relations**: *AtLocation*, *CapableOf*, *Causes*, *CausesDesire*, *CreatedBy*, *DefinedAs*, *DerivedFrom*, *Desires*, *Entails*, *ExternalURL*, *FormOf*, *HasA*, *HasContext*, *HasFirstSubevent*, *HasLastSubevent*, *HasPrerequisite*, *HasProperty*, *InstanceOf*, *IsA*, *MadeOf*, *MannerOf*, *MotivatedByGoal*, *ObstructedBy*, *PartOf*, *ReceivesAction*, *SenseOf*, *SymbolOf*, and *UsedFor*

12

## Sample of ConceptNet's "knife" Info

**knife is used for...**

- stabbing
- butter
- cutting food
- carving wood
- slicing
- boning
- cut meat with
- cut string
- cutting
- cutting steak
- killing
- pare an apple
- scratching
- slicing bread

**knife is capable of...**

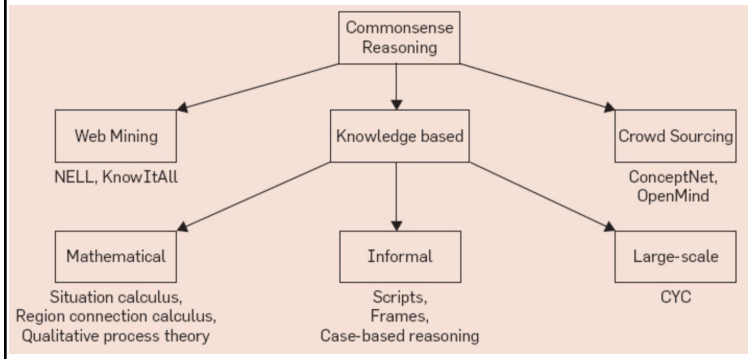
- spread butter
- spread peanut butter
- butter bread
- cut
- cut that apple
- hurt a dog
- hurt
- butter bred
- butter a piece of toast
- cut that cake
- cut cheese
- cut leather
- cut a man's hand

**Location of knife**

- the kitchen
- a drawer
- the kitchen drawer
- a pocket
- your back
- a backpack
- a drawer in a kitchen
- your ex husbands back
- a fishing boat
- a knife block
- a knife-holder
- a knife store
- a plate
- a sheath

13

## Categories of Commonsense KBs




14

## Reporting Bias

- A dream for NLP is to conquer the *knowledge acquisition bottleneck* by automatically extracting common sense knowledge from texts.
- We often assume that the more often we read something, the more likely it is to be true. Sometimes this is the case, but not always!
  - For example, some body parts are mentioned in text much more often than others. The Knext system found > 1 million instances of "*people may have eyes*" but < 1,500 instances of "*people may have a spleen*". But all body parts are equally likely in people!
- The discrepancy between reality and coverage in text has been called "**reporting bias**" [Gordon & Van Durme].
  - The problem is especially acute for common sense knowledge because these facts are so obvious to people that they are rarely mentioned!

15

## Examples of Reporting Bias Issues

- Events like murders and robberies are reported far more often than events like sleeping and breathing, which we all do every day.
- Airplane crashes are mentioned far more often than motorcycle crashes, but the latter are much more common.
- There are far more reports of people winning races than losing races, but there are many more losers than winners.
- In a forest fire story, it's common to mention that homes were destroyed and people killed, but rare to mention that deer, raccoons, and squirrels were killed.
- We're more likely to mention someone's hair color if it's purple or red, than if it's brown or black.
- If we mention a grocery store trip, we rarely mention details that almost certainly happened, such as: grabbing a shopping cart, walking down the aisles, putting items in the cart, standing in the checkout line, paying for the groceries, etc.

16

## ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning

- [Sap et al., 2019] create a large common sense knowledge base called ATOMIC that contains 877k textual descriptions of inferential knowledge.
- ATOMIC was built semi-automatically by extracting a large set of events from text corpora and crowd-sourcing information about event relationships.
- The goal is to capture common world knowledge and enable inferences about everyday events, their causes, and their effects.

17


18

## Types of Knowledge

1. **If-Event-Then-Mental-State**: captures mental pre- and post-conditions of an event: <X, action, Y>
   a) **X Intent** : likely reason why the agent performed the action
   b) **X Reaction** : emotional reaction of the agent
   c) **Other Reaction** : emotional reaction of others

For example, given event: X compliments Y

a) X wants to be nice
b) X feels good
c) Y feels flattered

19

## Types of Knowledge

2. **If-Event-Then-Event**: probable preceding and following events.
   a) **X Need** : pre-condition for the event for the agent
   b) **Effect on X**: voluntary post-conditions for the agent
   c) **Effect on Other**: voluntary post-conditions for other
   d) **X Want** : involuntary post-conditions for the agent
   e) **Other Want**: involuntary post-conditions for other

For example, given event: *X makes Y's coffee*

a) X needs to put coffee in the filter
b) X adds cream and sugar
c) Y drinks coffee
d) X gets thanked by Y

20

## Types of Knowledge

**3. If-Event-Then-Persona**: stative relation that captures how the agent is described or perceived.

**a) X Attribute** : resulting perceived attributes for the agent

For example, given the event: X calls the police

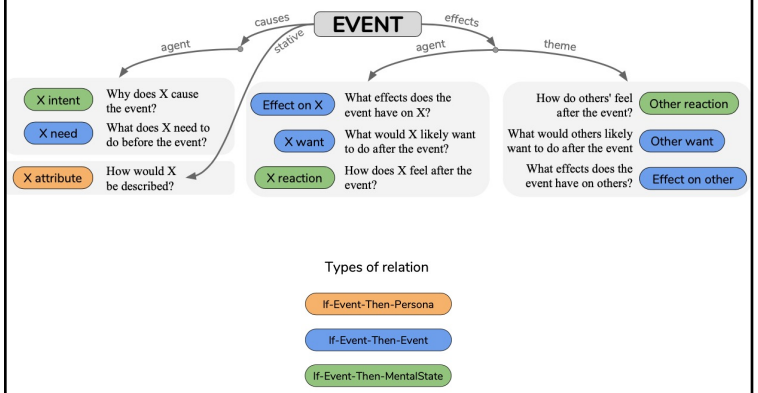
a) X is lawful

For example, given the event: X returns wallet containing money

a) X is honest

21

## 9 Types of Knowledge in All




22

## Harvesting Events

- ATOMIC is populated with events by automatically extracting events from large text corpora.
- Events are defined as a verb predicate and its arguments. For example: *he **drinks** dark roast in the morning*
- 24,000 event phrases were extracted from stories, books, Google Ngrams, and Wiktionary idioms. Frequency thresholds were applied.
- Tokens that refer to people were replaced with the special symbol **Person**, to make the event phrases more general. For example: *PersonX buys PersonY coffee*.

