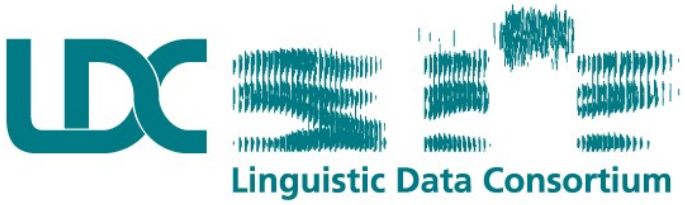


Schema Learning Corpus: Data and Annotation Focused on Complex Events

Song Chen, Jennifer Tracey, Ann Bies, Stephanie Strassel

Linguistic Data Consortium

Linguistic Data Consortium

Introduction: KAIROS

- DARPA's Knowledge-directed Artificial Intelligence Reasoning Over Schemas (KAIROS) program aims to
  - Build technology capable of understanding and reasoning about complex real-world events in order to provide actionable insights to end users
  - KAIROS systems utilize formal complex event representations in the form of schema libraries
  - Schemas are used in combination with event extraction to characterize and make predictions about real-world events

Introduction: Schema Learning Corpus

- Schema Learning Corpus (SLC) is designed to support the development and evaluation of technology capable of generalizing schemas
  - Background data: a large and varied corpus with data from many different domains in multiple languages for users to generalize schemas
    - Diversity in media, genre, formality etc.
    - Three languages: English, Spanish and Russian
  - Complex Event data: contains source data on 100 complex events, with exemplar annotation
    - Annotation in English and Spanish: entities, events, relations, and links of events and relations to complex event steps

SLC: Background Data

- 20 existing LDC corpora and new Spanish collection. Data not annotated.

	Text-only Docs	Multimedia docs	Total doc count
English	11,877,669	100,282	11,977,951
Spanish	3,822,559	14,942	3,837,501
Russian	435,017	12,652	447,669
All Languages	16,135,245	127,876	16,263,121

- ~15,000 Spanish multimedia documents newly collected include:
  - Instructional documents
  - Business and logistics domains
  - Multimedia data
- Russian documents added during the 2nd phase of KAIROS

Linguistic Data Consortium

SLC: Complex Event Data

- Purpose: Wide variety of complex events, not focused on a specific scenario
- Source Data: broad variety of languages, domains, data types
  - English, Spanish
  - Wide range of domains
  - Multimedia (text, image, video, audio)
  - Mostly typical news, blog, forum documents
- Annotation: handful of labeled instances per CE as exemplars
  - Not one real-world incident with complete annotation coverage
  - Variety of exemplars across multiple specific incidents per CE

Develop Complex Events

- Twelve general domain areas
- 100 CEs defined, each of which may fit into multiple domains

Business Workings
Civil Unrest
Conflict/Threat
Disaster
Government Workings
Cyber/Information
Illegal Activities
Legal Proceedings
Medical Intervention
Movement/travel
New Capability Development
Social Life

Disease outbreak
Evacuation
Provide/distribute disaster relief
Search and rescue
Supply shortage

Quarantine
Obtain or provide medical treatment
Develop a biological agent
Develop a new medicine
Experiment

Develop Complex Events

- Profiles for each CE type are used to inform data scouting and annotation

ID: ComplexEvent008

Title: Obtain or Provide Medical Treatment

Description: Medical treatment is applied to one
condition or injury. This event can include comm
offering help, transportation to a medical facility
with medical personnel or others as part of the p
treating the patient's condition. The focus of this
treatment by professionals such as doctors, nur
some treatments may also be applied by non-pr
may perform CPR while waiting for an ambulanc

Scope of event: The event begins when someo
treatment (or someone else seeks treatment on
arise from illness, injury, accident, or other caus
is not part of the event. The event ends when tr

Step 1: Request Treatment: Medical treatme
Expected kinds of events: Communication

Step 2: TravelForTreatment: Travel to bring patient and medical personnel together. This may involve medical personne location and/or the patient traveling to a me
Expected kinds of events: Movement (of p

Step 2.1 TransportMedicalPersonnel: Medica location of the person needing treatment
Expected kinds of events: Movement (of pe

ID and Title unique to each Complex Event

Description contains basic information about this Complex Event: the types of activities it includes and how it differs from other events

Scope indicates the boundaries of where the event begins and ends **for purposes** of SLC instances

Steps that are typically part of this CE, in their typical order, and **event types** that are likely to appear for each step

Linguistic Data Consortium

SLC Multi-layer Annotation

Document Scouting

Collect data in multiple langs, modalities depicting every CE step

Provenance Linking

Point to doc mention(s) for each step and select event type(s)

Event and Relation Mention

Complete the event/relation frame: args, attributes, temporal info

Result: At least 5 labeled instances for each Complex Event, with multiple examples of each step drawn from different instances, linked to CE steps

Linguistic Data Consortium

Data Scouting for Instance Variety

- Manual data scouting for multiple documents that contain specific, varied instances of general CE
- Variety in data sources, modalities, languages, etc.
- Also variety rather than uniformity in ***realization*** of instances

Lang	CE009 provide and distribute disaster relief
Eng	Local community aids victims of Doncaster Flooding
Eng	A group people raise money for Hurricane Sandy
Eng	Collecting money to help flood victims in
Eng	New Zealand farmers received aid for dr
Spa	Ship supplies to Bahamas after Hurrican
Spa	Red Cross' response to 2017 earthquake City
Spa	Humanitarian aid arrives in Venezuela

