Resources for Adding Semantics to Machine Translation

Jan Hajič
Charles University in Prague
Computer Science School
Institute of Formal and Applied Linguistics

Major contributions by:
E: Silvie Cinková, Jana Šindlerová, Josef Toman, (J. Semecký)
C: Marie Mikulová, Zdeňka Urešová, Jan Štěpánek
Today...

- The family of Prague Dependency Treebanks
  - Incl. the Prague (Czech-)English Dependency Treebank
- English “Tectogrammatical Representation” (TR)
  - Annotation layers
  - From Penn Treebank+ to PDT-style English annotation
  - TR annotation of interesting English phenomena
- Spoken language annotation
  - “Speech reconstruction”
- Current status + to take home + pointers
The Family of Prague Dependency Treebanks

- Prague Dependency Treebank (Czech)
  - 2001: version 1.0 (no deep syntax/semantics)
  - 2006: version 2.0 (w/deep syntax, semantics: “tectogrammatics”)

- Prague Czech-English Dependency TB 1.0
  - 2004: automatic annotation
  - English: PTB, Czech: 1/3rd of PTB translated

- Prague Arabic Dependency Treebank 1.0
  - 2004: ~ PDT 1.0 (no deep syntax)
The Prague Cze-Eng Dependency Treebank

• Penn Treebank
  + PropBank
  + BBN (co-reference and Named Entities)
  + NP structure (D. Vadas, J. R. Curran, ACL’07)
  + “Czech-like” tectogrammatics

• Translation to Czech
  – Manual annotation (with auto pre-annotation)
    • Morphology, Syntax, Tectogrammatics (TR)
Example: English TR

- Words
- Dependencies
- Sem. function
- Valency (predicates)
- Coref (BBN)
- Named Entities (BBN)
Layers of Annotation

- **t-layer**
  - tectogrammatics
- **a-layer**
  - (surface) syntax
- **m-layer**
  - Morphology (POS)
- **w-layer**
  - words (tokens)
English Surface Syntax

- From PTB:
  - Form
  - POS Tag
  - Function label
  - (Structure)

- Added
  - Lemma
  - Heads
Head Determination Rules

• Exhaustive set of rules
  – By J. Eisner + M. Čmejrek/J. Cuřín
  – 4000 rules (non-terminal based)
    • Ex.: (S (NP-SBJ VP .)) → VP
  – Additional rules
    • Coordination, Apposition
    • Punctuation (end-of-sentence, internal)

• Original idea (possibility of conversion)
  – J. Robinson (1960s)
Example: Head Determination Rules

Rules:

\[(\text{NP (DT NN)}) \rightarrow \text{NN}\]

\[(\text{VP (VB NP)}) \rightarrow \text{VB}\]

\[(\text{VP (MD VP)}) \rightarrow \text{VP}\]

\[(\text{S (\ldots VP \ldots)}) \rightarrow \text{VP}\]
Conversion: Analytic Structure, Functions

• Syntactic Function assignment (conversion)
• Rules
  – based on PTB functional tags:
    - SBJ Sb    - PRD Pnom    - BNF Obj    - DTV Obj
    - TMP Adv

  – Ad-hoc rules (if functional tags missing)
  – Lemmatization (years → year)
Structure & Functions: PTB to P(E)DT

Penn Treebank structure (with heads added)

PDT-like Analytic Representation
(automatic pre-annotation)
Predicative Complement

- Free (non-valency) modification (of both a noun and a verb)
- attribute `compl.rf` (green arrow to the noun)

He spoke of him as of his father.
We have not answered your question completely, for which we apologize.
The machine does work, but not well enough, and therefore we either must repair it or buy a new one.
Mary sang *in the same way* like John.
English TR IV: Restriction ("Exclusion")

She does nothing but complain all day long.

except, with the exception of, excluding, (all/none) but, beyond, apart from, unless, bar, barring, besides
English TR annotation

- **TrEd**
  - Pre-annotated
  - Graphical
    - TR dep. tree is primary
  - Text + TR
  - Czech translation

- **Valency (a.k.a. “propbanking”)**
  - During TR annotation
  - Propbank origins and examples
    - Linked, displayed
EngVallex (give)