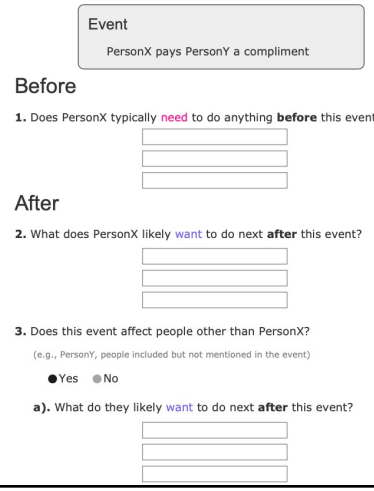
23

## Crowd-Sourcing Content

- A massive crowd-sourcing effort was used to generate the relations associated with the events.
- Workers were given questions associated with the targeted relations, and provided free-form answers.
- Workers were asked to list 3-4 likely answers for each question. Since not all dimensions apply to every event, workers could also indicate that a dimension didn't apply at all.
- To assess the quality of the results, they ran another task asking a worker to judge the validity of an annotation produced by a different worker. For a random sample of events, validity was judged to be 86%.

24

## Example of Crowd-Sourcing Interface




25

## Examples of If-Event-Then-X rules in ATOMIC

| Event | Type of relations | Inference examples | Inference dim. |
|---|---|---|---|
| "PersonX pays PersonY a compliment" | If-Event-Then-Mental-State | PersonX wanted to be nice<br>PersonX will feel good<br>PersonY will feel flattered | xIntent<br>xReact<br>oReact |
| | If-Event-Then-Event | PersonX will want to chat with PersonY<br>PersonY will smile<br>PersonY will compliment PersonX back | xWant<br>oEffect<br>oWant |
| | If-Event-Then-Persona | PersonX is flattering<br>PersonX is caring | xAttr<br>xAttr |
| "PersonX makes PersonY's coffee" | If-Event-Then-Mental-State | PersonX wanted to be helpful<br>PersonY will be appreciative<br>PersonY will be grateful | xIntent<br>oReact<br>oReact |
| | If-Event-Then-Event | PersonX needs to put the coffee in the filter<br>PersonX gets thanked<br>PersonX adds cream and sugar | xNeed<br>xEffect<br>xWant |
| | If-Event-Then-Persona | PersonX is helpful<br>PersonX is deferential | xAttr<br>xAttr |
| "PersonX calls the police" | If-Event-Then-Mental-State | PersonX wants to report a crime<br>Others feel worried | xIntent<br>oReact |
| | If-Event-Then-Event | PersonX needs to dial 911<br>PersonX wants to explain everything to the police<br>PersonX starts to panic<br>Others want to dispatch some officers | xNeed<br>xWant<br>xEffect<br>oWant |
| | If-Event-Then-Persona | PersonX is lawful<br>PersonX is responsible | xAttr<br>xAttr |

26

## Statistics for ATOMIC's Content

The resulting knowledge graph contains over 300k nodes for the original 24k events. (A node is a short phrase, 2.7 words on avg.)

Each triple is of the form <event, relation, event>

| | Count | #words |
|---|---|---|
| # triples: If-Event-Then-* | 877,108 | - |
| - Mental-State | 212,598 | - |
| - Event | 521,334 | - |
| - Persona | 143,176 | - |
| # nodes: If-Event-Then-* | 309,515 | 2.7 |
| - Mental-State | 51,928 | 2.1 |
| - Event | 245,905 | 3.3 |
| - Persona | 11,495 | 1.0 |
| Base events | 24,313 | 4.6 |
| # nodes appearing > 1 | 47,356 | – |

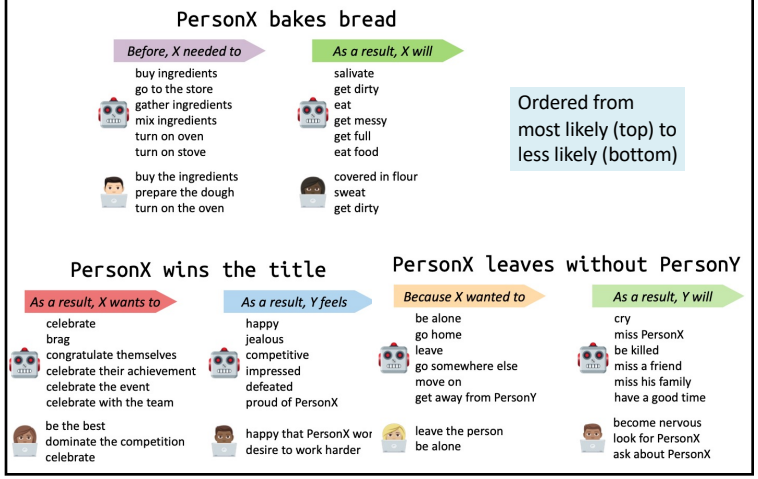
27

## Learning to Perform Commonsense Inference

- They framed the inference task as conditional sequence generation: given an event phrase **e** and an inference dimension **c**, train a model to generate the target (inference) **: $f_\theta(e,c)$**
- The architecture is an **encoding-decoding** framework:
  - First, they create an embedding representation for the event phrase beginning with pre-trained GloVe vectors concatenated with pre-trained ELMo contextualized embeddings and then further encoded with a bidirectional GRU.
  - Next, they train a decoder (unidirectional GRU) to generate an output string given the event phrase's encoding.

28

## Machine-generated Inferences




29

## Experiments

- They split the seed events into 80% training, 10% validation, and 10% test.
- Given a test event, the model generated phrases for each of the 9 dimensions of If-Then inferences.
- Automatic Evaluation: BLEU scores (n-gram matching) for the system's top 10 predictions compared with the answers from the crowd workers. (Hard to interpret results, see the paper for more.)
- Human Evaluation: randomly selected 100 events and generated 10 most likely system predictions. 5 crowd workers judged how many of them were valid.
  - Results presented as **Precision@10** : average number of correct predictions in the top 10.

30

## Experimental Results

| Model | xNeed | xIntent | xAttr | xEffect | xReact | xWant | oEffect | oReact | oWant | average |
|---|---|---|---|---|---|---|---|---|---|---|
| 9ENC9DEC | 48.74 | 51.70 | 52.20 | 47.52 | 63.57 | 51.56 | 22.92 | 32.92 | 35.50 | 45.32 |
| EVENT2(IN)VOLUNTARY | 49.82 | **61.32** | 52.58 | 46.76 | 71.22 | 52.44 | **26.46** | **36.04** | 34.70 | **47.93** |
| EVENT2PERSONX/Y | **54.04** | 53.93 | **52.98** | **48.86** | 66.42 | **54.04** | 24.72 | 33.80 | 35.08 | 46.41 |
| EVENT2PRE/POST | 47.94 | 57.77 | 52.20 | 46.78 | **72.22** | 47.94 | 26.26 | 34.48 | 35.78 | 46.76 |
| gold ATOMIC annotations | 81.98 | 91.37 | 78.44 | 83.92 | 95.18 | 90.90 | 84.62 | 86.13 | 83.12 | 86.18 |

↑ human agreement

The middle rows represent multitask learning where encoders are shared when learning related dimensions.

