MinIE – Minimized Facts for Open Information Extraction

Kiril Gashteovski

University of Mannheim
Chair of Practical Informatics I

Motivation

- $\sim 90\%$ of the world's data is held in unstructured formats
- Can we make this knowledge accessible?

Open Information Extraction (OIE)

- Extract relations and their arguments from natural language text in unsupervised manner
- In its simplest form, a triple of
  - Subject (S)
  - Relation (R)
  - Object (O)
- Example input sentence:
  - "AT&T, which is based in Dallas, is a telecommunication company."
- Possible facts expressed in this sentence
  - "AT&T" "is based in" "Dallas"
  - "AT&T" "is" "telecommunication company"

State of the art: ReVerb and OLLIE

ReVerb

- Extract verb mediated relations
- *"Early astronomers believed that the earth is the center of the universe."*
  - *("the earth","be the center of", "the universe")*

OLLIE

- Successor of ReVerb
- Bootstrapping to learn other relation patterns
- Context around the triple: attribution
  - *("the earth","be the center of", "the universe")*
  - Attribution: Early astronomers believed
- Context around the triple: clausal modifier – dependent clause modifying the main extraction
  - *"If he wins five key states, Romney will be elected President."*
  - *("Romney", "will be elected", "President")*
  - ClausalModifier: (if; he wins five key states)

State of the art: ClausIE

- ClausIE: clause based OIE system
- Detect clauses and extract propositions from them
  - exactly 7 clause types in English (SV, SVA, SVO, SVC, ...)
- "Donald Trump is the president of the United States."
  - clause type: SVC(A)
  - ("Donald Trump"; "is"; "the president")
  - ("Donald Trump"; "is"; "the president of the United States")

State of the art: NestIE

- Nested representations
- Sentence: *"After giving 5,000 people a second chance at life, doctors are celebrating the 25th anniversary of Britain's first heart transplant."*
- Extractions:
  - **P1:** (doctors, are celebrating, the 25th anniversary of Britain's first heart transplant)
  - **P2:** (doctors, giving, second chance at life)
  - **P3:** *(P1, after, P2)*

Common problems with OIE

- Relations can be too short
  - just verbs, making them highly polysemous
  - e.g. *"make"* has 49 meanings in WordNet
    v.s. the more informative *"make a deal with"*
- Arguments/relations can be overly specific
  - e.g. *"the extraordinary Richard Feynman"*
  - e.g. *"make a very good deal with"*

("'The great R. Feynman"; "worked jointly with"; "F. Dyson")

("Richard Feynman"; "worked with"; "Freeman Dyson")

- Lack of expressiveness of a triple
  - *("North Korea", "attack", "Guam")*
  - is this certain or merely a possibility?
  - who/what is the source of this triple?

MinIE - OIE system trying to tackle these challenges

MinIE: minimize by annotating and structuring information

- **Factuality:** information about the triple's polarity and modality
- **Polarity:** is the triple positive (+) or negative (–)?
  - *("H. Clinton"; "is not president of"; "U.S.")* $\Rightarrow$ *("H. Clinton"; "is president of"; "U.S.")* **(–)**
- **Modality:** is the triple a certainty (CT) or a possibility (PS)?
  - *("Bill Cosby"; "may go to"; "jail")* $\Rightarrow$ *("Bill Cosby"; "go to"; "jail")* **(PS)**
- **Attribution:** the supplier of the information and its factuality
  - *("**D. T.**"; "said that"; "B. O. may have been born in Kenya")* $\Rightarrow$ *("B. Obama"; "have been born in"; "Kenya")* **(+, PS)**
    Attribution: (Donald Trump, +, CT)

MinIE: minimize by annotating and structuring information

- **Quantities:** phrases expressing an amount of something
  - e.g. 9 cats, all cats, almost about 100 cats
    ⇒ *QUANT cats*
- *"F.B.I. official said that at least two e-mails were probably not marked as confidential."*
  - *("Q_1 e-mails"; "were marked as"; "confidential")*
    Factuality: (PS, –)
    Attribution: (F.B.I. official, (+, CT))
    Quantities: Q_1 = at least two;

