COMP 3225

Natural Language Processing

Information Extraction

Stuart E. Middleton

sem03@soton.ac.uk

University of Southampton

Overview

- Information Extraction
- Relation Extraction Datasets
- Pattern-based and Supervised Relation Extraction

- <break - discussion point>

- Semi-supervised and Unsupervised Relation Extraction
- Evaluation of Relation Extraction
- Temporal and Event Extraction
- Knowledge-based Population

Information Extraction

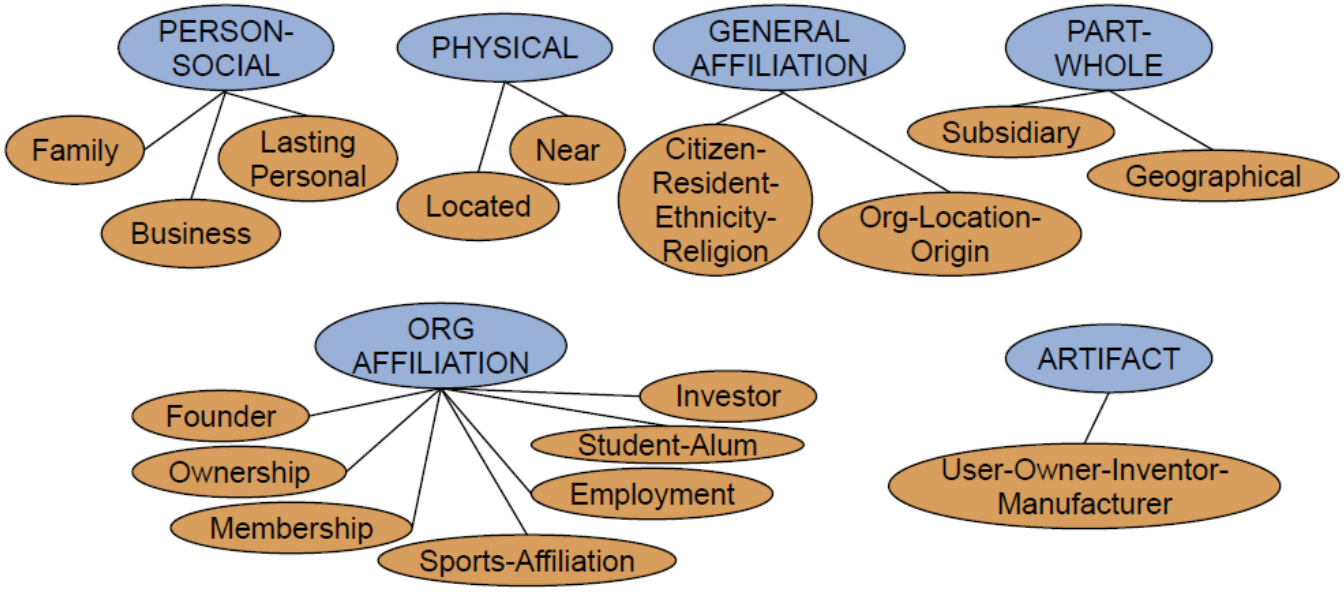
- Information Extraction turns unstructured text into structured data, such as a relational database or set of extracted tuples
- Relation Extraction (RE) finds semantic relations among text entities, such as parent-child, part-whole or geospatial relations
  - Relation words are typically action verb focused and often have noun phrase arguments
  - Knowledge-graphs can be constructed to encode relational information

    American Airlines supported the move by chairman Wagner
    American Airlines **supported** the move by chairman Wagner
    supported(American Airlines, the move by chairman Wagner)

- Event Extraction finds events in which entities participate
- Temporal Extraction finds times and dates
- Knowledge Base Population (KBP) populates knowledge bases from unstructured text using extracted information

Relation Extraction Datasets

- ACE relation extraction task - 17 relations, 7 entity types
  https://www.ldc.upenn.edu/collaborations/past-projects/ace



Relations	Types	Examples
Physical-Located	PER-GPE	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**XYZ**, the parent company of **ABC**
Person-Social-Family	PER-PER	**Yoko**'s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...

