English Understanding: From Annotations to AMRs

Nathan Schneider

August 28, 2012 :: ISI NLP Group :: Summer Internship Project Presentation
Current state of the art: syntax-based MT

- Hierarchical/syntactic structures on source and/or target side
- Learn string-to-tree, tree-to-string, or tree-to-tree mappings for a language pair
- Syntax good for linguistic well-formedness
U.S. maternal birth to 12 kg giant baby choose not to anesthesia delivery
Why go deeper than syntax?

**FRAGMENTATION**

I lied to her.
She was lied to.
I told her a lie.
I told a lie to her.
She was told a lie.
A lie was told to her.
Lies were told to her by me.
What she was told was a lie.

**CONFLATION**

She lies all the time
...to her boss.
...on the couch.
• How to get from the source sentence to target meaning, and from target meaning to target sentence?
  ‣ graph transducer formalisms & rule extraction algorithms *(previous talk!)*
  ‣ **designing English meaning representation & obtaining data**
  ‣ English generation from meaning representation *(next talk!)*

U.S. maternal birth to 12 kg giant baby choose not to anesthesia delivery
AMR Goals

- Meaning representation for English which is “more logical than syntax,” yet close enough to the surface form to support consistent annotation (not an interlingua)
  - Principally: **PropBank event structures with variables** (allowing entity and event coreference)
  - + special conventions for named entities, numeric and time expressions, modality, negation, questions, morphological simplification, etc.
  - in a **unified** graph structure
AMR Working Group

- ISI, U Colorado, LDC, SDL Language Weaver
- This summer: fine-tuning the AMR specification to the point where we can train annotators and expect decent inter-annotator agreement
  - Practice annotations, heated arguments!
  - Expanding to genres besides news
AMRs

(l / like-01
 :ARG0 (d / duck)
 :ARG1 (r / rain-01))

- ducks like rain
- the duck liked that it was raining
AMRs

(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))

(s2 / see-01
  :ARG0 (i / i)
  :ARG1 (d / duck
    :poss (s / she))))

- I saw her duck
(I / like-01
 :ARG0 (d / duck)
 :ARG1 (r / rain-01))

AMRs

(s2 / see-01
 :ARG0 (i / i)
 :ARG1 (d / duck
 :poss (s / she)))

(s2 / see-01
 :ARG0 (i / i)
 :ARG1 (d / duck-01
 :ARG0 (s / she)))

- I saw her duck (alternate interpretation)
She saw her (own) duck

AMRs

(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))

(s2 / see-01
  :ARG0 (i / i)
  :ARG1 (d / duck
    :poss (s / she)))

(s2 / see-01
  :ARG0 (s / she)
  :ARG1 (d / duck
    :poss s))

- She saw her (own) duck
She saw her (someone else’s) duck
Ducks who like rain are happy
Ducks who like rain are happy
Happy ducks like rain
Getting the AMRs we want

- **Ideal goal:** Learn a string-to-graph transducer using parallel data with Chinese string and gold-standard AMRs
Getting the AMRs we want

- **Ideal goal:** Learn a string-to-graph transducer using parallel data with Chinese string and gold-standard AMRs predictions of an English semantic analyzer that was trained on gold standard AMRs
Getting the AMRs we want

- **Ideal goal:** Learn a string-to-graph transducer using parallel data with Chinese string and gold-standard AMRs predictions of an English semantic analyzer that was trained on gold-standard AMRs hand-coded (rule-based)

- **Intermediate goal:** Build a rule-based English semantic analyzer for data that already has some gold-standard semantic representations

- **Next:** Fully automate so an AMR can be generated for any sentence (with existing tools and/or bootstrapping off of gold-standard annotations)
Combining Representations

In practice, working with the many different file formats and representational details is very tedious.
Our solution: a single JSON file for each sentence with many (gold & automatic) annotations

- For WSJ, required a lot of massaging to ensure compatibility across annotations

• Credits: Christian Buck, Liane Guillou, Yaqin Yang
AMR Generation

- Rule-based integration of OntoNotes annotations (+ some output of existing tools)
- The sentence below will illustrate the pipeline and the kinds of annotations it exploits
  - The AMR is built up incrementally as each new piece of annotation is considered
  - This is the actual system behavior ...albeit on a short and easy example!
  - Mr. Stearn, 46 years old, couldn’t be reached for comment.
nes: BBN Corpus

- BBN Pronoun Coreference & Entity Type Corpus: fine-grained named entity labels and anaphoric coreference for WSJ
  - Entity categories include refinements of the standard PERSON/ORG/LOCATION (e.g. LOCATION:CITY) as well as other categories (LAW, CHEMICAL, DISEASE, ...)  
  - BBN IdentiFinder tagger

Mr. Stearn, 46 years old, couldn’t be reached for comment.
timex: Stanford sutime

