Neural Relation Extraction from Unstructured Texts

Jinhua Du

ADAPT Centre, Dublin City University, Ireland

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Outline

- ✓ What is Relation Extraction
- ✓ Distant Supervision for RE
- ✓ Multi-Level Structured Attention
- ✓ Application

E-Mail: jinhua.du@adaptcentre.ie

An Example: What is Relation Extraction

Company report: "International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)..."

Standard Information Extraction Task:

Company	IBM
Location	New York
Date	June 16, 1911
Original-Name	Computing-Tabulating-Recording Co.

Relation Extraction Task:

Founding-year (IBM, 1911)

Founding-location (IBM, New York)

E-Mail: jinhua.du@adaptcentre.ie

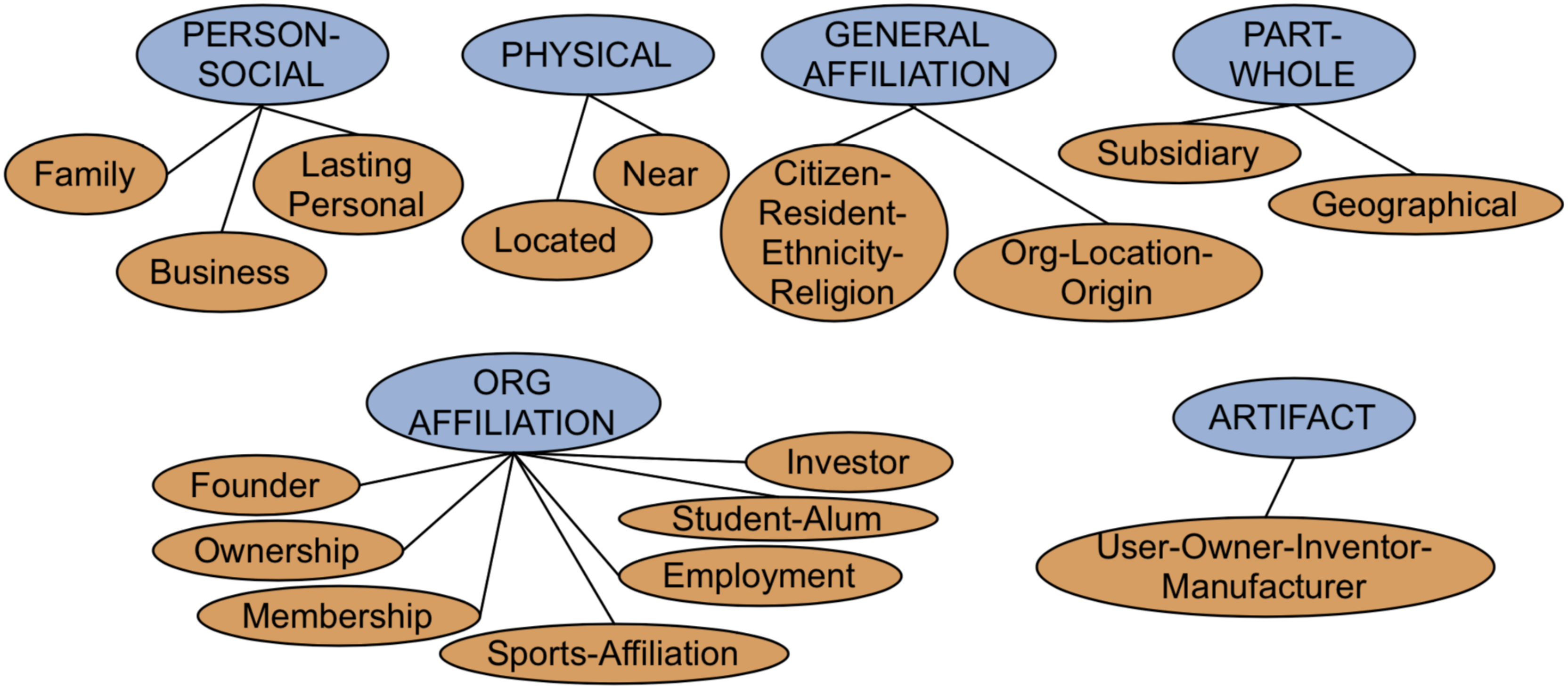
What is Relation Extraction

- What is it?
  - A fundamental task in Information Extraction
  - Definition: Given a sentence S with the annotated pairs of nominals e_1 and e_2, the goal is to identify the relation r from a predefined relation set $\boldsymbol{R}$ for the entity pair (e_1, e_2), written in the form of a triple (e_1, r, e_2) or (e_2, r, e_1)