Relation Extraction Datasets

- ACE relation extraction task - 17 relations, 7 entity types
  https://www.ldc.upenn.edu/collaborations/past-projects/ace
- Wikipedia info boxes
  DBpedia 2 billion+ entries RDF https://wiki.dbpedia.org/datasets
  Wikidata 100 million+ entries JSON https://www.wikidata.org/wiki
- WordNet - is-a, hypernym, part-of relations
  100,000+ synsets https://wordnet.princeton.edu/
- TACRED dataset - 41 relations
  100,000+ examples JSON https://catalog.ldc.upenn.edu/LDC2018T24

Example	Entity Types & Label
Carey will succeed Cathleen P. Black, who held the position for 15 years and will take on a new role as chairwoman of Hearst Magazines, the company said.	PERSON/TITLE Relation: *per:title*
Irene Morgan Kirkaldy, who was born and reared in Baltimore, lived on Long Island and ran a child-care center in Queens with her second husband, Stanley Kirkaldy.	PERSON/CITY Relation: *per:city_of_birth*
Baldwin declined further comment, and said JetBlue chief executive Dave Barger was unavailable.	Types: PERSON/TITLE Relation: *no_relation*

Relation Extraction Datasets

- ACE relation extraction task - 17 relations, 7 entity types
  https://www.ldc.upenn.edu/collaborations/past-projects/ace
- Wikipedia info boxes
  DBpedia 2 billion+ entries RDF https://wiki.dbpedia.org/datasets
  Wikidata 100 million+ entries JSON https://www.wikidata.org/wiki
- WordNet - is-a, hypernym, part-of relations
  100,000+ synsets https://wordnet.princeton.edu/
- TACRED dataset - 41 relations
  100,000+ examples JSON https://catalog.ldc.upenn.edu/LDC2018T24
- SemEval 2010 task 8 - 9 relations
  10,000+ examples https://www.aclweb.org/anthology/S10-1006/

Pattern-based and Supervised Relation Extraction

- Pattern-based RE >> hand-crafted patterns
  - Hearst patterns - Hearst 1992
    Lexico-syntactic hand-crafted patterns

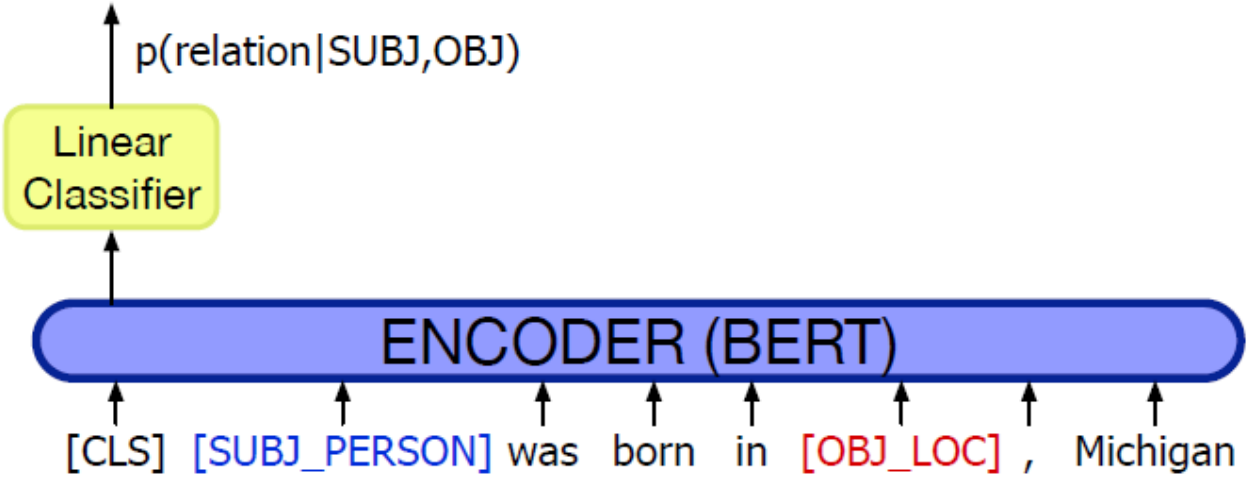
NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP}* {(or|and)} NP	European countries, especially France, England, and Spain

- Hand-crafted rules can be tailored to specific domain lexicons
- High precision, but often low recall and expensive to make rules

Cite: Hearst, M.A. Automatic Acquisition of Hyponyms from Large Text Corpora, COLING 1992, https://www.aclweb.org/anthology/C92-2082/

Pattern-based and Supervised Relation Extraction

- Supervised RE
  - Find pairs for named entities (in same sent) + classify relation word for pair
  - Any supervised ML >> logistic regression; random forest; RNN; Transformer
- Example >> BERT encoder + classifier
  - Source sent >> NER >> select NER pair >> create feature template
    entity 1 (phrase, NER type, context window around entity) features +
    entity 2 (phrase, NER type, context window around entity) features +
    features for words in-between entities
  - X = Feature set for the entity pair (embedding vector or one-hot vector)
  - Y = Prediction of relation (for provided entity pair) from 42 TAC relations (and a class for no relation)