EVENT2(IN)VOLUNTARY : voluntary vs. involuntary dimensions
EVENT2PERSONX/Y : agent dimensions vs. other dimensions
EVENT2PRE/POST : cause dimensions vs. effect dimensions

31

## ConceptNet Comparison

Mapping to ConceptNet's relations:

- **Wants**: MOTIVATEDBYGOAL, HASSUBEVENT, HASFIRSTSUBEVENT, CAUSESDESIRE
- **Effects**: CAUSES, HASSUBEVENT, HASFIRSTSUBEVENT, HASLASTSUBEVENT
- **Needs**: MOTIVATEDBYGOAL, ENTAILS, HASPREREQUISITE
- **Intents**: MOTIVATEDBYGOAL, CAUSESDESIRE, HASSUBEVENT, HASFIRSTSUBEVENT
- **Reactions**: CAUSES, HASLASTSUBEVENT, HASSUBEVENT
- **Attributes**: HASPROPERTY

They measured the overlap between ConceptNet's triples with ATOMIC's triples: 7% wants, 6% effects, 6% needs, 5% intents, 2% reactions, 0% attributes.

Furthermore, only 25% of ATOMIC's events were found in ConceptNet. => there is substantial room for knowledge expansion.

32

# Conclusions

- ATOMIC was among the first large-scale efforts to build a knowledge graph related to events and common sense inferences associated with events.
- In principle, the dimensions capture many different aspects of common sense reasoning associated with events (causes, effects, emotional reactions, etc.)
- They then used ATOMIC to show how a neural network can be trained to generate common sense inferences for previously unseen events.
- This effort represents a major push toward trying to capture common sense knowledge, but still relied on crowd-sourcing to populate the information in the knowledge graph.
  - Nevertheless, this is a proof-of-concept showing that a model can be trained to generate such inferences for new events.
- Results are still coarsely evaluated (precision@10), although it is an open question of how to properly evaluate common sense knowledge!

33

## Using Commonsense Knowledge to Assess Semantic Plausibility

- [Wang et al., NAACL 2018] investigated the problem of recognizing the *semantic plausibility of novel events*. For example:

  *the boy swallowed a bottlecap* → semantically plausible

  *the boy swallowed a desk* → NOT semantically plausible

- Many unusual events are perfectly understandable to people, even though they are unlikely to appear in a text corpus so our models may conclude that they are impossible.
- **NLP challenge**: can we develop systems that can distinguish between an event that is impossible from an event that is entirely possible but hasn't been seen before!

1

## Semantic Plausibility ≠ Selectional Preferences

- There has been prior research on learning the selectional preferences for verbal arguments.
  - **Refresher**: selectional preferences characterize the types of semantic entities expected by a verb's argument. For example, consider *eat*: its agent is typically ANIMATE and its object is typically FOOD.
- Semantic plausibility goes beyond selectional preferences, including things that are possible even if atypical.

| *man-swallow-** | PREFERRED? | PLAUSIBLE? |
|---|---|---|
| *-candy* | ✓ | ✓ |
| *-paintball* | ✗ | ✓ |
| *-desk* | ✗ | ✗ |

2

## Knowledge about Physical Attributes

- Wang et al. argued that NLP systems need knowledge about physical properties associated with objects for this task. In particular: sentience, mass-count, phase, size, weight, and rigidity.
- Why? Because these properties often characterize the types of objects that can be sensibly involved in an action. For example:

  *throw X* → X must be solid

  *X eats* → X must be sentient

  *put X in Y* → X must be smaller than Y

3

## Gold Plausibility Annotations

- They used Amazon's Mechanical Turk to crowd-source gold labels.
- The authors collected 150 concrete verbs and 450 concrete nouns based on Brysbaert et al.'s (2014) word list with concreteness ratings. Five turkers then labeled S-V-O (Subject-Verb-Object) triples:

  (a) Have Turkers write down plausible or implausible S-V and V-O selections;
  (b) Randomly generate S-V-O triples from collected S-V and V-O pairs;
  (c) Send resulting S-V-O triples to Turkers to filter for ones with high agreement (by majority vote).

The final gold data set consisted of 3,062 S-V-O triples for which at least 3 Turkers agreed on the label.

4

## Baselines for Physical Properties

- They experimented with an existing neural network model for identifying selectional preferences to see how well it would work for this new task.
- As input, they concatenated the 300D GloVe embeddings of the 3 words in the S-V-O triple.
- The model achieved 68% accuracy.
- The data has a roughly 50/50 split of plausible vs. implausible, so this is better than baseline but still underwhelming.
- CONCLUSION: distributional similarity alone is not sufficient.
- NEXT AGENDA: if we had commonsense knowledge of physical properties, would it help?

5

## Using Landmarks to Represent Properties of Physical Objects

Instead of comparing object pairs with respect to physical properties, they proposed defining **landmark categories** and binning objects. This approach is natural for people and avoids arbitrary numeric ranges or pairwise comparisons.

- SENTIENCE: *rock, tree, ant, cat, chimp, man.*
- MASS-COUNT: *milk, sand, pebbles, car.*
- PHASE: *smoke, milk, wood.*
- SIZE: *watch, book, cat, person, jeep, stadium.*
- WEIGHT: *watch, book, dumbbell, man, jeep, stadium.*
- RIGIDITY: *water, skin, leather, wood, metal.*

For example, *dog* would be assigned to the *wood* landmark for PHASE and the *cat* landmark for SIZE.

6

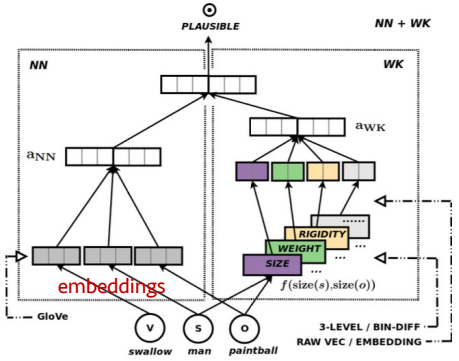
## Gold Physical Properties

- To generate "gold" knowledge about physical objects, 5 Turkers labeled the 450 concrete nouns used to create the S-V-O triples.
- Given a noun and a physical property, the annotator has to decide which of the landmarks for the property the noun is closest to.
  - For example: the landmarks for SIZE are: {*watch, book, cat, person, jeep, stadium*}
  - Given dog, the landmark *cat* should be selected.
  - Given shed, the landmark *jeep* should be selected.

7

## Injecting World Knowledge

They created a neural network model for the semantic plausibility task and showed that adding features that captured the physical property knowledge improved its performance.




