PRISMATIC - Automatic knowledge extraction from documents in IBM Watson

Paper Presentation

Antoni Maciąg, May 2022

The task

- Given a corpus of e.g. Wikipedia articles, extract knowledge that can later be used offline to answer Jeopardy! questions.
- To extract knowledge, first we need to define it. Knowledge and are too versatile to use a database of any sort.
- What we want to extract from the sentence ‘Nabokov wrote Lolita in 1953’ is that there was an action of ‘writing’, whose subject was ‘Nabokov’, object was ‘Lolita’, and time was the year ‘1953’.
- This can be expressed well with syntactic relations like subject, object, complement…
- If we see many sentences about Nabokov writing something, we also want to infer that Nabokov was probably a writer.

Concepts

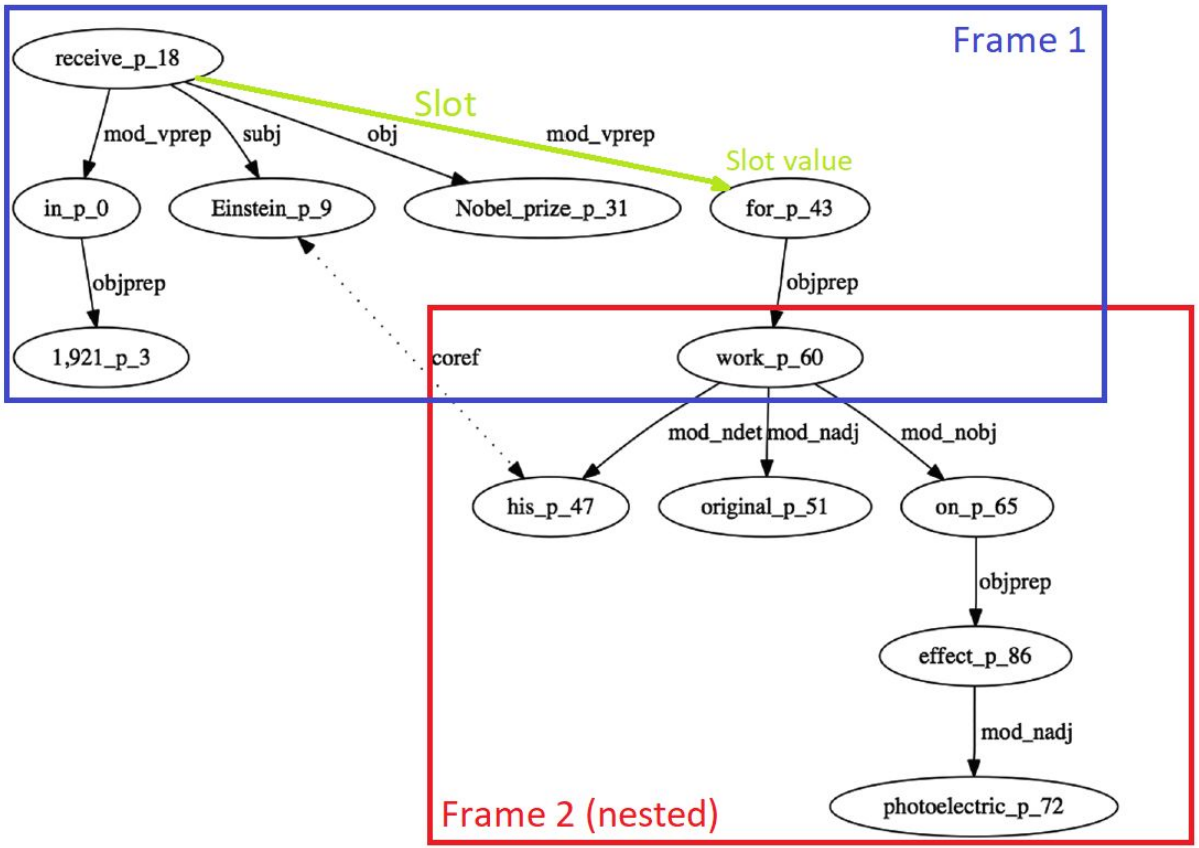
- Slot - a binary relation (but think of tuples). Can be syntactic, related to types (is of type, type is subtype of) or an is-a relation.
- Slot value - the entity in a relation.
- Frame - a set of slots and values for a given sentence.
- Frame projection - a frequent and important subset of a frame, for example V-OT (verb - object type) or S-V-O (subject - verb - object). Here an example of an S-V-O-OT frame:

  {<verb, “write”>, <subject, “nabokov”>, <object, “lolita”>, <object-type, “Book”>}

- And projected to S-V-O:

  {<verb, “write”>, <subject, “nabokov”>, <object, “lolita”>}

In 1921, Einstein received the Nobel Prize for his original work on the photoelectric effect.



