A construction-driven, MetaNet-based approach to metaphor extraction and corpus analysis

Elise Stickles, Ellen Dodge, and Jisup Hong
UC Berkeley and International Computer Science Institute

CSDL
November 5, 2014
Overview

• MetaNet Analysis project
• Metaphor constructions
• Automated metaphor extraction
• Corpus analysis case study
Overview: MetaNet Analysis project

• Goals
  • Development of a multi-lingual repository to support cognitive linguistic analysis of metaphor
  • Automated metaphor extraction system

• Repository
  • Networks of frames and conceptual metaphors
    • Structured relations between frames and metaphors
    • Frames organized into families of semantically coherent domains
  • English repository currently contains 647 frames and 754 metaphors
  • Includes comprehensive coverage of primary metaphors
Disease family: network of frames
Frame structure

Frame: Treating a Physical Affliction

Roles
- treatment provider
- patient
- affliction
- ...

Lexical Units
- alleviate.v
- assuage.v
- cure.n
- treat.v
- ...

Other Frame Info
- Inferences
- Relations to other frames
- Family: Disease
...neither the free market nor central planning had been able to alleviate unemployment and poverty (BNC:HKT)
Metaphoric constructions

• Conceptual metaphors are typically expressed in particular syntactic patterns.
  • Target and source lexemes reliably occupy certain grammatical slots. (Goldberg 1995; Croft 2002; Sullivan 2007, 2013)
  • E.g. Source is verb, Target is argument

• Metaphoric constructions in the MetaNet system are primarily formalized using dependency parse and part of speech tag information
neither the free market nor central planning had been able to alleviate unemployment and poverty (BNC:HKT)
Combining metaphor and construction
neither the free market nor central planning had been able to alleviate unemployment and poverty (BNC:HKT)
Metaphoric constructions:
Source is verb, Target is argument

• **Subject (Target) + Verb (Source)**
  • poverty and early death threatened family life (BNC:AP7)

• **Verb (Source) + Direct Object (Target)**
  • The measure was principally designed to tackle homelessness (BNC:G20)
Metaphoric constructions:
Source and Target are both nominals

• **N of N**: N1 (Source) + of + N2 (Target)
  - *end the twin scourges of mass poverty and mass unemployment* (BNC:CAK)

• **N-N compound**: N1 (Target) + N2(Source)
  - *Caught in the poverty trap, they are unable to make the savings necessary for business ventures.* (BNC:AN3)
Multi-lingual extraction system

- English, Spanish, Russian, and Persian
  - Taggers:
    - English, Spanish, Russian: TreeTagger
    - Persian: Custom
  - Dependency Parsers:
    - English: RASP
    - Spanish: Freeling
    - Russian/Persian: MALT

- Large corpora
  - Gigaword corpora for English, Spanish, Russian
  - 100 Million+ word corpora for Persian
Extraction process overview

- Corpus pre-processing using standard NLP methods

Text Input → POS Tagging (TreeTagger) → Lemmatization (TreeTagger) → Dependency parsing (RASP)
Extraction process overview

• Metaphor extraction using repository and constructions

1. Identify target expressions in the corpus, e.g. poverty
2. Construction matching patterns find source expressions grammatically related to target expressions
3. Metaphor candidates (target and source word pairs) are evaluated for metaphoricity
   • Frame relations in the metaphor repository used to determine non-metaphorical relatedness or metaphorical relatedness
4. Filter out low-scoring candidates to create database of linguistic metaphors
Extraction output

```json
{
  "extractor": "CMS",
  "name": "poverty hurt",
  "source": {
    "end": 13,
    "form": "hurts",
    "lemma": "hurt",
    "framenames": ["Experience pain", "Harm to living entity"],
    "lpos": "hurt.v",
    "start": 8,
  },
  "score": 0.93,
  "cms": ["POVERTY_IS_PHYSICAL_HARM",
    "EXPERIENCING_A_NEGATIVE_STATE_IS_EXPERIENCING_HARM"],
  "cxn": "**T-subj_S-verb",
  "target": {
    "concept": "POVERTY",
    "end": 7,
    "framefamily": "Economic inequality frames",
    "form": "Poverty",
    "lemma": "poverty",
    "start": 0,
    "framename": "Poverty",
  }
}
```
Overview: Case study

- **Goal:** Illustrate system’s potential
  - Large-scale automated extraction
  - Maintain analytic depth
  - Enabled by repository framework and data

- **British National Corpus**
  - 100 million words
  - Late 20th century (1970s-1990s)
  - 90% written, 10% spoken