- TIMEX3 is a markup format for time expressions (last Tuesday, several years from now, 7:00 pm, Tuesday, Aug. 28)

  - Stanford sutime tagger produces XML, e.g.: `<TIMEX3 tid="t1" value="P46Y" type="DURATION">46 years old</TIMEX3>

  - We implemented rules to handle different kinds of normalized time expressions

    DURATION:P46Y

  - Mr. Stearn, 46 years old, couldn’t be reached for comment.
PropBank annotations from OntoNotes provide the main skeleton of the sentence

- AMR has a somewhat different set of non-core roles; here :ARGM-PNC ought to be replaced with :purpose

- Note that the :ARG1 is a fragment from a previous module. Done with variable-to-token alignments and head finding for phrases.

Mr. Stearn, 46 years old, couldn’t be reached for comment.
Mr. Stearn, 46 years old, couldn’t be reached for comment.
verbalize: NomBank nouns to PropBank verbs

- AMR uses only verbal predicates, so mappings in the NomBank lexicon are used to convert nouns to verbs where possible
  - Here, we know `comment.n.01` corresponds to `comment.v.01`
  - Some nouns refer to a verb’s argument: a *filter* in AMR essentially becomes a *thing that filters*
  - Deciding when to convert a noun to a verb is often tricky, even for humans!

- Mr. Stearn, 46 years old, couldn’t be reached for comment.
conjunctions

- Identify coordinate structures based on the dependency parse (Stanford dependency converter). No coordination in this sentence.

copulas

- Predicate nominals/adjectives and nominal appositives. None in this sentence.

- Mr. Stearn, 46 years old, couldn’t be reached for comment.
Mr. Stearn, 46 years old, couldn’t be reached for comment.
Mr. Stearn, 46 years old, couldn’t be reached for comment.

- Coreferent expressions (typically, pronouns and their antecedents) are marked. N/A here.

- Heuristically designates a main concept based on the dependency parse (here, incorrectly).

  (7 / possible-or-permit-01 :domain (4 / reach-02-ROOT ...))

- Produces the final version of the AMR for human eyes...

  » Mr. Stearn, 46 years old, couldn’t be reached for comment.
Mr. Stearn, 46 years old, couldn’t be reached for comment.
Mr. Stearn, 46 years old, couldn’t be reached for comment.
AMR Generation

• 13 modules, each addressing some part of the meaning by consulting annotations and updating the working AMR

  ‣ Ulf has built a similar pipeline; ours uses more preexisting semantic representations (e.g. NomBank), Ulf’s is more fine-tuned and relies more heavily on lexical lists and specialized rules

• The system produces something reasonable for a cherry-picked example. But overall?
  
  ‣ Do we gain anything from NomBank?
Effect of NomBank

- Shu Cai’s smatch metric applied to compare 73 generated vs. gold-standard AMRs
  - Precision, recall, $F_1$ of graph edges under best matching of nodes
  - Daniel Bauer’s implementation

- Baseline: Pipeline – NomBank (no predicates for `comment`, `filter`, `president`)
  - $58 \quad 57 \quad 57$

- Full NomBank with verbalization: `comment-01`, `(thing :ARG0-of filter-01), president-n-01`
  - $57 \quad 53 \quad 55$

- Only NomBank predicates that are verbalized: `comment-01, (thing :ARG0-of filter-01), president`
  - $60 \quad 58 \quad 59$
Taking stock

• We get decent AMRs given gold annotations, but there is room to grow
  ‣ Lots of obvious tweaks that can be made
  ‣ Interesting NLP subproblems: prepositions/relations between nominals, modality/negation, etc.
    ‣ Complementary techniques from Ulf’s approach
• Automating the process so we can get AMRs for non-OntoNotes parallel data
• End-to-end MT!
Contributions

- Understanding of English semantic annotation schemes, corpora, and tools
- A tool for integrating several kinds of English annotations (OntoNotes, NomBank, automatic) into a single JSON file (with compatible indexing!)
- Manual AMR annotations & improvements to the AMR specification
- A prototype AMR generator that is highly modular and leverages many existing representations
- Understanding of the major challenges that remain for automatic AMR generation
PropBank 1.0

- Annotates the WSJ corpus (1M words) for verb propositions, applying a lexicon of verb frames to predicates and their arguments in the text. Included (and expanded to other data) in OntoNotes.
  - Mappings from PropBank frame rolesets to VerbNet verb classes/thematic roles are available in: the standalone PropBank release (though with limited coverage); the SemLink 1.1 release; and the PropBank lexicon within OntoNotes. Mappings from PropBank to FrameNet are available in OntoNotes, and (indirectly) via VerbNet in SemLink. OntoNotes sense entries map to PropBank, as do related NomBank rolesets.
- [web](#), [web](#), [LDC](#)
- [annotation manual](#), [frame creation manual](#)
- [download](#)
- [API in NLTK](#)
- A more recent version of PropBank is included within OntoNotes.