Methodology

- First, the shallow knowledge extraction phase, resulting in knowledge represented as projected frames.
- Stages: parsing, frame extraction, frame projection.
- Then, the obtained frames are used in the aggregation phase, producing more knowledge on the basis of aggregate statistics like frequency and conditional probability of frames.

Parsing

- An instance of dependency parsing, a well-known problem in NLP.
- The preexisting English Slot Grammar parser used here.
- Parsing enhanced with the recognition of some relations not recognized by ESG, e.g. is-a.
- Also enhanced with type detection based on a Named Entity Recognition model. This model is independent from the rest of the system and can be substituted.
- Proper nouns found and marked.

Frame extraction

- Frames are trees or subtrees. The limit for frame depth is two levels, but can be nested.
- The depth limit helps reduce the number of frames with parse errors.
- Also, bigger frames are harder to analyze and not really necessary, because frame projections are small.
- Extensional frames - values are instances: {<verb, “write”>, <object, “Lolita”>}
- Intentional frames - values are types: {<verb, “write”>, <object-type, “Book”>}

Frame projection

- We try to project every frame onto types like S–V-O, N-Isa-Mod. V-OT.
- For example, the frames constructed for both the sentences “Nabokov wrote Lolita” and “Witkacy wrote Nienasycenie” will be projected onto V-OT as {<verb, “write”>, <object-type, “Book”>}.
- Thus, we will learn that “books” are “written”, i. e. the object of the “write” action is a Book.
- In addition, we can specify projection constraints. For example, if the S and O in S-V-O are constrained to be proper nouns, we will be gathering concrete and useful data in the form of extensional frames, which can later be used directly to answer questions like “who wrote Lolita?”

Aggregate statistics

- Frequency - if we see a lot of frames of the form {<subject, “Freud”>, <verb, “take”>, <object, “cocaine”>}, we can conclude that Freud was known for taking cocaine.
- Conditional probability between frames. If 80% of frames of the form {<subject, “Freud”>, <verb, “take”>} co-occur with frames of the form {<subject, “Freud”>, <verb, “take”>, <object, “cocaine”>}, we can conclude that if Freud took something, it was probably cocaine (but in 20% of cases, something else).
- Normalized pointwise mutual information - ranging from -1 to 1, it gauges show often frames co-occur in comparison to their total popularity. For example, if Freud is well-known for taking cocaine, but a huge number of other people also are, then the co-occurrence between {<verb, “take”>, <object, “cocaine”>} and {<subject, “Freud”>} is low. However, the co-occurrence between {<subject, “Freud”>, <verb, “take”>} and {<object, “cocaine”>} is still high.

Evaluation

- The task is non-standard, there is no established evaluation metric like there are e.g. for popular NLP tasks, even precision and recall are hard to estimate because ‘frame correctness’ is hard to define.
- The best that the authors could come up with was number of frames generated per sentence (1.3) and the percentage of named entities included in any frame (94%).

Usage

Type inference

- If we know that the object of writing is a Book, but the object of directing is a Film, we can correctly discern that in the sentence “Nabokov wrote Lolita”, Lolita is a book, but in “Adrian Lyne directed Lolita”, Lolita is a movie.
- This is obviously useful for answering questions in which ambiguous entity names (like Lolita) appear.
- Also coreference resolution: in “Lolita was published by Olympia Press before it was translated into Russian”, we can conclude that the “it” refers to the book rather than the publisher, because publishers do not tend to be translated.

Type coercion

- Suppose we have a question: “This playwright is considered a leading figure in German literature” and two candidate answers: Goethe and Schiller. We can conclude that Goethe is more likely, because he is more often referred to as “playwright”.

Generating candidate answers

- Expanding upon that: Inowing that we are looking for a German playwright, the candidate answer “Goethe” can be independently generated by PRISMATIC, because Goethe is one of the most frequently referenced German playwrights in our corpus.
- Results show that 1 in 57 candidates generated by PRISMATIC is correct, compared to an average of 1 in 134 for the other candidate generation modules.

Finding links

- Let us consider the question: “Neither he nor Vergil are daunted by the warning sign near the entrance to Hell”.
- The correct answer, Dante, is linked with Vergil and Hell by an entity which does not appear in the sentence - the Divine Comedy.
- If we know that Vergil and Hell can be associated with this work, we can suppose that Dante is a possible answer.

So is it outdated yet?

- This is NLP, and NLP is nothing like it was in 2010.
- The article suggests that successful knowledge extraction for QA has to involve parsing, NER, typing, detecting relationships, co-reference resolution…
- SOTA QA models for short texts do none of these things.
- Watson was built for something else. But given that models based on extensive rule programming have been all but driven out by deep learning, we can suppose that this approach has no future.
- Individual tasks like dependency parsing and NER have improved considerably since 2010.
- Interestingly enough, the authors fail to mention that we are still kinda bad at co-reference resolution (and more so back then).

Thank you