- **Target domain: Poverty**
  - Defined by poverty frame family in repository
  - Extraction results
    - 698 sentences
    - 209 source domain lexical units
    - 134 source frames; 957 frame instances
Case study: Constructions

Frequency of Constructions

<table>
<thead>
<tr>
<th>Construction</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>N of N</td>
<td>290</td>
</tr>
<tr>
<td>V-DO</td>
<td>275</td>
</tr>
<tr>
<td>S-V</td>
<td>190</td>
</tr>
<tr>
<td>N-N</td>
<td>130</td>
</tr>
<tr>
<td>N's N</td>
<td>10</td>
</tr>
<tr>
<td>N is N</td>
<td>5</td>
</tr>
</tbody>
</table>
Case study: Source family data

- Highest frequency source domain frames
  - Identified 3 major source frame families
  - Account for 42% of frame instances

<table>
<thead>
<tr>
<th>Disease family</th>
<th>Impediments to motion family</th>
<th>Harm family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis of affliction</td>
<td>Antagonistic force</td>
<td>Attacking</td>
</tr>
<tr>
<td>Disease</td>
<td>Burden</td>
<td>Danger</td>
</tr>
<tr>
<td>Health treatment</td>
<td>Constructed confinement</td>
<td>Destroying</td>
</tr>
<tr>
<td>Physical affliction</td>
<td>Constructed restraints</td>
<td>Harm</td>
</tr>
<tr>
<td>Treating a physical affliction</td>
<td>Remove burden</td>
<td>Harm to living entity</td>
</tr>
<tr>
<td></td>
<td>Sunken confinement</td>
<td>Impact</td>
</tr>
<tr>
<td></td>
<td>Water confinement</td>
<td>Killing</td>
</tr>
</tbody>
</table>
Case study: Major source family data

Frequency of Major Source Families

Frequency

Disease
Harm
Impediments to motion
Constructions in Source family data

Distributions of Constructions by Source Family

- Disease
- Harm
- Impediments to motion

% of family

- Symptoms of poverty
- Alleviate poverty
- Impact of poverty
- Tackle homelessness
- Poverty trap

n.s.

**
Frames within Disease family

- Metaphors:
  - **POVERTY IS A PHYSICAL AFFLICTION**
  - **ADDRESSING POVERTY IS TREATING A PHYSICAL AFFLICTION**
Constructions in Disease family

Distribution of Constructions in Disease Family

**

Frequency

Treating a Physical Affliction

Physical Affliction

Disease

Health Treatment

Diagnosis of Affliction

N of N
V-DO
S-V
N-N
Constructions in Disease data: Examples

Source frame: Treating a Physical Affliction
Construction: Verb(Source)-Direct Object(Target)
  • *palliate* serious *destitution* (BNC:EE9)
  • *eradicate* basic *poverty* (BNC:A65)
  • *alleviate the poverty and suffering of the world* (BNC:ALH)
  • *cure deep poverty* (BNC:ABJ)
  • *relieve homelessness* (BNC:A4K)

Source frame: Physical affliction
Construction: Noun(Source) of Noun(Target)
  • *the Tory scourge of homelessness* (BNC:HHV)
  • *the sickness of poverty* (BNC:ABE)
  • *a symptom of poverty* (BNC:APN)
Frames in Target family data

Distribution of Source Families between Target Frames

- Disease
- Harm
- Impediments to motion

Frequency

- Impoverished
- Poverty
Constructions in Target family data

Distribution of Constructions in Poverty

- N-N
- Subj-V
- V-DO
- N of N

Poverty
Impoverished

alleviation of poverty
clobbering the poor
Conclusions and further progress

• Proof of concept
• Enables rich semantic analysis at different levels of granularity
• Further progress:
  • Defining additional constructions:
    • V(S)-Prep-IO(T), V(S)-Part-Prep-IO(T), Adj-N, N(S)-Prep-N(T), N(S)-Part-Prep-N(S)
  • Continuing to expand the repository’s semantic coverage
  • Further testing and refinement of system
  • Cross-linguistic and cross-cultural comparisons
• Initial release of metaphor repository in 2015
References

• The British National Corpus, version 3. (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/


Acknowledgements

• The MetaNet Analysis and Repository teams: George Lakoff, Eve Sweetser, Oana David, Karie Moorman, Luca Gilardi
• Collin Baker, Jim Hieronymous, and the rest of the MetaNet project members
• Comments and suggestions from Michael Ellsworth

• Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0022. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.
Thank you!

• Contact:
  • elstickles@berkeley.edu
  • edodge@icsi.berkeley.edu
  • jhong@icsi.berkeley.edu