Acquisition of Multiword Lexical Units for FrameNet

Silvana Hartmann & Iryna Gurevych
Part 1

MWEs and FrameNet: state of affairs
FrameNet

- Lexical semantic resource modeling „frame semantics“
- Frames group words in classes by „typical situations“
- Frame-specific Roles are assigned to participants in the situation (syntactic arguments)

**CCM bought Chrysler from Daimler**

**FRAME**

`Commercial Transaction`

**ROLE**

- **Buyer**
- **Goods**
- **Seller**
Multiword Expressions

MWE

Multiword Expression

- A phrase that stands out because of it is lexicalized or institutionalized (Sag et al., 2002)
- Lexicalized: idiosyncratic syntax, semantics, etc.: **up and away, to spill the beans**
- Institutionalized: marked by frequent use: **leather jacket**
Modeling MWEs in FrameNet

- Lexical Entries
  - contain MWE constituents („lexemes“)
  - and marks MWE heads
- MWE Semantics:
  - Incorporated Roles (ICRs) can be marked
  - Make implicit relations between constituents explicit!
  - See Calzolari et al. (2002)
- Example:
## MWEs in FrameNet I

MWE Senses in FrameNet 1.5

<table>
<thead>
<tr>
<th>POS</th>
<th># MWEs</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>432</td>
<td>sell short, stay in touch</td>
</tr>
<tr>
<td>Noun</td>
<td>246</td>
<td>sugar daddy, rush hour</td>
</tr>
<tr>
<td>Adjective</td>
<td>125</td>
<td>next generation, heart-warming</td>
</tr>
<tr>
<td>Other</td>
<td>63</td>
<td>beyond compare</td>
</tr>
<tr>
<td>TOTAL</td>
<td>866</td>
<td>(806 distinct lemmas)</td>
</tr>
</tbody>
</table>
MWEs in FrameNet II

English MWE Lemmas in other resources

<table>
<thead>
<tr>
<th>Resource</th>
<th># MWEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>FrameNet</td>
<td>806</td>
</tr>
<tr>
<td>WordNet</td>
<td>69,466</td>
</tr>
<tr>
<td>VerbNet</td>
<td>105</td>
</tr>
<tr>
<td>Wiktionary EN</td>
<td>67,548</td>
</tr>
<tr>
<td>OmegaWiki EN</td>
<td>21,000</td>
</tr>
<tr>
<td>ALL</td>
<td>139,281</td>
</tr>
</tbody>
</table>

UBY

numbers from UBY
MWE Semantics in FrameNet: Incorporated Roles

- Senses with ICRs in FrameNet 1.5: 1145 (56 MWEs)
- Most of the ICRs refer to the full mwe, not a constituent:

  - Bacterial Meningitis
  - Medical_Conditions
  - Ailment

  - to act up
  - Conduct
  - Manner

- What is lacking: ICRs referring to constituents

  - leather jacket
  - Clothing
  - Material

*(leather jacket does occur in the example sentences, but not in the lexicon)*
MWEs and Constructions

- Productive MWE patterns

- Don’t want to list them all
- derive constructions?
Other Languages

- Example: German

  Kalbs|schnitzel
  (veal / schnitzel)

  Made from veal

  Kinder|schnitzel
  (children / schnitzel)

  For children

- Also productive: Kinder|döner, Studenten|döner
- Also events: Wetter|vorhersage (weather/ forecast)
Summary: Situation of MWEs in FrameNet

Currently it doesn't look like a love affair…

SOLUTION

- Add more (A) MWEs and (B) MWE semantics to FrameNet

GOALS

- Improved representation of MWE semantics and constructions
- Improved Semantic Role Labeling (SRL)
  - Erk & Padó (2006): lack of MWEs harms SRL performance
Part 2

Acquisition of MWEs and MWE semantics
(A) MWE Acquisition for FrameNet

- Assigning MWEs to Frames
- MWEs from: other lexical semantic resources or corpora
- How?
  - Distributional Methods (see Pennacchiotti et al., 2008)
- Also works for other languages...
  - e.g., Italian (see Pennacchiotti et al., 2008)
  - e.g., alignment to Wikipedia for Italian (Tonelli et al., 2013)
  - e.g., alignment to Wiktionary for German (Hartmann & Gurevych (2013), to appear)
(B) Acquisition of MWE Semantics

- How to automatically discover ICRs? Use paraphrases!