Predicate: give
give_away
give_back
give_off
give_out
give_up
give_in

transfer

Arg 0: giver
Arg 1: thing given
Arg 2: entity given to

double object: The executives gave the chefs a standing ovation.
ditransitive: She had given the answers to two low-ability geography classes.
give way: The Beatles give way to baseball in the Nipponese version.
Odd discourse-level usage: This is very odd, *trace* given uncertainties about the auction.
another discourse-level usage: Given the weakness in both the junk bond market and the stock market, traders fear that these transactions may be revised yet again.
<frame status="reviewed" roleset_name="acknowledge truth" id="sv-w55zj1">
 <frame_elements>
   <element functor="ACT" type="oblig">...</element>
   <element functor="PAT" type="oblig">...</element>
   <element functor="ADDR" type="oblig">...</element>
 </frame_elements>
 <examples>
   <example>
     <externallink type="PB example" fileref="propbank/admit.xml" rolesetref="admit.01" name="with hearer"/>
     <text>John admitted to Mary that he secretly loathed her haircut.</text>
     <arg n="0"/>
     <arg n="2" f="to"/>
     <arg n="1"/>
   </example>
   <example>
     <externallink type="PB example" fileref="propbank/admit.xml" rolesetref="admit.01" name="admit a full S, inverted"/>
     <text>"We've had a few bombs," admits "trace" Mr. Peters.</text>
     <arg n="1"/>
     <arg n="0"/>
   </example>
   <example>
     <externallink type="PB example" fileref="propbank/admit.xml" rolesetref="admit.01" name="admitted fragmentary S"/>
     <text>The charity admits spending a lot on fund raising, but says that was necessary to establish ...</text>
     <arg n="0"/>
     <arg n="1"/>
   </example>
 </examples>
 <alternations author="v-u_nobody">Basic Transitive, Sentential Complement</alternations>
 <local_history>...</local_history>
 <note>...</note>
 <externallink type="PB roleset" fileref="propbank/admit.xml" rolesetref="admit.01"/>
</frame>
Valency in Translation

- **leave-1**
  - ACT() PAT() LOC() → nechat-3
    - ACT(.1) PAT(.4) LOC()

- **leave-2**
  - ACT() DIR1(from.) → odjet-1
    - ACT(.1) DIR1(z.[.2])

- nechat-3 → **leave-1**

- odjet-1 → **leave-2**
Interannotator Agreement

2007-2009:
- New annotators (lower numbers)
- Annotation “by phenomenon”
- Restarting now
Prague English Dependency Treebank

• Availability
  – Version 1.0 now (PTB license needed)
    • 250k words
  – Full version (parallel with Czech): early 2011

• Size
  – Full WSJ portion of PTB (2312 files)
    – 49208 sentences, 1253013 tokens
Czech PDT-style Annotation

- All layers
  - morphology, syntax, tectogrammatical
- So far…
  - Automatic (many tools by many authors)
- Manual annotation
  - Complete now, co-reference annotation finishing
  - Top-down
    - Tectogrammatical first (*lower layers automatically*)
    - … then syntactic structure and morphology
Spoken corpus: Speech Reconstruction

• Beyond disfluency removal: an idea by F. Jelinek:
  – Transcription, even if perfect, is hard to analyze
  – ~ “people [when speaking] are ungrammatical”
  – ~ editing recorded dialogs for print

• Example:

  Transcript:
  [breath] *i think I th - see Si I think in this picture*

  *...after speech reconstruction:*
  *I think I see Si in this picture.*
Speech Reconstruction Annotation

- Multilevel audio/text editor “MEd”
  - Linking words, free movement of words
  - Editing, inserting, deleting words
  - Manual/auto transcripts (simultaneously visible)
  - Listening (as in transcription)
Speech Reconstruction Corpus: “Companions”

- English, Czech dialogs
  - “Wizard-of-Oz” setting for recording
  - Topic: Reminiscing over photographs
  - Uses in the EU FP6 “Companions” project
  - English: 20h, Czech: 120h
  - Manual transcription
  - Double or triple SR annotated
  - Release: spring 2011

Connecting speech and language understanding

- **Full annotation over speech data:**
  - “Companions” corpus → PDT-like annotated
    - All levels (morphology, syntax, semantics, valency)
    - Over reconstructed speech (“easy”)
    - Sample published: PDTSE corpus

  **Deep syntax / tectogrammatics**

- **POS, surface syntax, …**
- **“Reconstructed”**
- **transcript**
- **audio**

He is a member of the Club – they were the Yankees.

He is a member they’re [UN] yeah, the yankees member of the club
Summary

• PDT is/has (a)...
  – (Family of) dependency-based treebanking project(s)
    • Czech (English, Arabic, ...)
  – ~ 1mil. words
    • sufficient size for ML experiments
  – 4 interlinked layers of annotation
    • token, morphology, syntax, deep syntax/semantics++
    • independent and “full” information at all levels
    • interlinked (for the development of parsers/generators)
  – Parallel corpus Cze <-> Eng -> Machine Translation

• PDTSL adds...
  – Speech, transcription, speech reconstruction
Pointers, Acknowledgements

- http://ufal.mff.cuni.cz/pedt
- http://ufal.mff.cuni.cz/pdtsl
- http://ufal.mff.cuni.cz/pdt2.0

Acknowledgements

- FP7 – Network “META-NET”
- FP6-IST “Euromatrix”, Companions
- FP7-IST “Euromatrix+”, “Faust”
- LC536 (Center for Computational Linguistics)
- GAČR 405/06/0589 (Speech and deep syntax)
- MŠMT: MSM0021620838, ME838, ME09008