8

## The WK Neural Net

- The WK (World Knowledge) NN encodes the physical properties associated with the Subject and Object. They experimented with 2 ways of encoding the knowledge for each property.
- **3-LEVEL** represents whether the Subject is <, =, or > the Object.

$$f_{3\text{-L}}(\text{SIZE}(s), \text{SIZE}(o)) \in \{-1, 0, 1\}$$

- **BIN-DIFF** captures the relative distance between the landmark categories for the Subject (S) and Object (O):

$$f_{\text{BIN}}(\text{SIZE}(s), \text{SIZE}(o)) = \text{BIN}(s) - \text{BIN}(o)$$

Example: consider SIZE: {*watch, book, cat, person, jeep, stadium*}
ant would be labeled as *watch*; man would labeled as *person*
→ BIN-DIFF(ant, man) = 1-4 = -3

9

## Experimental Results

- They present results for 10-fold cross-validation for the semantic plausibility task.
- For comparison, they show a Random model, Logistic Regression (LR), and the Selectional Preference NN alone.
- The BIN-DIFF encoding outperformed 3-Level.

| MODELS | ACCURACY |
|---|---|
| Random | 0.50 |
| LR baseline | 0.64 |
| NN (Van de Cruys, 2014) | 0.68 |
| **NN + WK-GOLD** | **0.76** |

*Physical object knowledge clearly improves performance!*

10

## Commonsense Knowledge of Quantities

- How much does a housecat weigh?
- How tall is a housecat?
- When do people typically eat breakfast?
- When do people typically sleep?
- How long are basketball games?
- What is the typical temperature on Christmas?
- How big is a ball? A house?
- How much does a ball cost? A house?

11

## How Large are Lions?

- [Elazar et al., ACL 2019] tackled the problem of learning "**quantitative attributes**" for words in their paper "*How Large Are Lions? Inducing Distributions over Quantitative Attributes*".
- They developed an unsupervised method to extract quantitative information from large text corpora and coalesce the results into reliable data.
- This paper includes valuable discussions about general challenges in acquiring commonsense knowledge from text corpora:
  - dealing with noisy data because IE from text is never perfect
  - challenges with reporting bias

12

## Task

- Acquire quantitative distributions over 10 dimensions: TIME, CURRENCY, LENGTH, AREA, VOLUME, MASS, TEMPERATURE, DURATION, SPEED, and VOLTAGE.
- Distributions are learned for:
  - **Nouns** For example: elephant, airplane, NBA game
  - **Adjectives** For example: cold, hot, lukewarm
  - **Verbs** For example: eating, walking, running
- The resulting resource is called **Distributions over Quantities (DOQ)** and contains over 350k entries (triples) that were each observed at least 1,000 times.

13

## Extracting Measurement Information

- They use a rule-based approach to detect and extract measurement information from a large text corpus.
- First, they extract all measurement mentions that they can detect. Units serve as key anchors!
- Second, they associate the measurements with nearby objects and aggregate the results.
- They intentionally aimed for a simple approach that requires only shallow resources so that this method can be applied to different languages.

14

## Extracting Measurements

- They wrote a context-free grammar to identify measurement expressions with a parser.
- They also created a mapping table between units and dimensions.

  Examples: *inch* → LENGTH *acre foot* → VOLUME
- The mapping table also defines each unit in terms of a standard unit for normalization purposes. For example, all TEMPERATURE mentions are normalized in terms of degrees Kelvin and SPEED mentions are normalized in terms of meters per second.

  Examples:

  *inch* = 0.02524 meters

  *acre foot* = 1233.48 cubic meters

15

## Extracting Objects

- All nouns, adjectives, and verbs are extracted as the objects of measurements. (NOTE: "object" in this paper means the target of the measurement, not syntactic object.)
- For each 1-word object, they also look to create a multi-word object by extracting its syntactic head. If its head is also a targeted POS, then a 2-word phrase is created from both words.

  Example: "*the fast car was driving at 100 mph*."

  *1) "fast", "car",* and *"driving"* are each extracted as objects.

  *2) "fast car"* is also extracted because the head of fast is car (noun).

NOTE: there are no details about the syntactic heads in the paper, but I'd guess that the head needs to be a noun or maybe adjective.

16

## Generating Measurement/Object Pairs

1. They collected billions of English web pages and set up a framework that allows for parallel processing.
2. Extract and normalize measurement phrases.
3. Apply a POS tagger and dependency parser.
4. Extract object phrases that occur within close proximity (same sentence or within a distance threshold) to a measurement.
5. Discard all measurements that occur in the same sentence as a negation word, because determining the precise scope of negation is a non-trivial problem.
6. Aggregate all instances with the same object head and measurement unit to obtain a distribution over values.

17

## Dataset Statistics

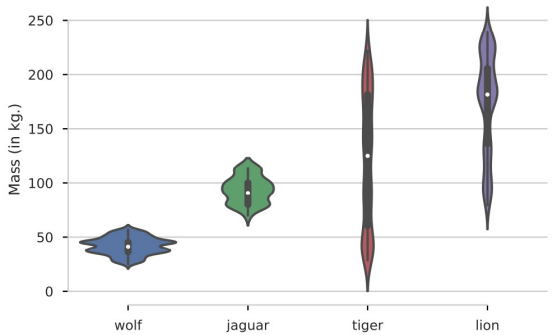
The table below shows the number of distinct object/measurement tuples that were extracted:

| Filter/Type | Nouns | Adjectives | Verbs |
|---|---|---|---|
| none | 117,953,900 | 2,513,033 | 2,121,448 |
| 5 | 16,188,215 | 598,563 | 603,799 |
| 100 | 1,497,753 | 130,534 | 160,060 |
| 1000 | 266,655 | 40,518 | 51,625 |

has at least this number of occurrences

18

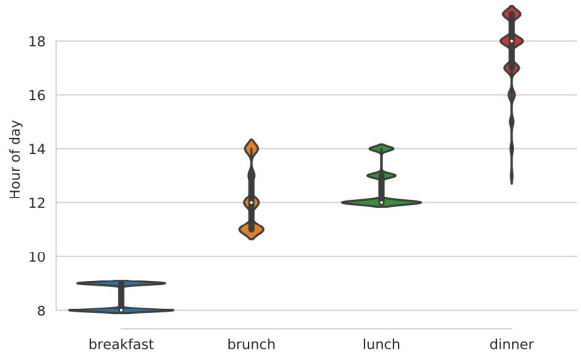
## Mass Distributions for Animals




These are “violin” plots. The white dot is the median. Wider areas represent values with a higher probability and skinnier areas represent lower probabilities

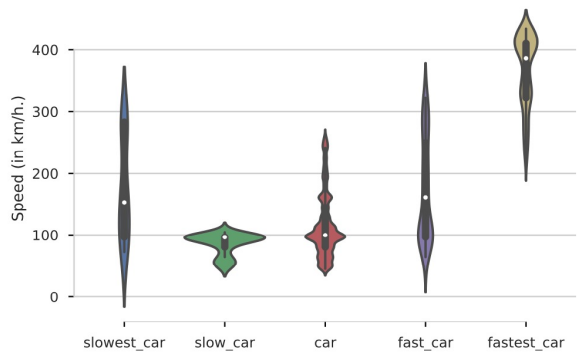
19

## Distributions over Hours of the Day




20

## Speed Distributions for Car Modifiers




21

## Evaluation

- They evaluated the quality and utility of the DoQ by applying it to several existing datasets and also with an intrinsic evaluation.
- The first evaluation focuses on the task of identifying the relative physical relationships between object pairs (as in [Forbes & Choi, 2017]). Formally:

  Given $O_1$ and $O_2$, predict the relation { < , =, > }

- To apply the DoQ, they look up the distributions for the designated property for $O_1$ and $O_2$ and compare their medians.