MinIE on WikiPedia: example triples

- Combinations of **factuality + frequency**
  - *("Barack Obama";"be";"president")* **(+, CT)**: 8,930
  - *("Barack Obama";"be";"president")* **(–, CT)**: 1
  - *("Barack Obama";"be";"president")* **(–, PS)**: 1
- Consider the source of information: **factuality + attribution**
  - *("Barack Obama"; "be born in"; "U.S.")* **(–, CT)**: 1
    - **attribution:** Orly Taitz **(+, CT)** ← conspiracy theorist
  - *("Barack Obama"; "be born in"; "U.S.")* **(+, CT)**: 1
    - **attribution:** Joshua A. Wisch **(+, CT)** ← special assistant to attorney general of State of Hawaii
- How reliable is the **attribution**?
  - E.g: what is attributed to **Donald Trump**?
  - *("Donald Trump"; "be"; "pro-choice")* **(+,CT)**: 1
  - *("Donald Trump"; "be"; "pro-life")* **(+,CT)**: 1
  - *("Barack Obama"; "be born in"; "Kenya")* **(+, PS)**: 1
  - *("Barack Obama"; "be born in"; "United States")* **(+, CT)**: 1

MinIE on WikiPedia: relating results to a Knowledge Base

- Facts which can be found in both DBPedia and MinIE's output
  - Example: Carl Benz's birth place
  - MinIE: *("Karl Benz"; "was born in"; "Mühlburg")* **(+,CT)**: 1
  - DBPedia: *(dbp:Carl_Benz; dbp:birthPlace; dbp:Mühlburg)*
- Facts with relations not found in DBPedia
  - Example: Carl Benz's invention
  - MinIE: *("K. B."; "be inventor of"; "automobile")* **(+,CT)**: 2
  - DBPedia/YAGO: no relation "be inventor of"
  - MinIE: "be inventor of" appears as relation in 6,316 triples
- MinIE is schema-free, so it could be useful for
  - discovering new facts for already established entities in a KB
  - discovering new entities/relations for a KB

MinIE: minimize by dropping overly specific words

- Identify and remove words that are considered overly specific without damaging the meaning of the phrase.

Input sentence:
"The great Richard Feynman worked jointly with Freeman Dyson."

Output triple:
*("**The great** R. Feynman"; "worked **jointly** with"; "F. Dyson")*

$\Downarrow$ minimize

("Richard Feynman"; "worked with"; "Freeman Dyson")

$\downarrow$ $\downarrow$

"great" **"jointly"** $\Rightarrow$ keep dropped words as annotations

MinIE: minimize by dropping overly specific words

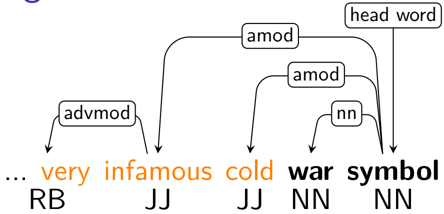
- Danger: over-minimizing might change the semantics

Is this the place to learn about mining?

MinIE: several modes of minimization

- Minimization modes with different levels of aggressiveness
  - effectively control the minimality-precision trade-off
- **MinIE-C** (Complete Mode)
  - prunes all the extractions that contain subordinate clauses
  - does not otherwise modify the annotated extractions
- **MinIE-S** (Safe Mode)
  - drops words that are considered to be safe to drop
  - e.g. determiners, adverbs modifying verbs, ...
- **MinIE-D** (Dictionary Mode)
  1. run the safe mode on a corpus
  2. construct a dictionary of collocations $\mathcal{D}$ with frequent args/rels
  3. find candidate words for dropping (e.g. adj. modifying NPs)
  4. generate sub-constituents (see next slide)
  5. drop candidates not found in the dictionary
- **MinIE-A** (Aggressive Mode)
  - all words for which we are not sure if they need to be retained are dropped

Sub-constituent generation



Possibile sub-constituents (22 combinations):

- combinations from **stable constituents** (1 combination)
  - war symbol
- combinations from one dependency path (9 combinations)
  - [very] infamous war symbol, [very] infamous war, [very] infamous symbol (6 combinations)
  - **cold war**, cold symbol, cold war symbol (3 combinations)
    - "cold war" found in dictionary ⇒ "cold" is marked as "stable" and is not dropped
- combinations from several dependency paths (12 comb.)
  - [very] infamous cold war, [very] infamous cold symbol, [very] infamous cold war symbol
  - cold [very] infamous symbol, cold [very] infamous war, cold [very] infamous war symbol

MinIE: several modes of minimization

Input: *"The big celebration on the campus lasted for 2 days."*

Output:

("*The big celebration on the campus*"; "*lasted for*"; "Q_1 *days*") **MinIE-C**

$\Downarrow$

("*big celebration on campus*"; "*lasted for*"; "Q_1 *days*") **MinIE-S**

$\Downarrow$

("*celebration on campus*"; "*lasted for*"; "Q_1 *days*") **MinIE-D**

$\Downarrow$

("*celebration*"; "*lasted for*"; "*days*") **MinIE-A**

Experiments: triples length and annotations

- Dataset: 10,000 random sentences from the N. Y. Times Corpus
- Triples length (word count)
- Number of triples with annotations