1. **bread container**

   - FRAME: Container
     - ICR: Contents

   - FRAME: Container
     - ICR: Contents

2. **that contains bread**

   - FRAME: Container
     - ICR: Contents

---

- Related Task: Noun Compound Interpretation (Nakov, 2008)
- use datasets from Noun Compound Interpretation
Proposed Setup

- Annotate paraphrases of MWEs with SRL system
- Identify ICRs based on annotations of the paraphrasing words
- Alternative: develop specialized classifier

- MWE + Paraphrase
- Input
- Annotate with SRL System
- Classifier on SRL Annotations
- Evaluation: Comparison with GS
- MWE + Paraphrase
- Input
- Classifier
- For ICRs
- Evaluation: Comparison with GS
Proposed Setup

- Annotate paraphrases of MWEs with SRL system
- Identify ICRs based on annotations of the paraphrasing words
- Alternative: develop specialized classifier

**MWE + Paraphrase**

**Input**

**that contains bread**

**FRAME**

**Container**

**ICR**

**Contents**

**Evaluation:** Comparison with GS

**Evaluation:** Comparison with GS
Part 3

First experiments on the acquisition of MWE semantics
First Experiments

- Annotate paraphrases of MWEs with SRL system (SEMAFOR, Das et al., 2010)
- Identify ICRs based on annotations of the paraphrasing words

Hybrid System:

1. Paraphrase Input
2. Baseline Input
3. Annotate with SRL System
4. Classifier on SRL Annotations
5. Evaluation: Comparison with GS
Data

- Noun Compounds and Paraphrases from Semeval Shared tasks:
  - Semeval 2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions (Butnariu et al., 2010)
  - Semeval 2013 task 4: Free Paraphrases of Noun Compounds

- Compounds:
  - Noun compounds with two constituents: adventure story, advertising agency..
  - 812 noun compounds

- Paraphrases:
  - Collected via crowd-sourcing, ranked by frequency
  - Example:

  motor yacht \[\rightarrow\] yacht propelled by a motor
  yacht with a motor
Gold Standard

- **Problem 1**: Minimal overlap with shared task MWEs and FrameNet lexicon: „child abuse“
- **Problem 2**: Few MWEs with ICRs in FrameNet (and no paraphrases for these..)
- **Solution**: annotation of a gold standard (GS) on shared task MWEs
  - Match MWE to candidate frame based on head
  - Two-class annotation: MWE matches the frame (yes|no)
  - If yes: annotate whether non-head matches a role of the frame (= is an ICR)

- 412 MWEs with between 10 and 170 paraphrases each
- 225 have an ICR in the gold standard

```
cigarette burn
```

```
FRAME  Bodily_Harm
ICR    Injuring_E
```
Baseline Experiments

- **Question:** How well does Semafor annotate the MWEs out of context?

- **Experiment Setup**
  - Annotate MWEs with Semafor
  - Compare System Annotations S to Gold Standard Annotations G
  - **High-recall setup:**
    - Intersect G frame with S frames
    - If yes, intersect G roles with S roles
    - If set not empty=> Role-match

- **Results**
  - There are annotations for 50% of the GS
  - There is a Frame-match for 125 MWEs
  - There is a Role-match for 10 MWEs
Experiments on Paraphrases

- Same setup as baseline relying on Semafor annotations on paraphrases
- Variables:
  - All paraphrases (high recall) vs. best rated paraphrase
  - Most frequent annotation vs all annotations

- Best results for „all“ paraphrases:
  - There are annotations for 60% of the Gold Standard
  - There is a frame-match for 148 MWEs
  - There is a role match for 42 MWEs (0.28 accuracy)

- Next steps:
  - Evaluate full sentences as paraphrases
  - Focus on precision by filtering paraphrases and/or annotations for constituents
  - Increase dataset and train classifier (next slide)
Acquisition of MWE Semantics using Paraphrases - revisited

- How to automatically discover ICRs?

1. **bread container**
   - Container
   - Contents

2. **leather jacket**
   - Clothing
   - Material

- More complex dependencies for certain paraphrases
More Paraphrases

- Semeval Shared 2013 Task for Generating Paraphrases of Noun Compounds is ongoing…
- Alternative: Generate Paraphrases from Parallel Text
  - Monolingual Translation Tables
  - Multilingual Translation Tables: EN→DE & DE→EN leads to EN paraphrases („Pivoting“)

- Resource based on Pivoting: PPDB („The Paraphrase Database“), Ganitkevitch et al. (2013)
  - Overlap with Gold Standard MWEs: 75 Paraphrases in PPDB
  - Paraphrases mainly nominal phrases, of similar length
  - Example: travelling companion → travel companion, best partner
  - Example: waste material → waste materials, solid waste, waste disposal, hazardous wastes, materials that are
Summary/Outlook

- Motivation of MWE acquisition and ICR detection for FrameNet
- Basic setup for automatic ICR detection
  - Preliminary results
  - Difficult task requires a more complex setup

Future work includes
- Evaluating more complex classification setup and additional sources of paraphrases
- Exploring the connection to relation detection for noun compounds (feature for ICR detection?)
- Exploring methods of generalizing patterns

MWE
Multiword Expression
Thank you for your attention!

Ubiquitous Knowledge Processing Lab
References


References