22

## Comparison with [Forbes & Choi, 2017]

- They discovered problems with the original [Forbes & Choi, 2017] set, so they re-labeled the data via a new crowd-sourcing process.
  - Their version of the data set is substantially smaller (because they tossed problematic cases), but overall they claim it has higher quality labels.
- Applying their DoQ to this data produces better results than the previous state-of-the-art system [Yang et al., 2018].

| Model/Dataset | F&C Clean | | New Data | |
|---|---|---|---|---|
| | Dev | Test | Dev | Test |
| Majority | 0.54 | 0.57 | 0.51 | 0.50 |
| Yang et al. (PCE LSTM) | **0.86** | **0.87** | 0.60 | 0.57 |
| DoQ | 0.78 | 0.77 | **0.62** | **0.62** |
| DoQ + 10-distance | 0.78 | 0.77 | **0.62** | **0.62** |
| DoQ + 3-distance | 0.81 | 0.80 | **0.62** | 0.61 |

23

## Similar Results on Size Data

They also evaluated their DoQ on a data set produced by [Bagherinezhad et al., 2016] of 486 object pairs labeled with respect to physical size.

| Model | Accuracy |
|---|---|
| Chance | 0.5 |
| Bagherinezhad et al. | 0.835 |
| Yang et al. (Transfer) | 0.858 |
| DoQ | 0.872 |
| DoQ + 10-distance | **0.877** |
| DoQ + 3-distance | 0.858 |

24

## Adjective Evaluation

- Prior work has focused on learning the relative intensities of adjectives, so they exploited some of that existing data to evaluate the quality of their adjective results.
- They collected adjectives that had comparative intensity labels and manually assigned the appropriate type of dimension. For example:

  Hot and Cold are not comparable

  Cold < Frigid : TEMPERATURE dimension

  Tiny < small : SIZE dimension

- They also adopted a different procedure for applying the DoQ because adjectives can lead to very different inferences depending on what they are modifying. For example, consider:

  small dog vs. small car

25

## Comparing Adjective Distributions

**Algorithm 1** Adjectives Comparison Inference

**Input:** adjectives $x$,$z$, dimension $d$ and object distributions $H$
**Output:** comparison label
**Procedure:**
Initialize $\hat{y}$, the predictions per head
$intersect \leftarrow$ findHeadIntersection($H$, $x$, $z$, $d$)
▷ the intersecting heads of $x$ and $z$
**for** $a_i, b_i \in intersect$ **do**
$\hat{y}_i \leftarrow$ compare($a_i$,$b_i$,$d$)
**end for**
Return majority($\hat{y}$)

26

## Results on Scalar Adjective Dataset

The table below shows results for the adjectives evaluation across 3 different data sets.

| Model | deMelo | Wilk-intense | Wilk-all |
|---|---|---|---|
| Global Ranking | **0.642** | 0.818 | - |
| Cocos et al. | 0.620 | **0.841** | - |
| DoQ | 0.617 | 0.700 | 0.870 |
| DoQ + 10-distance | 0.608 | 0.750 | **0.891** |
| DoQ + 3-distance | 0.567 | 0.500 | 0.761 |

27

## Intrinsic Evaluation

A sample of the DoQ was manually annotated to directly assess the quality of its entries.

| Mass | Length | Speed | Currency | All |
|---|---|---|---|---|
| .61 | .79 | .77 | .58 | .69 |

The annotators were from India, and the authors wondered if the low Currency results could be from cultural differences in perceived prices. So they had U.S. annotators re-label Currency. Their agreement = 76%!

Example: Indian annotators said that a suit could not cost between $1K-$10K, while U.S-based annotators reported it was possible.

We tend to forget about cultural differences when we crowd-source data but it can be a significant issue!

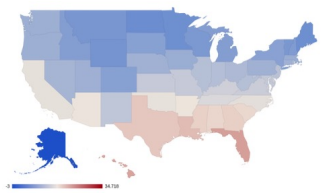
28

## Reporting Bias Discussion

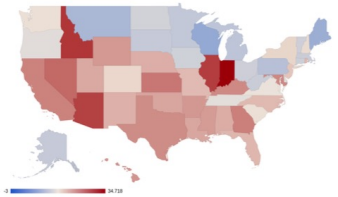
- They thought that their technique would be relatively robust to reporting bias issues because they used a massive text collection and because they focused on numeric measurement data.
- However they observed that:
  - People tend to discuss objects when they are exceptional (e.g., "*I saw an extremely tiny horse*").
  - People sometimes exaggerate measurements for rhetorical effect. For example, they found that people tend to exaggerate hot temperatures more than cold temperatures.

29

## Reporting Bias Example



(a) Measured average temperature

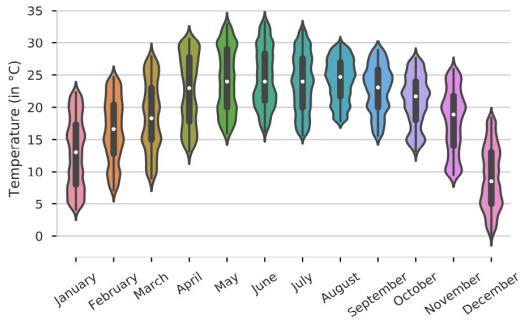


(b) Induced average temperature

Figure (b) shows that people tend to talk about or exaggerate hot temperatures more than cold ones!

30

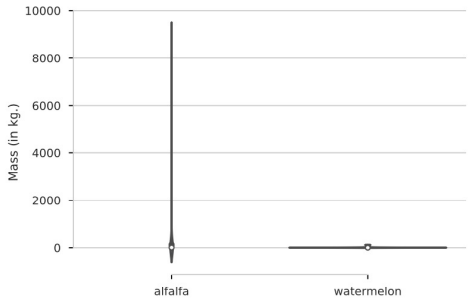
## Reporting Bias Example




These temperature patterns reflect a reporting bias for the Northern Hemisphere!