System	triples length ($\mu \pm \sigma$)	with attributions	with neg. polarity	with possibility	with quantities
OLLIE	9.9 ± 5.8	6.8%	-	-	-
ClausIE	10.9 ± 7.0	-	-	-	-
Stanford OIE	6.6 ± 3.0	-	-	-	-
MinIE-C	8.3 ± 4.9	**10.8%**	**3.8%**	**10.1%**	17.6%
MinIE-S	7.2 ± 4.2	**10.8%**	3.7%	9.9%	**17.8%**
MinIE-D	7.0 ± 4.1	10.7%	3.7%	10.0%	**17.8%**
MinIE-A	$\mathbf{4.7 \pm 1.9}$	**10.8%**	**3.8%**	9.7%	1.9%

μ – mean word count per triple
σ – standard deviation for word counts per triple

Experiments: number of extracted triples

- Dataset: 10,000 random sentences from the N. Y. Times Corpus
- **Redundant triple:** a triple t_1 is redundant if it appears as subsequence in some other triple t_2 produced by the same extractor from the same sentence
  - **Input:** "Richard Feynman lived in California in 1970."
    Output:
  - ("Richard Feynman"; "lived in California in"; "1970") → non-redundant
  - ("Richard Feynman"; "lived in"; "California") → redundant

System	# non-redundant extractions	# with redundant extractions
OLLIE	20,557	24,316
ClausIE	36,173	**58,420**
Stanford OIE	16,350	43,360
MinIE-C	**37,465**	47,637
MinIE-S	37,093	45,492
MinIE-D	36,921	45,318
MinIE-A	36,474	42,842

μ – mean word count per triple
σ – standard deviation for word counts per triple

Experiments: precision of labeled extractions

- Datasets: random samples of 200 sentences from
  - Wiki – Wikipedia
  - NYT – the New York Times Corpus
- Measures
  - factual precision: the fraction of correct triples out of all extractions
  - attribution precision: the fraction of correct triples that have correct attributions

System	Factual Precision (NYT/Wiki)	Attr. Precision (NYT/Wiki)
OLLIE	0.61 / 0.50	0.90 / **0.97**
ClausIE	0.61 / 0.63	-
Stanford OIE	0.50 / 0.43	-
MinIE-C	**0.75 / 0.75**	**0.94 / 0.97**
MinIE-S	**0.75** / 0.74	0.93 / 0.96
MinIE-D	0.74 / 0.73	0.93 / 0.96
MinIE-A	0.59 / 0.61	0.93 / 0.97

Experiments: recall of labeled extractions

- Datasets: random samples of 200 sentences from
  - Wiki – Wikipedia
  - NYT – the New York Times Corpus
- Measures
  - recall: the number of correct triples

System	*NYT*		*Wiki*	
	#non-redundant (correct/total)	#w/ redundant (correct/total)	#non-redund. (correct/total)	#w/ redund. (correct/total)
OLLIE	246/414	302/497	229/479	284/565
ClausIE	505/821	**792/1300**	424/704	628/1002
Stanford OIE	178/342	530/1052	217/398	**651/1519**
MinIE-C	**581/785**	727/970	**500/666**	635/851
MinIE-S	574/781	690/924	489/661	602/816
MinIE-D	569/777	681/916	486/669	593/816
MinIE-A	439/753	505/860	401/658	474/783

Experiments: comments

- Factual precision dropped when we use more aggressive modes
- The drop in precision between MinIE-C and MinIE-D was quite low, even though extractions get shorter
- The aggressive minimization of MinIE-A led to a more severe drop in precision
- For attribution precision, most of the sentences in our samples did not contain attributions; these numbers thus should be taken with a grain of salt
- For all modes, errors in dependency parsing transfer over to errors in MinIE
- MinIE-D sometimes drops adjectives which in fact form collocations (e.g., "*assistant* director") with the noun they are modifying
  - this happens when the collocation is not present in the dictionary; better collocation dictionaries may address this problem.

Take aways

- Extracting triples out of unstructured text
- Improve content by adding annotations on them
  - factuality: is the triple positive/negative?
    is it certainty/possibility?
  - attribution: who said what and how?
  - quantities: {*9 cats*, *almost 10 cats*, *few cats*} ⇒ *QUANT cats*
- Minimize the relations and arguments
  - e.g. *"Richard Feynman"* not *"the great Richard Feynman"*
  - e.g. *"made deal with"* not *"made a very good deal with"*
- Danger of over-minimization
  - e.g. *"data mining"* not *"mining"*
- Different levels of minimization: complete, safe, dictionary and aggressive