E-Mail: jinhua.du@adaptcentre.ie

➢ Automated Content Extraction (ACE): NIST

- 17 relations from 2008 "Relation Extraction Task"



E-Mail: jinhua.du@adaptcentre.ie

Relation Types (closed domain)

➢ SemEval 2010 Task 8

- 9 relations without directionality (plus ‘other’)
- 19 relations with directionality (including ‘other’)

Labels	Train	Dev	Test
Entity-Origin(e_1, e_2)	454	114	211
Entity-Origin(e_2, e_1)	118	30	47
Entity-Destination(e_1, e_2)	675	169	291
Entity-Destination(e_2, e_1)	0	1	1
Component-Whole(e_1, e_2)	376	94	162
Component-Whole(e_2, e_1)	376	95	150
Product-Producer(e_1, e_2)	258	65	108
Product-Producer(e_2, e_1)	315	79	123
Instrument-Agency(e_1, e_2)	77	20	22
Instrument-Agency(e_2, e_1)	325	82	134
Content-Container(e_1, e_2)	299	75	153
Content-Container(e_2, e_1)	132	34	39
Cause-Effect(e_1, e_2)	275	69	134
Cause-Effect(e_2, e_1)	527	132	194
Message-Topic(e_1, e_2)	392	98	210
Message-Topic(e_2, e_1)	115	29	51
Member-Collection(e_1, e_2)	62	16	32
Member-Collection(e_2, e_1)	489	123	201
other	1,128	282	454

E-Mail: jinhua.du@adaptcentre.ie

Relation Types (open domain)

- Freebase: thousand relations/million entities

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

E-Mail: jinhua.du@adaptcentre.ie

Two Sub-tasks in RE

- Entities recognition
  - Name entities: Person, Organization, Location, Times, Dates, etc.
  - Domain-specific nouns: genes, proteins, diseases, financial terms, etc.

- Relation extraction
  - Located in, employed by, married to, etc.

E-Mail: jinhua.du@adaptcentre.ie

Two Categories of Modelling

Pipeline Modelling

Text → Entity Recognition → Relation Extraction

Mark Elliot Zuckerberg is an American computer programmer and Internet entrepreneur. He is a co-founder of Facebook, and currently operates as its chairman and chief executive officer.

PERSON: Mark Elliot Zuckerberg
LOCATION: American
ORGANISATION: Facebook

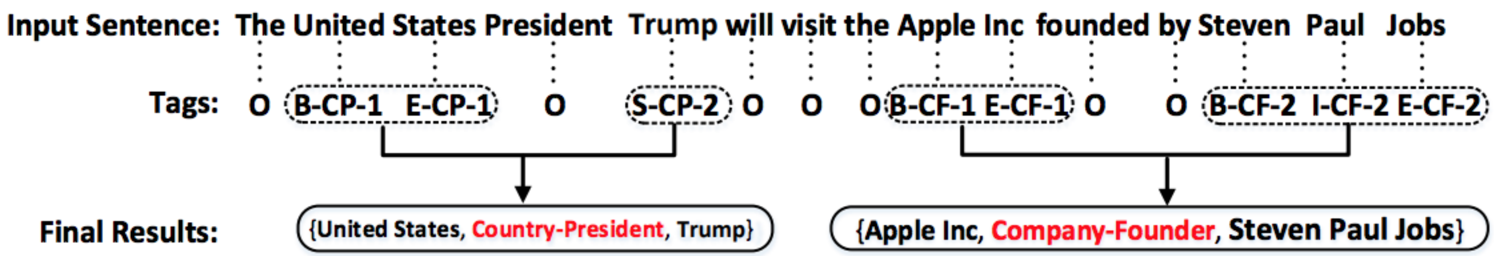
Mark Elliot Zuckerberg *is founder of* Facebook

