Cross-lingual Open Information Extraction with Neural Seq2Seq Models

Sheng Zhang, Kevin Duh, Ben Van Durme
(Johns Hopkins University)
Analyst: User with complex information need
Aid Worker: Which locations have immediate need?
Documents with these persons: \{ \}
Documents with a: 

CausalRelation(DISEASE, SYMPTOM)
Task Formulation

• Cross-lingual:
  – analyst speaks English, but document collection is in other languages

• Cross-lingual Information Retrieval?
  – Document unit is too large

• Cross-lingual Question Answering?
  – Difficulty in formulating questions

• Cross-lingual Information Extraction?
  – Close, but no fixed ontology
Information Extraction vs. Open Information Extraction

IE

Co-founder (Bill Gates, Microsoft)
Director-of (MacLorrance, Ciao)
Employee-of (MacLorrance, Ciao)
...

Open IE

(Bill Gate, be, Microsoft co-founder)
(Bill Gates, stepped down as, CEO)
(Bill Gates, was included in, the Forbes wealthiest list)
(Bill Gates, was, the wealthiest)
(IBM, would buy, Ciao)
(MacLorrance, has occupied, the corner office of the Hopkinton)
...

<table>
<thead>
<tr>
<th>IE</th>
<th>Open IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Sentences + Labeled relations</td>
</tr>
<tr>
<td>Relation</td>
<td>Specified relations in advance</td>
</tr>
<tr>
<td>Extractor</td>
<td>Specified relations</td>
</tr>
</tbody>
</table>

Figure from: Duc-Thuan Vo and Ebrahim Bagheri. (2016) “Open Information Extraction”
Input: Chinese sentence

克里斯想造一艘船。

Cross-lingual Open Information Extraction

Output: A set of English tuples, e.g. Relation(arg1,arg2)

Chris wants Chris build a boat
Cross-lingual *Open* Information Extraction

Visualization → Query →

RelationA(arg1, arg2)
RelationB(arg1, arg2)
RelationC(arg1, arg2)
Assumptions

1. Training data: Chinese-English bitext
2. Monolingual Open IE system in English
Monolingual Open IE System
PredPatt: https://github.com/hltcoe/PredPatt

• Based on Universal Dependencies
• Rules for:
  1. identifying predicate root and argument root:
     e.g. nsubj(s, v), dobj(o, v)
  2. resolving arguments:
     Chris expects to visit Pat → nsubj(Chris,visit)
     Chris likes to sing and dance → nsubj(Chris,dance)
  3. phrase extraction:
     PredPatt finds structure in text → ?a finds ?b in ?c
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

?a is/are 61 years old
  ?a: Pierre Vinken

?a will join ?b as ?c ?d
  ?a: Pierre Vinken, 61 years old
  ?b: the board
  ?c: a nonexecutive director
  ?d: Nov. 29

?a is/are nonexecutive
  ?a: a director
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.

?a is chairman of ?b
?a: Mr. Vinken
?b: Elsevier N.V.

?a is/are the Dutch publishing group
?a: Elsevier N.V.
Input: Chinese sentence

克里斯想造一艘船。

Cross-lingual Open Information Extraction

Output: A set of English tuples, e.g. Relation(arg1,arg2)
Chris wants to build a boat

克里斯 想 造 一艘 船。
Chris wants to build a boat.
JOINT SOLUTION

克里斯想造一艘船。

Cross-lingual Open IE

Chris wants to build a boat.
Neural Sequence-to-Sequence Model

\[ c = \text{RNN}(x) \]

\[ y \mid c \sim \text{RNNLM}(c) \]

Cho (2014); Sutskever (2014); Bahdanau (2015)
Neural Sequence-to-Sequence Model

"Attention model": Context vector $c$ is dynamic

Beginnings

START

Aller  Anfang  ist  schwer  STOP
Neural Sequence-to-Sequence Model

“Attention model”: Context vector $c$ is dynamic
Neural Sequence-to-Sequence Model

“Attention model”: Context vector $c$ is dynamic
Neural Sequence-to-Sequence Model

“Attention model”: Context vector $c$ is dynamic
Linearized OpenIE output as target

\[
[(\text{Chris}\!:\!a_h)\ \text{wants}:\!p_h\ [(\text{Chris}\!:\!a_h)\ \text{build}:\!p_h\ (a:a\ \text{boat}:a_h)]]
\]
Experiment Setting

- 1 million sentence English-Chinese bitext (GALE project; mixed domain)

# Predicates in OpenIE output

![Graph showing sentence length vs. number of predicates in OpenIE output]
Chris wants to build a boat

Machine Translation (Moses)

Dependency Parser (Parsey) + English Open IE (PredPatt)

克里斯 想 造 一艘 船。
PIPELINE: BLEU=17.2 / PredicateF1=27.2
JOINT w/ Moses: BLEU=18.3 / PredicateF1=25.4

克里斯想造一艘船。

Phrase-based Machine Translation (Moses)
PIPELINE: BLEU=17.2 / PredicateF1=27.2
JOINT w/ Moses: BLEU=18.3 / PredicateF1=25.4
JOINT w/ Neural: BLEU=18.9 / PredicateF1=28.0

克里斯想造一艘船。
Analysis & Next Steps

• % malformed output:
  – Pipeline: 15%
  – Joint w/ Moses: 84%
  – Joint w/ Neural: 1%

• Next steps:
  – Directly optimize IE objective, not max likelihood
  – Integration with analyst search engine

For more info, see our EACL2017 paper:
MT/IE: Cross-lingual Open Information Extraction with Neural Sequence-to-Sequence Models