Pattern-based and Supervised Relation Extraction

- Supervised RE
  - Find pairs for named entities (in same sent) + classify relation word for pair
  - Any supervised ML >> logistic regression; random forest; RNN; Transformer
- Example >> BERT encoder + linear classifier
  - Source sent >> NER >> select NER pair >> create feature template
    entity 1 (phrase, NER type, context window around entity) features +
    entity 2 (phrase, NER type, context window around entity) features +
    features for words in-between entities
  - X = Feature set for the entity pair (embedding vector or one-hot vector)
  - Y = Prediction of relation (for provided entity pair) from 42 TAC relations (and a class for no relation)
  - Transformer Model
    (a) replace SUBJ and OBJ with NER tags to avoid overfitting lexical terms
    (b) good idea to use RoBERTa or SPANbert pre-trained word embeddings, not vanilla BERT, as pre-training is done using single sentences not sentence pairs with a [SEP] in-between

Break

- Panopto Quiz - discussion point

- Why do named entities tend to overfit in deep learning supervised RE models?

There is a positive feedback loop between NER tagger and RE model
The RE labelled datasets are too small
The RE model cannot discover patterns to map named entities to arguments
The named entities are provided as input

Break

- Panopto Quiz - discussion point

- Why do named entities tend to overfit in deep learning supervised RE models?

There is a positive feedback loop between NER tagger and RE model >> No, although there will be some error propagation from NER tagging error

The RE labelled datasets are too small >> datasets are 10k-100k. there will not be examples of every possible named entity phrase (e.g. location), so model is likely to overfit to the dataset sample of the ones in the training set

The RE model cannot discover patterns to map named entities to arguments >> it could, but would need more training examples and deeper layers as this essentially would need to do two tasks NER + RE in one go

The named entities are provided as input >> Not relevant

Semi-supervised and Unsupervised Relation Extraction

- Semi-supervised RE using bootstrapping
  - Bootstrap using a (small) set of high-quality seed tuples (r, e1, e2)
  - Find sents that match >> extract patterns >> find new seed tuples

  - Confidence threshold for new tuples >> reduce semantic drift
  - Limit dep graph walk for new tuples >> reduce semantic drift
  - Generalize lexical terms in pattern >> avoid overfitting to corpus

function BOOTSTRAP(*Relation R*) **returns** *new relation tuples*

tuples ← Gather a set of seed tuples that have relation *R*
iterate
sentences ← find sentences that contain entities in *tuples*
patterns ← generalize the context between and around entities in *sentences*
newpairs ← use *patterns* to grep for more tuples
newpairs ← *newpairs* with high confidence
tuples ← *tuples* + *newpairs*
return tuples

Semi-supervised and Unsupervised Relation Extraction

- Distant supervision for RE
  - Use a knowledge-based (e.g. DBpedia) as a source of seed tuples (r, e1, e2)
  - Knowledge-base avoids semantic drift of bootstrapping
  - Text corpus >> NER tagger >> match entities with seed tuples
  - Add matches to training set as a feature set (with occurrence freq)

  - Generates a very large training set for supervised RE >> very noisy (low P)
  - GAN or incremental training approaches >> reduce noise during labelling

function DISTANT SUPERVISION(*Database D, Text T*) **returns** *relation classifier C*

foreach relation *R*
  foreach tuple (*e1,e2*) of entities with relation *R* in *D*
    sentences ← Sentences in *T* that contain *e1* and *e2*
    f ← Frequent features in *sentences*
    observations ← observations + new training tuple (*e1, e2, f, R*)
C ← Train supervised classifier on *observations*
return *C*

Semi-supervised and Unsupervised Relation Extraction

- Unsupervised RE extracts relations with no training data
  - Unsupervised RE is also called Open Information Extraction or OpenIE
  - ReVerb, ClauseIE, OpenIE5
    https://github.com/dair-iitd/OpenIE-standalone

  - Text corpus >> POS tagger >> Verb-based POS patterns
    >> Syntax & lexical constrained walk of tokens in sent (explore options)
    >> Relation tuple (r, e1, e2) + confidence

  - Unlimited number of relations and entity types (no training data)
  - Extracted phrases are not semantically grounded to a database entry
    >> harder for applications to use relations (e.g. inference)

V | VP | VW*P
V = verb particle? adv?
W = (noun | adj | adv | pron | det)
P = (prep | particle | inf. marker)

Evaluation of Relation Extraction

- Supervised RE
  >> P/R/F1 as we have a labelled corpus with a ground truth
- Semi-supervised or Unsupervised RE
  >> Random sample + human inspection
  compute precision of unique tuples
  >> P@Yield or P@R is precision at different levels of recall
  e.g. P@1000 extractions = precision of top 1000 extractions
  ranked by confidence

Temporal and Event Extraction

- Temporal Extraction extracts times, dates and durations
  - Can be absolute (easy) or relative (contextual to a reference point in text)

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of '77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

