Cross-lingual Open Information Extraction

Sheng Zhang  Kevin Duh  Ben Van Durme

Johns Hopkins University
**Analyst**: User with complex information need
Data
Entity Recognition
Documents with these persons: { }
Documents with a:

CausalRelation(DISEASE, SYMPTOM)
Documents with a CausalRelation(DISEASE, SYMPTOM)

Foreign Language Documents
Financial Analyst: Do I invest in commodity futures?
Trans-Sahara trafficking and threat finance
22 April 2015

Intelligence Analyst: How are the terrorists connected?
Aid Worker: Which locations have immediate need?
Outline

1. Motivation

2. Problem Definition

3. Pipeline vs. Joint Solution

4. Improved Joint Solution
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

Bill Gates, Microsoft co-founder, stepped down as CEO in January 2000. Gates was included in the Forbes wealthiest list since 1987 and was the wealthiest from 1995 to 2007...

It was announced that IBM would buy Ciao for an undisclosed amount. The CEO, MacLorrance has occupied the corner office of the Hopkinton, company

The company’s storage business is also threatened by new, born-on-the Web could providers like Dropbox and Box, and …

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>

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)
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

Chris wants to build a boat
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.
Outline

1. Motivation
2. Problem Definition
   3. Pipeline vs. Joint Solution
4. Improved Joint Solution
Chris wants to build a boat.
Chris wants to build a boat

Machine Translation

Dependency Parser + English Open IE

克里斯想造一艘船。
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
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

[(Chris:a_h) wants:p_h (Chris:a_h) build:p_h (a:a boat:a_h)]
Experiment Setting

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

# Predicates in OpenIE output
Chris wants to build a boat.

Machine Translation (Moses)

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

Chris 想 造 一 艘 船。

PIPELINE: BLEU=17.2 / PredicateF1=24.2
Chris wants Chris to build a boat.
PIPELINE: BLEU=17.2 / PredicateF1=24.2
JOINT w/ Moses: BLEU=18.3 / PredicateF1=25.1
JOINT w/ Neural: BLEU=18.9 / PredicateF1=25.8
Outline

1. Motivation
2. Problem Definition
3. Pipeline vs. Joint Solution
4. Improved Joint Solution
Sequence generation vs. labeling

• Previously, treat word:label as single token

\[(\text{Chris}:a_h) \text{ wants}:p_h \ [(\text{Chris}:a_h) \text{ build}:p_h \ (a:a \text{ boat}:a_h)]\]
Decompose generation and labeling

\[ P(Y, T \mid X) = \prod_{i=1}^{\left| Y \right|} P(y_i, t_i \mid y_{<i}, t_{<i}, X) \]

\[ = \prod_{i=1}^{\left| Y \right|} P(y_i \mid y_{<i}, t_{\leq i}, X) P(t_i \mid y_{<i}, t_{<i}, X) \]

Decoder depends on \( t_i \)

Predict label \( t_i \)
Decompose generation and labeling

\[ \prod_{i=1}^{\|Y\|} P(y_i \mid y_{<i}, t_{\leq i}, X) P(t_i \mid y_{<i}, t_{<i}, X) \]

- Decoder depends on \( t_i \)
- Predict label \( t_i \)

- Limits increase of target vocabulary
- Models generation process separately by type
  - Given previous word “wanted”:
  - predicate decoder generates “to”, “by”
  - argument decoder generates “a”, “him”
Selective Decoding

Selector

Predicate Decoder

Argument Decoder

$S_0$
In detail: at each decoding step

$s_i = \sum_{t_i \in T} m_i[t_i] f_t(y_{i-1}, s_{i-1}, c_i)$

Attention context ($c_i$) is omitted in the figure.
Results on Chinese-English task

BLEU  
-  Pipeline: 17.2
-  MosesSMT+PredPatt: 18.9
-  Joint Seq2Seq: 23.9
-  Joint Seq2Seq (previous section): 24.2
-  Joint Seq2Seq with Selective Decoding: 25.8

Predicate F1  
-  Pipeline: 25.8
-  MosesSMT+PredPatt: 23.9
-  Joint Seq2Seq: 31.7
-  Joint Seq2Seq (previous section): 24.2
If we only care about MT (not MT+IE)

Chris wants to build a boat

Standard seq2seq
BLEU = 24.92

Seq2Seq with selective decoding, then ignore labels
BLEU = 25.16
BLEU on Low-Resource Languages
(Data from DARPA LORELEI Project)

**NOT SHOWN:** The winner is less clear for Predicate F1 (8-14%)
Summary
Support users with complex information needs

Aid Worker

Intelligence Analyst

Financial Analyst
Input: Chinese sentence

克里斯 想 造 一艘 船。

Cross-lingual Open IE

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

Chris wants build a boat
Next Steps

• Integration with analyst search engine
• Directly optimize IE objective, not likelihood
• Explore Selective Decoding for other problems
Thanks!

• To Learn More:
  – Code on GitHub