31

## Reporting Bias Example




Alfalfa is typically mentioned in farming contexts, so harvesting large amounts (tons). In contrast, watermelon is typically mentioned in terms of individual items so small units (grams).

32

## Conclusions

- [Elazar et al., 2019] showed that a simple, unsupervised approach for extracting measurements from a large corpus can produce a resource that compares favorably against methods that require more resources and offer less coverage.
  - Extracting massive volumes of data in a simple way and taking care to clean it properly can be remarkably effective sometimes!
- However, the quality of the data is still mixed if you look through the resource. The notion of "object" probably should be refined.
- Also, the representation of ranges is still relatively crude.
- Even for this relatively straightforward kind of data, reporting bias is still an issue! This is an on-going challenge for nearly all problems related to harvesting commonsense knowledge.

33

## Commonsense Knowledge Base Completion

- In recent years, some researchers have tackled the problem of **commonsense knowledge base completion (CKBC)**, which is the task of automatically expanding an existing common sense KB.
- These models typically use supervised learning and are trained with an existing KB such as ConceptNet. A held-out sample of the KB is reserved as a test set.
- These methods have produced evaluation scores that can look quite impressive.
- However, a deeper dive into the results has shown that they are primarily rephrasing the relations found in the training set. *Training set leakage* is a common problem.

This begs the question: is supervised learning really the answer?

1

## Training Set Leakage Problem

- Training set leakage refers to situations were the training and test sets accidentally share information.

  *"information is revealed to the model that gives it an unrealistic advantage to make better predictions"*

- As a result, performance on the test set is artificially high! Applying the system to a truly blind test set yields much lower results.
- Some examples of how this can happen:
  - A single document is split across the training and test sets.
  - Duplicates may exist (this is very common with Twitter data!)
  - News articles from same time period.
  - Very similar cases exist (e.g., paraphrases). This case is hard to avoid, and likely a common reason why different data sources usually perform worse.
  - In general, train/test splits share more vocabulary than new data sets.

2

## Types of Neural Language Models

Causal language models process an input sequence from **left-to-right** and **predict the next word in the sequence.**

$$p(w_1, w_2, \ldots, w_m) = \prod_{}^{m} p(w_i | w_1, \ldots, w_{i-1})$$

GPT (Generative Pre-trained Transformer) models are a family of neural LMs trained on massive text collections.

---

Masked Language Models (MLM) are a new breed of neural LMs that are bidirectional and trained to predict a small number of words that have been "masked" : $p(w_i | w'_{1:i-1}, w'_{i+1:m})$

$w' \in \mathcal{V} \cup \{\kappa\}$ where $\kappa$ is a special Mask token

- The MLM returns a probability distribution over words in the Mask position.
- BERT (Bidirectional Encoder Representations from Transformers) is a MLM that is widely used and trained on massive text corpora.

3

## Recap: Masked Language Modeling

- **Masked Language Models** (**MLM**s) are trained by "masking" a position of the input.
- Causal transformer models are trained with a ***causal mask*** to predict the next word:

  *Peanut butter and _____*

- Bidirectional transformer models are trained with a **cloze test**, which is a "fill in the blank" mask:

  *Peanut butter and _____ sandwich*

- To train with a MLM, many input tokens are sampled. For BERT, 80% are then masked, 10% are randomly replaced, and 10% are left unchanged.

4

## Harvesting Commonsense Knowledge from Language Models

- Pre-trained language models (LMs) have become the backbone of many NLP models recently, to produce contextual embeddings that are used as rich semantic representations for downstream tasks.
- Researchers have also begun to explore whether large language models can be useful for recognizing commonsense knowledge.
- A key observation is that a LM can estimate the probability of a sentence based on its training corpus**. The likelihood of the sentence can serve as a proxy for the truth of the knowledge**.

*Dogs are mammals.* ✓ (high prob)

*Dogs can run.* ✓ (high prob)

*Dogs can fly.* ✗ (low prob)

5

## Exploiting Masked Language Models

- [Feldman et al., EMNLP 2019] explored the idea of using Masked Language Models (MLMs) to validate commonsense knowledge facts.
- Given a relational triple <E, R, E>, they present an approach for using a MLM to estimate the likelihood that the relation is a valid sentence (and therefore true knowledge).
- This approach could be used to validate any relational triple, but their focus is on evaluating commonsense knowledge relations.
- Similar ideas have been explored for assessing the plausibility of a relation, for tasks like textual entailment.

6

## Representing Facts

- Each candidate fact is represented as a **head-relation-tail triple**: **x = <h, r, t>**
- They assume that there is a fixed set of known relations **R**, so **r** ε **R**.
- The head and tail can be multi-word phrases, where each word comes from a known vocabulary V.

$\mathbf{h} = \{h_1\ h_2 \ldots h_n\}$ $\mathbf{t} = \{t_1\ t_2 \ldots t_m\}$

- The goal is to learn a function **f(x) = y** such that **y** reflects the model's confidence that **x** represents true knowledge.

7

## Generating Sentences

Language models expect sentences as input, so each triple needs to be converted into a grammatical sentence. But this is not as simple as it sounds! For example, consider:

(ferret, AtLocation, pet store)

Possible sentences:

*A ferret is at a pet store.*
*The ferret is at the pet store.*
*The ferret is in the pet store.*
*The ferret is located in the pet store.*

Possible sentences: (water, AtLocation, Mars)

*A water is at a Mars.*
*The water is at the Mars.*
*The water is in the Mars.*
*The water is located in the Mars.*

8

## Generating Sentences from Triples

1. They apply grammatical transformations to each head and tail:
   - If the first word is a N or ADJ, or the first word is a V and the second word is a N or ADJ, then prepend an indefinite or definite article.
   - If the first word is an infinitive form, make it a gerund (e.g., *jump* → *jumping*)
   - If the first word is a number, pluralize the next word (e.g., *two leg* → *two legs*)
2. For each relation, they manually define a set of templates.

   For example, the templates below are used for *<X, CapableOf, Y>* :

   *X can Y*  *An activity X can do is Y*

   *X often Y*  *X sometimes Y*

9

## Coherency Ranking

- All possible combinations of candidate sentences are generated.
- Then they select the candidate sentence with the highest log-likelihood score from a unidirectional LM: $\mathbf{P_{coh}}$.

$$S^* = \arg\max_{S \in \mathcal{S}} [\log P_{\text{coh}}(S)]$$

- The expectation is that natural, grammatical sentences will have higher LM scores than unnatural or ungrammatical sentence.
- In practice, they found that this approach produced "significantly higher quality" results than using deterministic rules alone.