  - Rule-based systems look for lexical triggers (with POS tags) encoded as temporal expressions

Category	Examples
Noun	*morning, noon, night, winter, dusk, dawn*
Proper Noun	*January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet*
Adjective	*recent, past, annual, former*
Adverb	*hourly, daily, monthly, yearly*

Temporal and Event Extraction

- Sequence labelling approaches use BIO tags
  - Feature templates encode features sets for a supervised classifier
    e.g. CRF, Transformer

Feature	Explanation
Token	The target token to be labeled
Tokens in window	Bag of tokens in the window around a target
Shape	Character shape features
POS	Parts of speech of target and window words
Chunk tags	Base phrase chunk tag for target and words in a window
Lexical triggers	Presence in a list of temporal terms

Temporal and Event Extraction

- Temporal normalization maps temporal expression to a point in time or quantified duration
- Approaches for temporal normalization tend to be rule-based
  - Fully qualified expressions (which are often rare in free text)

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

  - Temporal anchors

    We started yesterday, and will continue through the weekend
    Start = yesterday = time statement was made - 1 day
    End = the weekend = the coming weekend after start

Temporal and Event Extraction

- Event extraction identifies mentions of events within text
  - Events in English often correspond to verbs, but not always (nouns can introduce an event)

[EVENT Citing] high fuel prices, United Airlines [EVENT said] Friday it has [EVENT increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately [EVENT matched] [EVENT the move], spokesman Tim Wagner [EVENT said]. United, a unit of UAL Corp., [EVENT said] [EVENT the increase] took effect Thursday and [EVENT applies] to most routes where it [EVENT competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

Temporal and Event Extraction

- Event extraction identifies mentions of events within text
  - Events in English often correspond to verbs, but not always (nouns can introduce an event)
- Sequence labelling approaches use BIO tags
  - Feature templates encode features sets for a supervised classifier
    e.g. CRF, Transformer

Feature	Explanation
Character affixes	Character-level prefixes and suffixes of target word
Nominalization suffix	Character-level suffixes for nominalizations (e.g., *-tion*)
Part of speech	Part of speech of the target word
Light verb	Binary feature indicating that the target is governed by a light verb
Subject syntactic category	Syntactic category of the subject of the sentence
Morphological stem	Stemmed version of the target word
Verb root	Root form of the verb basis for a nominalization
WordNet hypernyms	Hypernym set for the target

Knowledge-based Population

- Knowledge-base population (KBP)
  - Partial KB + large corpus of text → Populated KB
- Slotfilling completes all known information about a query entity
  - Text + NER tags → Supervised RE → rel(subj,obj) x N
  - Compute 0-hop and 1-hop slots for all query entities
  - Errors in hop-0 predictions can easily propagate to hop-1 predict
  - Zhang 2017 https://www.aclweb.org/anthology/D17-1004/

query entity: **Mike Penner**

hop-0 slot: *per:spouse* ------→ **Lisa Dillman**

hop-1 slot: *per:title* ------→ **Sportswriter**

(query) (fillers)

Penner *is survived by his brother,* ***John****, a* ***copy editor*** *at the* ***Times****, and his former wife,* ***Times sportswriter Lisa Dillman****.*

Subject	Relation	Object
Mike Penner	per:spouse	Lisa Dillman
Mike Penner	per:siblings	John Penner
Lisa Dillman	per:title	Sportswriter
Lisa Dillman	per:employee_of	Los Angeles Times
John Penner	per:title	Copy Editor
John Penner	per:employee_of	Los Angeles Times

Knowledge-based Population

- Knowledge-base population (KBP)
  - Partial KB + large corpus of text → Populated KB
- Slotfilling completes all known information about a query entity
  - Text + NER tags → Supervised RE → rel(subj,obj) x N
  - Compute 0-hop and 1-hop slots for all query entities
  - Errors in hop-0 predictions can easily propagate to hop-1 predict
  - Zhang 2017 https://www.aclweb.org/anthology/D17-1004/
- Entity linking links entity text references to unique KB entities
- Resources for KBP
  - Text Analysis Conference Knowledge Base Population (TAC KBP) https://www.ldc.upenn.edu/collaborations/past-projects/tac-kbp
  - Stanford KBP system https://nlp.stanford.edu/projects/kbp/

Required Reading

- Information Extraction
  - Jurafsky and Martin, Speech and Language Processing, 3rd edition (online) >> chapter 17

Questions

- Panopto Quiz - 1 minute brainstorm for interactive questions

  Please write down in Panopto quiz in **1 minute** two or three questions that you would like to have answered at the next interactive session.

  Do it **right now** while its fresh.

  Take a screen shot of your questions and **bring them with you** at the interactive session so you have something to ask.