Complex Event Step Number	Complex Event Step Name	Docs with Text	Docs with Audio	Docs with Image	Docs with Video
1	CollectAid	4	1	1	2
1.1	AppealForAid	6	2	1	1
1.2	CollectPhysicalAid	4	1	2	3
1.3	CollectFinancialAid	4	3	1	1
1.4	PurchaseAidItems	1	1	0	1
2	TransferAidToRegion	3	1	0	2
2.1	TransportAidToRegion	3	1	1	2
2.2	TransferFunds	3	2	1	1
2.3	DistributeAid	5	1	1	3

MOU2

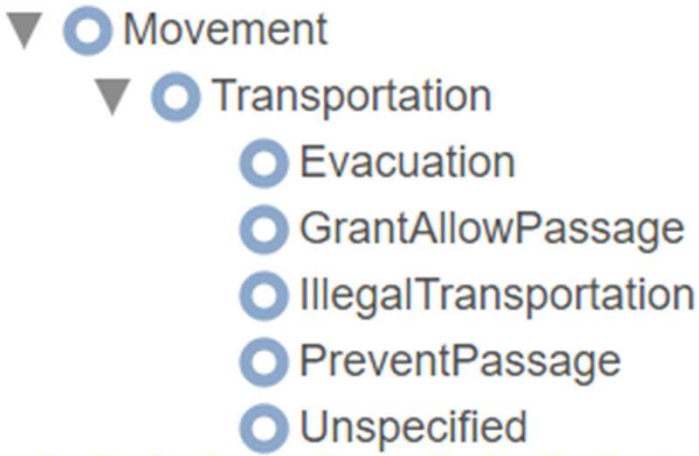
If short on time, skip this slide

Microsoft Office User, 11/1/2021

Annotation Tag Set

- KAIROS Phase 1 annotation tag set, used for both SLC and Evaluation
- Significant expansion from ACE, DEFT ERE
  - Cover CE-relevant areas, e.g. cognitive events
  - Finer grained subtypes to support schema instantiation, e.g. movement.transportation has 5 subsubtypes
  - Some event and relation types allow events as arguments

Category	SLC type count
Event types	67
Relation types	46
Entity/filler types	24
Total types	137



MOU3

Linguistic Data Consortium

Labeling Provenance

Doncaster Flooding, England, Nov 2019

Some residents in Fishlake near Doncaster are expected to be out of their homes for weeks

Event Type: Transaction.Donation

Linked step: **Step 1.2: Collect Physical Aid**

Provenance: video 0:58-1:05

Hurricane Dorian, Bahamas

Suministrar efectos de socorro por aire y por mar

USAID transportó vía aérea 47 toneladas métricas de suministros de socorro crítico el 4 de septiembre desde su almacén de emergencia en Miami a las Bahamas

Event Type: Movement.Transportation

Linked step: **Step 2.1: Transport Aid to Region**

Provenance: text 2229-2759

MOU3

If short on time, skip this slide

Microsoft Office User, 11/1/2021

Linguistic Data Consortium

Labeling Mentions

Hurricane Dorian, Bahamas

Suministrar efectos de socorro por aire y por mar

USAID transportó vía aérea 47 toneladas métricas de suministros de socorro crítico el 4 de septiembre desde su almacén de emergencia en Miami a las Bahamas

Event Type: Movement.Transportation

Linked step: **Step 2.1: Transport Aid to Region**

Provenance: text 2229-2759

trigger: transportó

Attribute: N/A

Movement.Transportation

- **SourcePlace:** de emergencia en Miami
- **DestinationPlace:** Bahamas
- **Agent:** USAID
- **MeansInstrument:** N/A
- **Things:** 47 toneladas métricas de suministros de socorro crítico

Timestamp

- **Start and End:** on 2019-09-04

MOU4 If short on time, skip this slide

Microsoft Office User, 11/1/2021

Quality Control

- Manual and automatic checks for consistency, completeness and data integrity
- Document level QC: manual 2nd pass annotation to correct errors
- CE level QC to check:
  - Sufficient coverage of steps, languages, media types, other features
  - Every doc contains good exemplars of CE/steps for system development
  - As a whole, docs and annotations reflect good range of diversity for CE instantiation
- Corpus wide QC:
  - Automated checks for invalid annotation
  - Quantitative and qualitative review and revision of trends across annotators

SLC Summary by Domain

		Released Docs		
CE Domain	**CEs**	**Source**	**Provenance Linked**	**Mentions Annotated**
Business Workings	15	614	195	54
Civil Unrest	3	101	40	13
Conflict/Threat	11	400	133	38
Cyber/Information	6	215	65	23
Disaster	5	181	63	23
Government Workings	10	400	116	38
Illegal Activity	7	247	89	26
Legal Proceedings	9	428	114	35
Medical Intervention	4	139	54	17
Movement	3	159	39	10
New Capability Development	14	516	169	41
Social Life	6	291	66	21
Total	**93**	**3691**	**1143**	**339**

Conclusions

- The Schema Learning corpus makes a unique contribution to the available resources on schema and event detection
  - Multilingual, multimedia background source data for schema induction
  - 100 CEs defined across multiple domains; SLC includes multilingual, multimedia source data and annotation for schema learning for 93 of these CEs
  - Event, relation and entity annotation of multimedia data in English and Spanish
  - Events and relations linked to schema steps to show schema instantiation by real-world event
- Currently available within the KAIROS program
- To be published in the LDC catalog once program needs permit