E-Mail: jinhua.du@adaptcentre.ie

Two Categories of Modelling

Joint Modelling

Text → Relation Extraction



E-Mail: jinhua.du@adaptcentre.ie

Outline

- ✓ What is Relation Extraction
- ✓ Distant Supervision for RE
- ✓ Multi-Level Structured Attention
- ✓ Application

E-Mail: jinhua.du@adaptcentre.ie

- Problems

Human Annotated Data

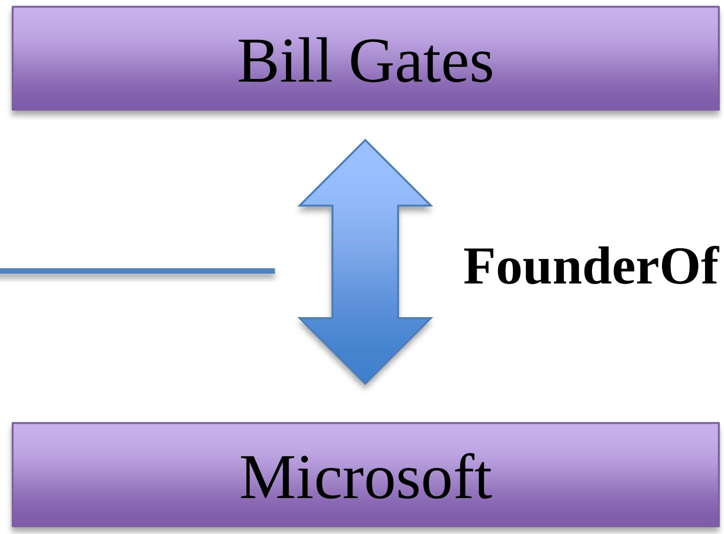
- **SemEval 2007**
- **SemEval 2010**
- **BioNLP Shared Task**
- **ADE-V2**

- Data is always important!
- Labeled data is not enough to train a good RE system with a good generalization capability

E-Mail: jinhua.du@adaptcentre.ie

- DS-RE:
  - Automatic labeling via knowledge bases, such as Freebase, DBpedia
  - It assumes that if one entity pair appearing in some sentences can be observed in a KB with a certain relationship, then these sentences will be labeled as the context of this entity pair and this relationship.
- Advantage
  - Effective and efficient method for automatically labeling large-scale training data
- Disadvantage
  - It introduces a severe mislabelling problem due to the fact that a sentence that mentions two entities does not necessarily express their relation in a KB

An Example



Relation	Instance
FounderOf	Microsoft was founded on April 4, 1975, by Bill Gates and Paul Allen in Albuquerque, New Mexico.
ChairmanOf	In February 2014 Gates stepped down as chairman from Microsoft but continued to serve as a board member.
Other/NA	Largely on the strength of Microsoft's success, Gates amassed a huge paper fortune as the company's largest individual shareholder.

E-Mail: jinhua.du@adaptcentre.ie

DS-RE is different from the traditional RE

- **It is a multi-instance learning problem**
- **It is a multi-label classification problem**
- **It contains a lot of noise**

Distant Supervision Data

- **New York Times**
- **Google's RE Corpus**
- **NIST KBP**
- **Portuguese DBpedia**