In a letter to Georgia Gulf President Jerry R. Satrum, Mr. Martin asked Georgia Gulf to answer its offer by Tuesday.

In a letter to Georgia Gulf President Jerry R. Satrum, Mr. Martin asked Georgia Gulf to answer its offer by Tuesday.

John translated his dissertation from English into Swahili, Chinese, Russian, and Yiddish.

Example frame annotation from propbank/frames/translate.xml:

http://tinyurl.com/semcorpora
### English

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Toks</th>
<th>Propositions</th>
<th>Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ (newswire; excludes 584 financial docs/375k tokens with .ont files only)</td>
<td>5433 v, 250 n</td>
<td>2683 v, 2194 n</td>
<td>1728, 597, 1728, 0, 1728, 1718, 1606, 0</td>
</tr>
<tr>
<td>Broadcast News (TDT-4)</td>
<td>900k</td>
<td>85k (82%) v</td>
<td>38k (93%(^1)) v, 51k (70%(^1)) n</td>
</tr>
<tr>
<td>Broadcast Conversation (50k EN-ZH, 50k ZH-EN)</td>
<td>200k</td>
<td>27k (89%) v</td>
<td>28k (93%) v, 28k (70%) n</td>
</tr>
<tr>
<td>English-Chinese Treebank (Xinhua newswire, Sinorama magazine)</td>
<td>200k</td>
<td>30k (95%) v</td>
<td>28k (90%) v, 17k (55%) n</td>
</tr>
<tr>
<td>P2.5 (80k ZH-EN, 65k AR-EN; 35k for each of nw, bn, bc, and wb genres)</td>
<td>325k</td>
<td>32k (90%) v</td>
<td>31k (86%) v, 48k (70%) n</td>
</tr>
<tr>
<td>Web (55k AR-EN, 75k ZH-EN)</td>
<td>145k</td>
<td>14k (87%) v</td>
<td>14k (81%) v</td>
</tr>
<tr>
<td>Selected Web sentences</td>
<td>200k</td>
<td>19k (75%) v</td>
<td>19k (73%) v</td>
</tr>
</tbody>
</table>

### Chinese

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Toks</th>
<th>Propositions</th>
<th>Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Chinese Treebank (100k Xinhua newswire, 154k Sinorama magazine)</td>
<td>20134 total</td>
<td>763 total</td>
<td></td>
</tr>
<tr>
<td>Broadcast News (TDT-4)</td>
<td>254k</td>
<td>40k (90%) v</td>
<td>32k (71%) v, 15k (20%) n</td>
</tr>
<tr>
<td>Broadcast Conversation (GALE; 50k ZH-EN, 55k EN-ZH)</td>
<td>296k</td>
<td>45k (88%) v</td>
<td>38k (75%) v, 12k (16%) n</td>
</tr>
<tr>
<td>Web (40k ZH-EN, 70k EN-ZH, 86k Dev09)</td>
<td>196k</td>
<td>15k (74%) v</td>
<td>26k (83%) v</td>
</tr>
<tr>
<td>P2.5 (nw, bn, bc, and wb genres)</td>
<td>40k</td>
<td>6k (63%) v</td>
<td>3k (33%) v</td>
</tr>
</tbody>
</table>

### Arabic

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Toks</th>
<th>Propositions</th>
<th>Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>An-Nahar (newswire; trees from Penn Arabic Treebank Part 3 v. 3.1)</td>
<td>623 a</td>
<td>150 v</td>
<td>111 n</td>
</tr>
</tbody>
</table>

\(\d\) WSJ sense coverage is out of the 300k-token (Year 1) portion that has been annotated for OntoNotes senses.

This table provides a breakdown of the resources available in OntoNotes 4.0. The portion on the left records counts and coverage statistics drawn primarily from the OntoNotes manual (details). To the right are counts of files in each subcorpus by filetype, computed with a script as described below. Hover over a row for a subcorpus to see its directories in the release.

### Explanations of statistics

For each language are counts of proposition frames and sense types. (Some proposition frames do not yet have any corresponding annotations.) For each subcorpus are counts of tokens, verb propositions, and verb and noun senses. (Noun proposition annotations are not included in OntoNotes 4. The word sense coverage figures give credit for monosemous words even if they are not explicitly annotated.)

---

http://tinyurl.com/on4stats
Contributions

• Understanding of English semantic annotation schemes, corpora, and tools

• A tool for integrating several kinds of English annotations (OntoNotes, NomBank, automatic) into a single JSON file (with compatible indexing!)

• Manual AMR annotations & improvements to the AMR specification

• A prototype AMR generator that is highly modular and leverages many existing representations

• Understanding of the major challenges that remain for automatic AMR generation
Thanks & Questions?