10

## Coherency Ranking Example

LM scores for several candidate sentences generated from the triple: *(musician, CapableOf, play musical instrument)*

| Candidate Sentence $S_i$ | $\log p(S_i)$ |
|---|---|
| "musician can playing musical instrument" | $-5.7$ |
| "musician can be play musical instrument" | $-4.9$ |
| "musician often play musical instrument" | $-5.5$ |
| "a musician can play a musical instrument" | $\mathbf{-2.9}$ |

11

## Scoring Sentences

- Once the best sentence has been identified, it needs to be scored to assess its validity.
- A sentence is scored based on the PMI of the head and tail conditioned on the relation.

$$\text{PMI}(\mathbf{t}, \mathbf{h}|r) = \log p(\mathbf{t}|\mathbf{h}, r) - \log p(\mathbf{t}|r)$$

- Ultimately they use a weighted PMI measure, where the lambda ($\lambda$) value is a tuned hyperparameter.

$$\text{PMI}_\lambda(\mathbf{t}, \mathbf{h}|r) = \lambda \log p(\mathbf{t}|\mathbf{h}, r) - \log p(\mathbf{t}|r)$$

- PMI is symmetric in theory, but this approximation is not. So they average $\text{PMI}_\lambda(\mathbf{t}, \mathbf{h}|r)$ and $\text{PMI}_\lambda(\mathbf{h}, \mathbf{t}|r)$ .

12

## PMI Redux

$$PMI(X,Y) = \log_2 \left( \frac{P(X, Y)}{P(X) * P(Y)} \right)$$

$$PMI(t,h \mid R) = \log_2 \left( \frac{P(t,h \mid R)}{P(t|R) * P(h|R)} \right)$$

$$\log_2 \left( \frac{P(t \mid h,R) * P(h|R)}{P(t|R) * P(h|R)} \right)$$

$$\log_2 \left( \frac{P(t \mid h,R)}{P(t|R)} \right)$$

$$= \log_2 (P(t \mid h,R)) - \log_2 (P(t \mid R))$$

13

## Estimating Probabilities from a MLM

- We can estimate the probability of the tail with a single mask position in a bidirectional MLM model ($P_{cmp}$)

$$P(\mathbf{t} \mid \mathbf{h},R) = P_{\text{cmp}}(w_i = \mathbf{t} \mid w_{1:i-1}, w_{i+1:m})$$

- If the tail is a multi-word term, then they use a greedy approach to estimate a probability for the phrase. For a phrase with j terms:
  - Mask each word, one at a time, and compute the probability for each.
  - Select the word with the highest probability and insert it in the phrase
  - Repeat this process j times.
- Finally, the probability for the phrase is the product of the probabilities found for the individual terms:

$$p(\mathbf{t}|\mathbf{h}, r) = \prod_{k=1}^{j} p_k$$

14

## Estimating Probabilities from a MLM

Computing P(**t** | R) is similar but it needs to be estimated over all possible heads. So the head must be masked throughout.

For example, consider the sentence:

*You are likely to find a ferret in the pet store.*

Initially, both the head and tail are masked:

*You are likely to find a $h_1$ in the $t_1$ $t_2$.*

The tail terms are gradually inserted based on their probabilities:

*You are likely to find a $h_1$ in the $t_1$ store.* (→ $P_{store} > P_{pet}$)

*You are likely to find a $h_1$ in the pet store.* (→ $P_{pet}$)

*The final probability is then $P_{store} * P_{pet}$*

15

## Experimental Set-up

- For sentence ranking, they use the GPT-2 Language Model. For the masked language model, they use BERT (large model).
- As a baseline for sentence generation, they simply split the words in the relation name and **concatenate** the head and tail.

  For example:
  *(ferret, AtLocation, pet store)* → "*ferret at location pet store*"
- They also evaluate against a Commonsense Knowledge Base Completion (CKBC) approach by instantiating a single **template** with the head and tail.

  For example: *(ferret, AtLocation, pet store)* →
  "*you are likely to find ferret in pet store*"

16

## Task 1: Commonsense Knowledge Base Completion

- Use test set from (Li et al., 2016), which contains 2400 triples that contain an equal number of **Valid/Invalid triples**.
- The Valid triples come from the crowd-sourced Open Mind Common Sense (OMCS) entries in the ConceptNet 5 dataset.
- The Invalid triples were produced by replacing one of the elements in a valid triple with a randomly selected item.
- The Coherency Rank model is applied to produce a score for each triple and the triples are grouped into two clusters based on their scores. All of the triples in the cluster with the highest mean PMI are labeled as Valid.

17

## Task 2: Mining Wikipedia

- (Li et al., 2016) curated a data set mined from Wikipedia primarily using part-of-speech patterns: 1.7M triples across 10 relations.
- For this work, 300 triples were sampled for each relation, producing a test set of 3,000 relation triples.
- The Coherency Rank method scored each triple, and the top-scoring 100 triples were manually reviewed by 2 human annotators on a scale from 0 to 4:

  0 : doesn't make sense
  1 : not true
  2 : opinion / don't know
  3 : sometimes true
  4 : generally true

NOTE: kappa agreement was only .23! But if only two buckets are used, then disagreement drops by 50%.

18

## Results

| Model | Task 1 | Task 2 |
|---|---|---|
| Unsupervised | | |
| CONCATENATION | 68.8 | $2.95 \pm 0.11$ |
| TEMPLATE | 72.2 | $2.98 \pm 0.11$ |
| TEMPL.+GRAMMAR | 74.4 | $2.56 \pm 0.13$ |
| COHERENCY RANK | 78.8 | $\mathbf{3.00} \pm 0.12$ |
| Supervised | | |
| DNN | **89.2** | 2.50 |
| FACTORIZED | 89.0 | 2.61 |
| PROTOTYPICAL | 79.4 | 2.55 |

For Task 1: The Coherency Rank approach performed better than the other unsupervised approaches, but not as well as the supervised CKBC systems.

For Task 2: The Coherency Rank approach outperformed all the other models, including the supervised CKBC systems.

19

## Breakdown of Results for Task 1 & Task 2

They analyzed their results based on whether the sentence had a grammatical error or misrepresented the relation's meaning.

For example, *(golf, HasProperty, good)* → "*golf is a good*"
Is grammatical but captures a different meaning.

| Task 1 | N (/100) | F1 Score |
|---|---|---|
| GRAMMATICAL | 75 | 79.1 |
| UNGRAMMATICAL | 25 | 66.7 |
| CORRECT MEANING | 91 | 77.6 |
| WRONG MEANING | 9 | 66.7 |
| **Task 2** | - | **Quality** |
| GRAMMATICAL | 83 | 3.01 |
| UNGRAMMATICAL | 17 | 2.88 |
| CORRECT MEANING | 88 | 3.22 |
| WRONG MEANING | 12 | 1.18 |

20

## Error Analysis

**Most Confident Mistakes**: the following triples were in the top 100 predictions but received scores < 3 by the human annotators.