E-Mail: jinhua.du@adaptcentre.ie

Evaluation Metrics of RE

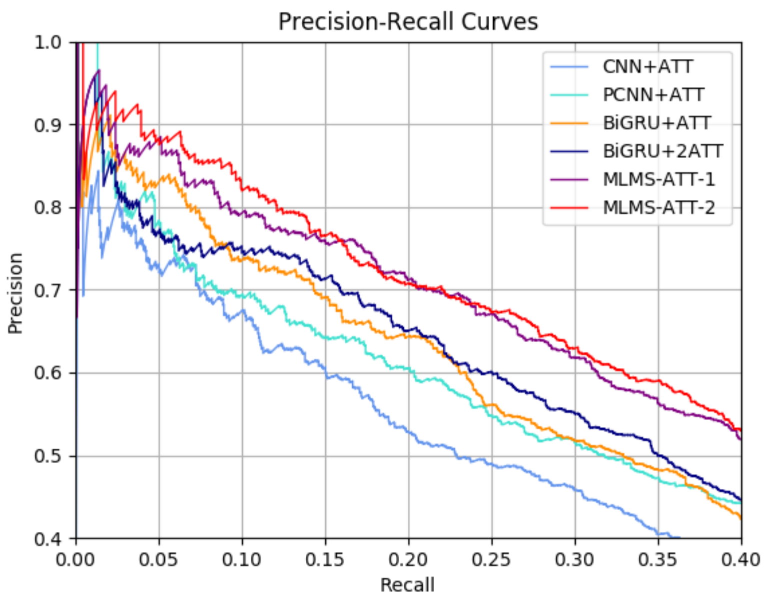
- Compute P/R/F1

$$P = \frac{\text{\# of correctly extracted relations}}{\text{Total \# of extracted relations}}$$

$$F_1 = \frac{2PR}{P+R}$$

$$R = \frac{\text{\# of correctly extracted relations}}{\text{Total \# of gold relations}}$$

- PR Curve
- AUC



E-Mail: jinhua.du@adaptcentre.ie

Outline

- ✓ What is Relation Extraction
- ✓ Distant Supervision for RE
- ✓ Multi-Level Structured Attention
- ✓ Application

