Aligning English Strings with Abstract Meaning Representation Graphs

Nima Pourdamghani  Yang Gao  Ulf Hermjakob  Kevin Knight
Information Sciences Institute
University of Southern California

**Overview**

Abstract Meaning Representation (AMR) [1]:
- Logical meaning of sentences
- Directed acyclic graphs with labeled edges

The boy wants to go

:w / want-01
:arg0 (b / boy)
:arg1 (g / go-01)

- Find alignment links between English tokens and AMR concepts
- Alignments are required for: Semantic parsing, English generation

**Approach**

- Similar to Statistical Machine Translation
- Linearize AMR graph (not obvious how)
- Use string / string alignment
- Easier than SMT
- AMR and English are highly cognate
- Harder
- AMR is a graph with unordered nodes
- Much less training data than in SMT

**Corpus**

- 13050 public AMR/English sentence pairs
- Hand Aligned 200
  - 100 dev, 100 test
- Ratio of aligned tokens in the gold data
  - English: ~ 3/4
  - AMR: ~ 1/2
- AMR role tokens: ~ 1/4

**The process**

**Preprocess**

- The boy wants to go
  - :w / want-01
  - :arg0 (b / boy)
  - :arg1 (g / go-01)

- Linearize AMR:
  - w / want-01
  - :arg0 b / boy
  - :arg1 g / go-01

- Remove stopwords
  - English: boy wants go
  - AMR: want-01 boy go-01

- Remove word sense indicator, etc. in AMR
  - want boy go

- Stem both English and AMR to first four letters
  - English: boy want go
  - AMR: want-01 boy go-01

**Extend Parallel Corpus**

- Tokens that look the same after stemming
- boy
- want
- go
- go

- English tokens that map to multiple AMR ones
  - bigger big
degree more
- month
- month

- November

**Symmetrized EM**

- Word alignment is symmetric
  - Training should be symmetric as well
  - New objective:
    \[ \theta_{AB} = \arg\max \{ L_{AB}(A|E) + L_{BA}(E|A) \} \]

- Subject to:
  - \( \theta_{AB} = \theta_{BA} \)
  - \( \theta_{AB} \)
  - \( \theta_{BA} \)

- Approximate solution:
  - 1- optimize \( \theta_{AB} = \arg\max L_{AB}(A|E) \)
  - 2- satisfy constraint, initialize \( \theta_{BA} \approx \theta_{AB} \)
  - 3- optimize \( \theta_{AB} = \arg\max L_{BA}(E|A) \)
  - 4- satisfy constraint, initialize \( \theta_{BA} \approx \theta_{AB} \)
  - 5- Iterate
  - Steps 1 and 3: EM (IBM models)
  - Steps 2, 4: simple initialization
  - No extra code needed

**Training**

- Based on IBM word alignment models [2]
- Use EM to maximize likelihood:
- Generating AMR from English
  - \( \theta_{AB} = \arg\max L_{AB}(A|E) \)
- Or, generating English from AMR
  - \( \theta_{BA} = \arg\max L_{BA}(E|A) \)
- Decoding: get the most probable alignments given parameters using Viterbi algorithm

**Postprocess**

- Goal: rebuild the aligned AMR graph
- Restore stopwords, change alignments
- Rebuild graph using recorded original structure

**Experiments**

**Precision, Recall, F-measure**

- We used Mgliza++ as implementation of IBM models
- Experiment setup (Model 4+):
  - 5 × Model 1 + 5 × HMM + 4 × symmetrized Model 4

<table>
<thead>
<tr>
<th>Model</th>
<th>precision</th>
<th>recall</th>
<th>F score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>70.9</td>
<td>71.1</td>
<td>71.0</td>
</tr>
<tr>
<td>HMM</td>
<td>87.6</td>
<td>80.1</td>
<td>83.7</td>
</tr>
<tr>
<td>Model 4</td>
<td>89.7</td>
<td>80.4</td>
<td>84.8</td>
</tr>
<tr>
<td>Model 4+</td>
<td>94.1</td>
<td>80.0</td>
<td>86.5</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>74.8</td>
<td>71.6</td>
<td>73.2</td>
</tr>
<tr>
<td>HMM</td>
<td>83.8</td>
<td>73.8</td>
<td>78.5</td>
</tr>
<tr>
<td>Model 4</td>
<td>85.8</td>
<td>74.9</td>
<td>80.0</td>
</tr>
<tr>
<td>Model 4+</td>
<td>92.4</td>
<td>75.6</td>
<td>83.1</td>
</tr>
</tbody>
</table>

**Error Sources**

- Most of the error is on role tokens
  - role tokens don’t have a specific translation in English
  - some can align to many different English words
  - they can match to part of an English word
  - or the connection might be very implicit

- About 35% of recall loss is due to removing aligned stop words

**Conclusions Future Work**

We have presented:
- The first set of manually aligned English/AMR pairs (available in amr.isi.edu)
- The first system, and a strong baseline, for learning alignments between English sentences and AMR graphs
- The system is adaptable to any domain and any language
- First step for parsing AMR from English and generating English from AMR

**References**

Griffitt, U. Hermjakob, K. Knight, P. Koehn, M.
Palmer, and N. Schneider. 2013. Abstract meaning
representation for sembanking. Linguistic
Annotation Workshop (LAW VII-ID), ACL.

and R. L. Mercer. 1993. The mathematics of
statistical machine translation: Parameter
estimation. Computational linguistics, 19(2):263–
311.

**Acknowledgements**

This work was supported by DARPA contracts
HR0011-12-0-0014 and FA-8750-13-2-0045.
The authors would like to thank David
Chiang, Tomer Levinboim, and Ashish
Vaswani for their comments and suggestions.