(atomic nucleus, IsA, atom)
(negative number, HasProperty, positive)
(the harbor, HasA, island)
(the substance, HasA, drug)
(plurality voting, HasPrerequisite, majority)
(function, ReceivesAction, element of a)
(minister, ReceivesAction, member of parliament)
(bombing, IsA, war crime)
(prime minister, ReceivesAction, head of state)
(film, Causes, silent version)
(island, AtLocation, other side)
(subset, ReceivesAction, element of s)
(monarchy, ReceivesAction, form of government)
(law, ReceivesAction, cause of action)
(weather, UsedFor, heavy rain)
(example, UsedFor, word processing)

21

## Conclusions

- The Coherency Rank method is a creative approach for exploiting massive pre-trained LMs to confirm or disconfirm whether a fact is likely true or not.
- This is a powerful idea, since these LMs are essentially encoding massive amounts of textual data.
  - Instead of explicitly harvesting facts from a large corpus, can we essentially probe a LM to discover what it knows?
- This approach is also exciting because it does not use supervised learning so no manually annotated data is needed.
- We've just begun to scratch the surface of how these large LMs may be used! This direction seems very promising.

22

## Prompting Methods

- ***Prompting methods*** have become a hot area in NLP that exploit pre-trained LMs as an alternative to supervised learning.
- Large pre-trained language models can be viewed not only as a repository of language but also as a *repository of knowledge*. *Prompt-based methods* aim to extract knowledge from pre-trained LMs with carefully designed inputs that essentially query the LM for missing information.
- Typically, a *template* is combined with an input value to create a *prompt string* that has one or more unfilled positions. The language model can then return the words or phrases that are most likely to occur in those positions with probability estimates.
- The trick is to design good templates for your task!

23

## Example: Sentiment Analysis

Suppose you want to get a sense of the popularity of a movie, such as "Arrival". The typical approach would be to collect a corpus of reviews for the movie and apply a sentiment classifier, ideally one trained on movie reviews.

Alternatively, you could use prompting with the template:

*"<MOVIE> was a <MASK> movie"*

The input to the LM would be "*Arrival was a [X] movie*", and the LM would then return a probability distribution over words that could fill the X position. For example:

*great* (.10), *terrific* (.09), *boring* (.07), *sci-fi* (.06), *good* (.06) …

You could then use a sentiment lexicon to assess the overall polarity.

24

## Prompt Engineering

- Designing an effective prompt for your task is key. Language models can be sensitive to the specific words used in the prompt, even punctuation can matter.
- There are different types of prompts, such as:
  - **Prefix prompts**: the LM completes the input
  - **Cloze prompts**: the LM fills a blank in the middle
  - **Zero-shot prompts**: no examples are given
    Example: "*The capital of Utah is [X]*"
  - **Few shot prompts**: one or more examples are given.
    Example: "*A list of cities in Utah. 1. Salt Lake City*"

25

## Challenges with Prompting

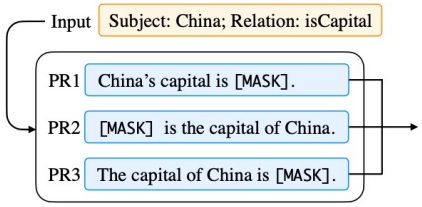
Prompting methods can sometimes work very well and their potential is intriguing, but there are also major challenges.

- Multi-word answers can be difficult to extract with a mask.
- Hard to control syntactic expectations (many types of answers are possible due to complex syntactic constructions).
- The most common answers dominate. It can be difficult, perhaps impossible, to extract uncommon/rare information.
- Complex prompts can be challenging because the LM often focuses on the local context around the mask and may ignore the full context of the prompt.

26

## Prompt Ensembling

Since any one prompt may not be ideal, multiple prompts can be used in a sort of "ensemble". The results are then pooled and can be ranked based on voting or more complex ranking methods.



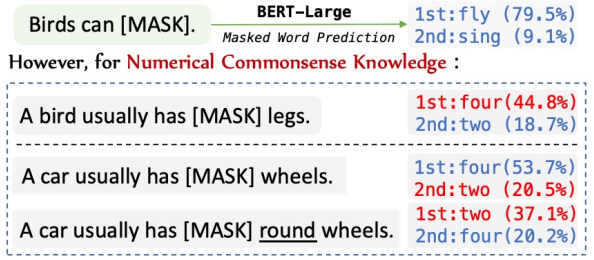

(a) Prompt Ensembling.

See *"Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing"* [Liu et al., 2021] for a comprehensive survey of prompting methods.

27

## Also … LMs alone may not be enough

[Lin et al., EMNLP 2020] recently showed that masked language models are not very good at learning numerical common sense knowledge.




Even when the predictions are correct, they can be brittle.

28

## LMs can have strong biases

The table below shows the top 3 predictions when the [x] variable is instantiated with 1k random words.

| Template: | a [x] usually has [MASK] legs. |
|---|---|
| BERT–L | four: 39.3%, two: 18.3%, three: 10.1% |
| RoBERTa–L | four: 20.8%, two: 9.0%, three: 8.1% |
| **Template:** | **most [x] have [MASK] wheels.** |
| BERT–L | four: 25.3%, two: 14.1%, three: 5.1% |
| RoBERTa–L | four: 9.2%, two: 7.8%, three: 4.6% |
| **Template:** | **all [x] have [MASK] sides.** |
| BERT–L | two: 28.3%, three: 12.9%, four: 12.9% |
| RoBERTa–L | two: 16.6%, no: 2.9%, three: 2.3% |

There is often one dominant value that seems to emerge based on the most common cases.

29

## Summary: Commonsense Knowledge

- We have a long way to go to be able to accurately acquire the vast amounts of commonsense knowledge that are needed.
- We need tons of knowledge, but it also must be organized and represented in a useful way.
- However, there are many promising avenues for acquiring specific types of knowledge, automatically and semi-automatically.
  - Researchers nearly always focus on fully automatic techniques, but in the real-world semi-automatic techniques can be extremely valuable!
  - A small amount of human effort to "curate" automatically extracted information can insure high-integrity data and often produce substantially more data than manual efforts would.

30

## Summary: Information Extraction

- IE is a rich area of NLP that covers a wide variety of problems!
  - Some topics correspond to fundamental aspects of text understanding, such as Named Entity Recognition, Semantic Class Learning, and Temporal IE.
  - Other topics are closely tied to real-world applications, such as Relation Extraction, Event Extraction, Opinion Extraction.
- For most sequence labeling tasks, NLP models can be trained to perform reasonably well for a specific domain *given sufficient domain-specific training*.
- IE tasks tend to be more challenging than non-sequential classification tasks, and it is difficult to achieve high recall and high precision at the same time.
- Many challenges for the future, including cross-sentence (document-level) models, learning with small amounts of labeled data, and IE methods to harvest domain-specific knowledge!

New IE problems are always around the corner! You'll see some in the projects. ☺

31