E-Mail: jinhua.du@adaptcentre.ie

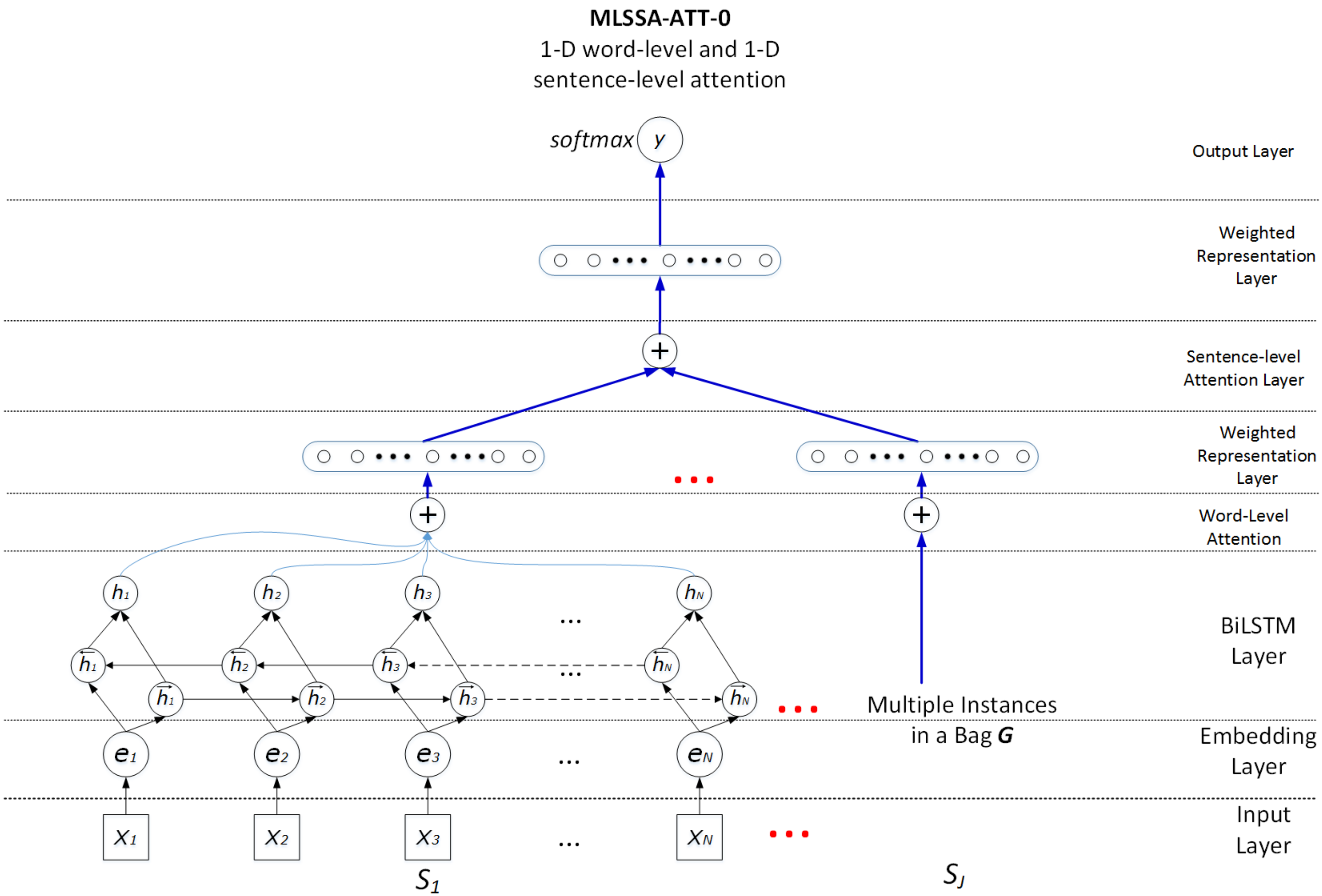
Deep Learning for Supervised RE

- Neural Networks for RE
  - Convolutional NN
  - Recurrent NN (LSTM, GRU, Bidirectional RNN)
  - Attention mechanism

- Like other NLP tasks, neural relation extraction has become the state-of-the-art.

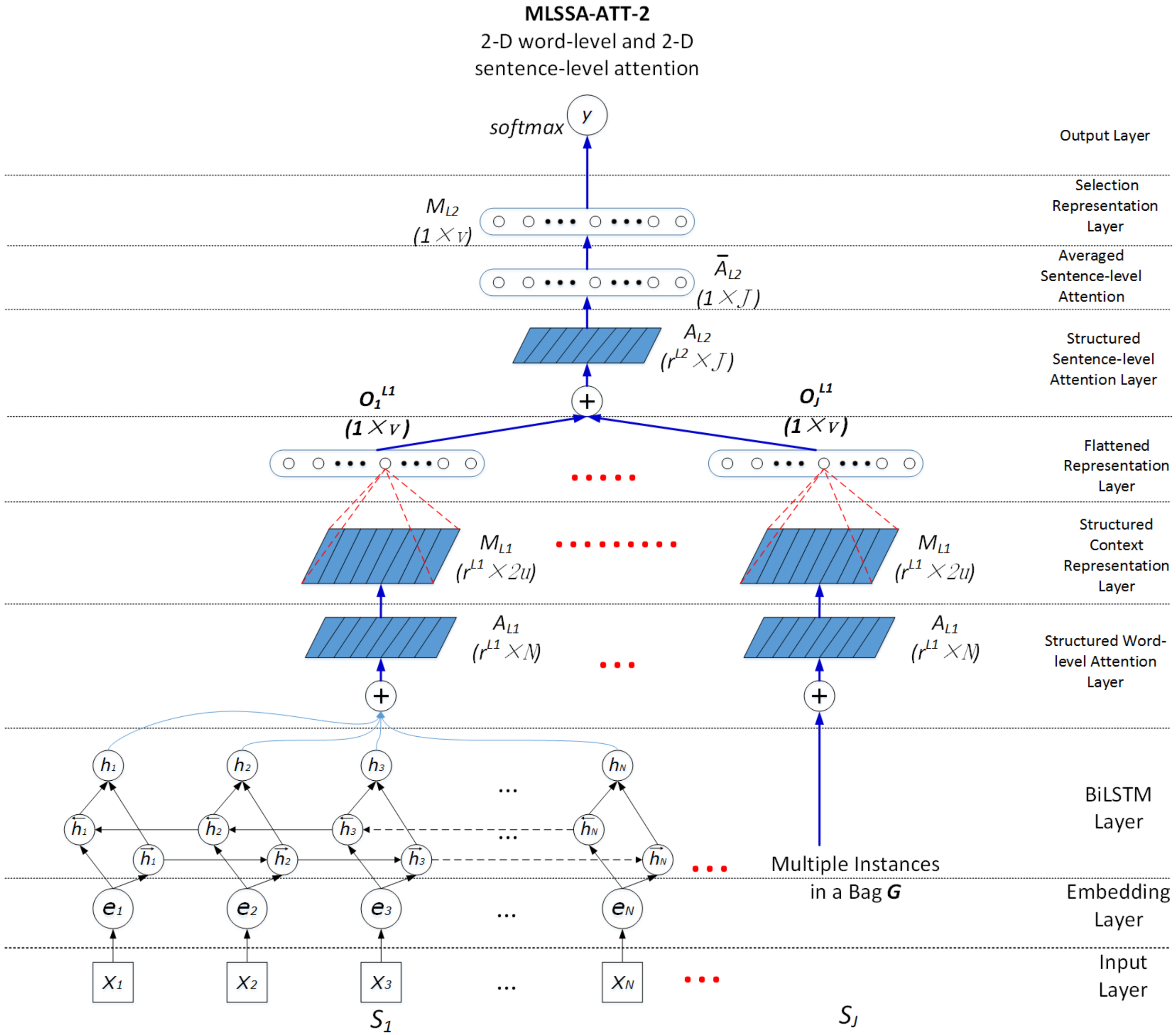
E-Mail: jinhua.du@adaptcentre.ie

Baseline: RNN with Muti-Level Attentions for DS-RE



E-Mail: jinhua.du@adaptcentre.ie

Our Work: Multi-Level Structured Self-Attention Mechanism



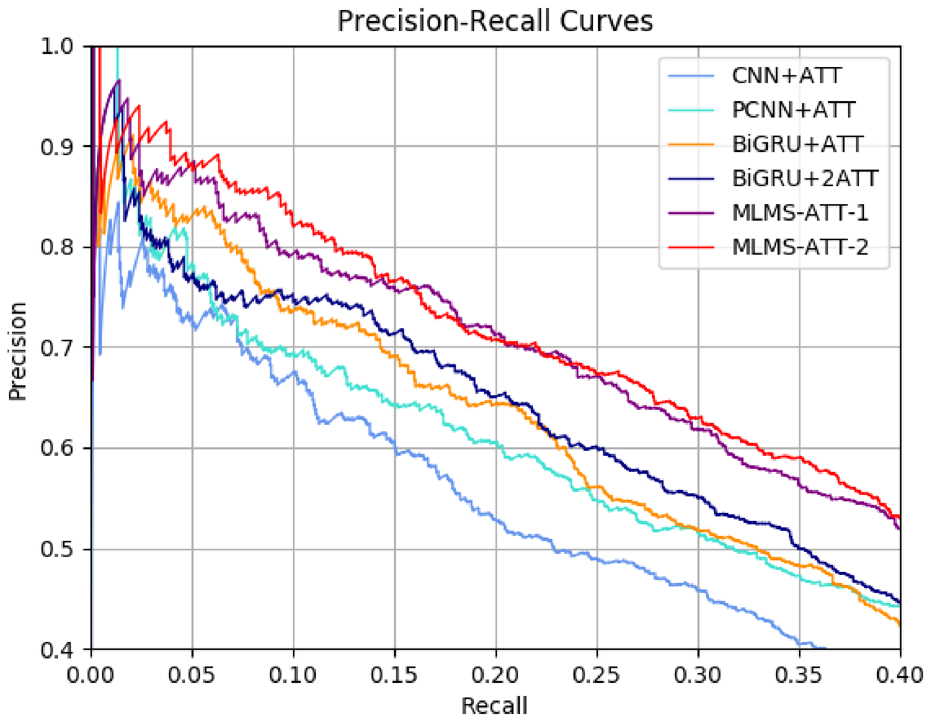
Jinhua Du et al. Multi-Level Structured Self-Attentions for Distantly Supervised Relation Extraction. Accepted by EMNLP 2018.

E-Mail: jinhua.du@adaptcentre.ie

Our Work in Accenture for RE

Relation Extraction with Multi-Level and Multi-Scale Self-Attention

- Motivation:
  - fully use contextual knowledge in the input sentence
  - select valid instances, and surpass the noisy instances
- Results:



E-Mail: jinhua.du@adaptcentre.ie

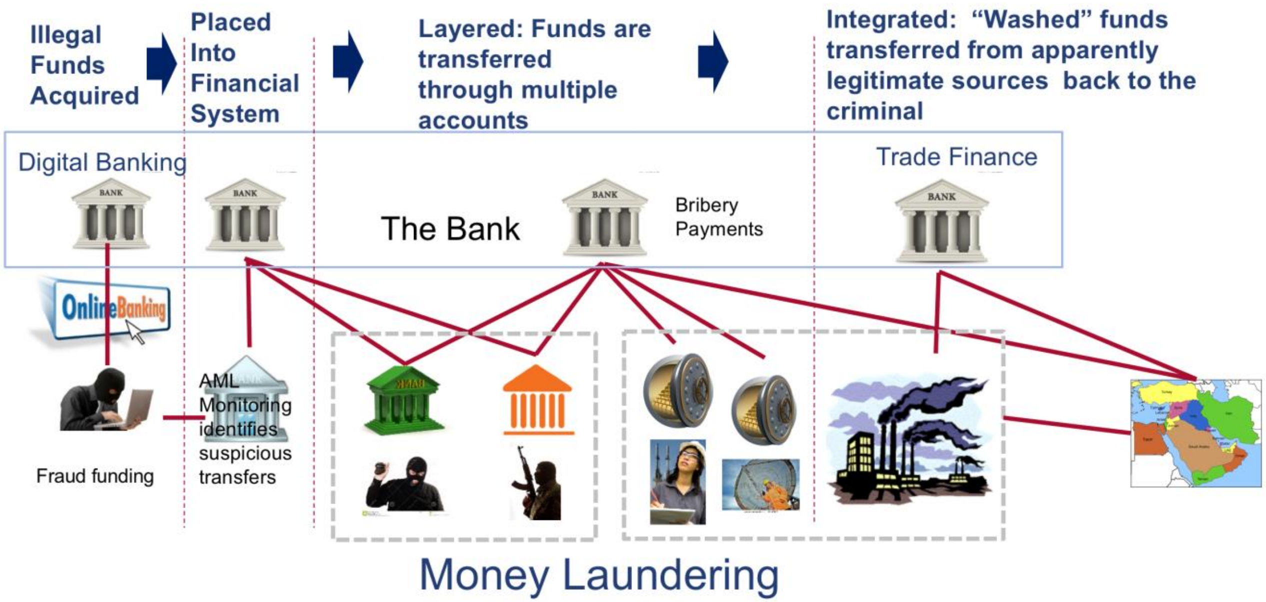
Outline

- ✓ What is Relation Extraction
- ✓ Distant Supervision for RE
- ✓ Multi-Level Structured Attention
- ✓ Application

E-Mail: jinhua.du@adaptcentre.ie

Applications of Relation Extraction: A Case Study: Anti-money Laundering Monitoring

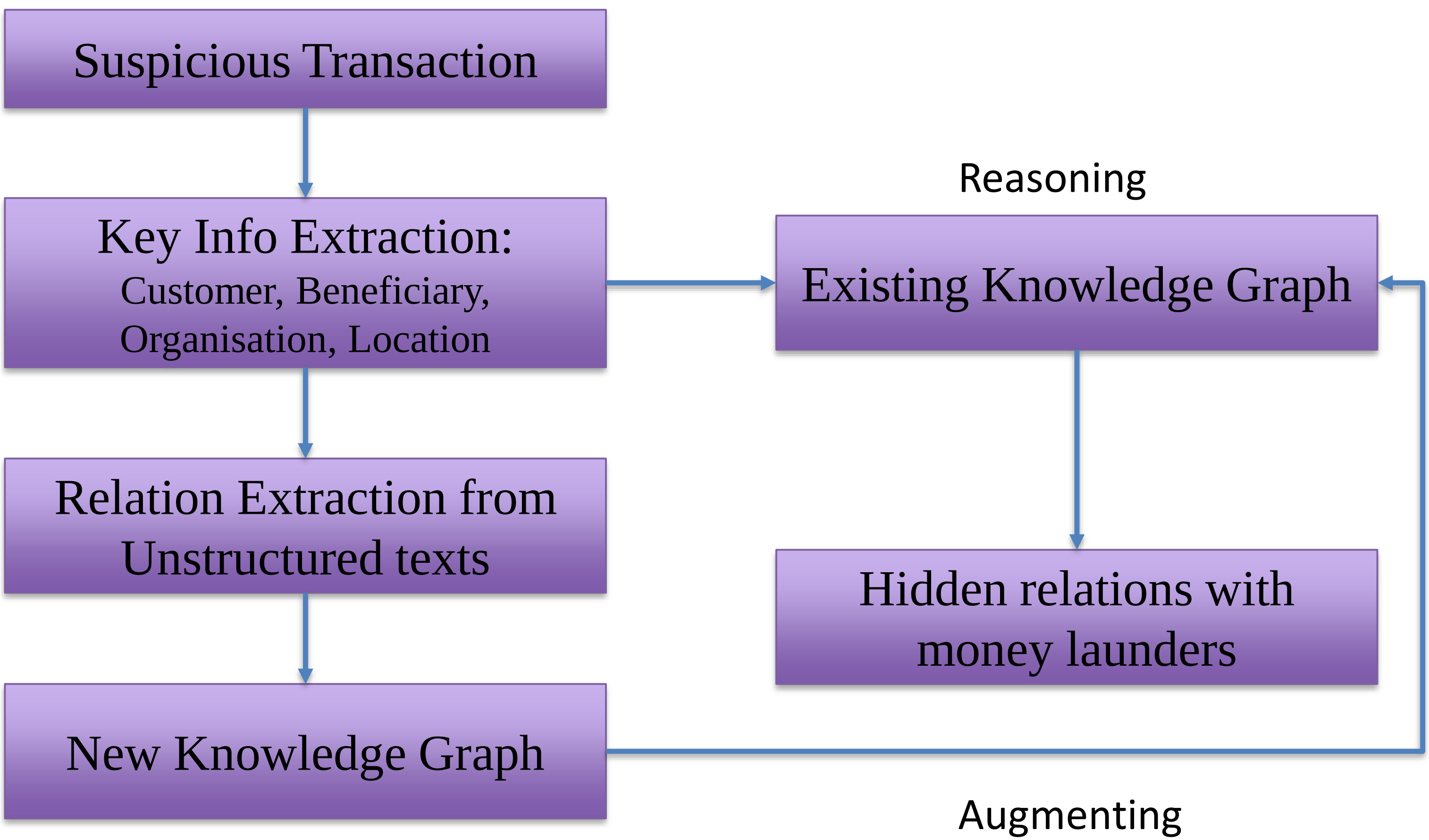
- Money Laundering: three stages
  - Placement
  - Layering
  - Integration



- Transaction Monitoring solutions: attempt to detect high risk or out of character funds transfers that may indicate money laundering activity

E-Mail: jinhua.du@adaptcentre.ie

➢ Basic Workflow



Jinhua Du et al. NextGen AML: Distributed Deep Learning based Language Technologies to Augment Anti Money Laundering Investigation. Proceedings of ACL 2018.

E-Mail: jinhua.du@adaptcentre.ie

THANKS FOR YOUR ATTENTION!

This research is supported by Science Foundation Ireland through the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University and Trinity College Dublin, and by SFI Industry Fellowship Programme 2016 (Grant 16/IFB/4490).

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.