Supervised relation extraction for ontology learning from text based on a cognitively plausible model of relations

Massimo Poesio and Eduard Barbu and Claudio Giuliano and Lorenza Romano
Outline

- Introduction
- Feature Norms and the feature Classification Schema
- Kernel methods for relation extraction
- Experimental setup
- Results and conclusions
- Future work
One of the most successful approaches for semantic relation learning exploit an idea first articulated by Hearst.

The idea is that some semantic relations can be learnt using precise lexico-syntactic patterns.

For example the hypernym is expressed with good precision using a pattern like: $NP_0$ such as {NP$_1$, NP$_2$,...,(and/or) }NP$_n$.

Many researchers used this technique for learning other relations than IS-A:

- The **Born** relation (a relation holding between people and their date of birth) (Ravichandran and Hovy)
- The **Entailment** relation (Szpektor and al.)
Kinds of structures: Qualia Structures

- The next relevant step was to use this method for the extraction of structures.

- Inspired by the work of Pustejovský (who in turn was inspired by Aristotle) Cimiano and Wenderoth use a pattern-based approach to learn Qualia Structures.

- **Qualia Structures** define the meaning of a lexical element using four roles (types of properties):
  
  - **Constitutive Roles**: Include physical properties of an object (usually parts).
    
    - Example: *Knife* has \{blade, handle\}
  
  - **Agentive Roles**: Factors involved in bringing about an object.
    
    - Example: *Knife* created \{make\}
Kinds of structures: Qualia Structures

- **Formal Role**: Properties which distinguish the object in a larger domain (usually its Super-concept).
  - Example: *Knife* is an *Artifact*

- **Telic Role**: Describes the purpose of an object.
  - Example: *Knife* has purpose *Cut*

- **Learning**: They made patterns for each roles and acquired Qualia Structures from the Web
Kinds of structures: AP classification of attributes

- Inspired by Guarino and Pustejovsky, Abdulhraman and Poesio proposed the following attribute schema
  
  - **Qualities**: Analog with Guarino and Pustejovsky formal role
  
  - **Related objects**: Include the parts of an object and other objects related with it.
    
    - Example: Car has \{door, seat, window\}
    
    - Example: Car has \{keys, registration documents\}
Kinds of structures: AP classification of attributes

- **Activities.** Activities are part of Pustejovsky’s telic role and agentive roles
  - Example: *Car* has related activities \{*Repairing, Driving, Manufacturing*\}

- **Related agents.** Agents that are related with the objects.
  - Example: *Car* has related agents \{*Driver, Owner*\}

- **Learning.** They collected from the web instances of these relations and then based on the number of features that differentiate the relations they trained two classifiers (a 5 way classifier and a binary classifier) to distinguish between the attributes.
Feature Norms

- We are searching for better founded structures (based on empirical evidence) and so we turned to feature norms.

- The main assumption behind feature norm collection is that concepts can be at least partially characterized in terms of features.

- The feature norms provide insights into mental representation of concepts.

- In a task called feature generation, subjects list what they believe are the most important properties for a set of concepts (usually the concepts are basic level concepts and they represent concrete objects).
The Feature Norm by McRae et. al.

- The feature norm collection starts with the effort of Rosch and Mervis in mid 70.

- The researchers in cognitive psychology made available some feature norms (Garrard feature norm, Vinson and Vigliocco feature norm, McRae feature norm etc...)

- In our experiments we use McRae Feature Norm
  - 725 subjects listed features for 541 living and not living basic level concepts
  - Example of the featural description of the concept knife: \{has a handle, has a blade, made of steel, is shiny, used for cutting, is sharp, a cutlery etc.. \}
Wu and Barsalou taxonomy

- The features in McRae feature norm are annotated with a modified version of Wu and Barsalou (WB) taxonomy

- The taxonomy is derived from studies of human perception.

- The principles for the taxonomy construction are
  - the introspective experience of subjects,
  - the modality specific regions of the brain,
  - the frame theory of Filmore
Part of the conceptual description for the concept **AXE**
The features are annotated using WB taxonomy
The relations for our experiment I

- **External Surface Property**. External surface properties are those properties of an entity the entity’s surface, including shape, color, pattern, texture, size, touch, smell, taste. These properties usually denote the qualities of a concrete object.
  
  - Example: The car is red

- **Function**. Function properties are properties denoting the function an entity typically fulfills.
  
  - Example: Airplane is used for transportation

- **Internal Component** : These properties denote internal components of an object
  
  - Example: The car has an engine
The relations for our experiment II

- **Origin.** Origin properties are those properties denoting the origin of an entity.
  - Example: *Cigar made in Cuba*

- **Participant.** Participant properties are those properties denoting the agents who perform an action with a certain entity.
  - Example: *Desk is used by students*

- **Superordinate.** The superordinate is the well-known IS-A relation.
Kernel methods for relation extraction

- A **kernel** is a function that computes the similarity between instances
- **Kernel Methods** map the original (nonlinear) data into a higher dimensional space
- Subsequently one uses a linear algorithm in the higher dimensional space to discover nonlinear patterns in the input space
- In our experiments we use **SVM** as learning algorithm.
Kernel construction I

- We combine two families of kernel functions:
  - **Global Context Kernels**
  - **Local Context Kernels**
- The kernels are computed like this:

\[
K(x_1, x_2) = \frac{\langle \phi(x_1), \phi(x_2) \rangle}{\|\phi(x_1)\| \|\phi(x_2)\|},
\]

\(\phi(\cdot)\) is the embedding vector and \(\| \cdot \|\) is the 2-norm.
Global Context Kernel

- Relations between entities are generally expressed using words appearing in one of the three contexts.
  - Fore-Between: Tokens before and between the entities. Example: the head of [ORG] Dr. [PER]
  - Between: Tokens between entities. Example: [ORG] spokesman [PER]
  - Between-After: Tokens between and after the two entities. Example: [PER], a [ORG] professor

- The **Global Context Kernel** is defined as:
  \[ K_{GC}(R_1, R_2) = K_{FB}(R_1, R_2) + K_B(R_1, R_2) + K_{BA}(R_1, R_2), \]
  where \( K_{FB}, K_B \) and \( K_{BA} \) are the n-gram kernels that operate on the Fore-Between, Between and Between-After patterns.

- \( R_1 \) and \( R_2 \) are the examples compared by kernel function
Other kernels used

- The **Local Context Kernel** considers the ordering of tokens. We look at a context window of 2 tokens to the left and to the right around the candidate entity.

\[ K_{LC}(R_1, R_2) = K_{left}(R_1, R_2) + K_{right}(R_1, R_2) \]

- The **Shallow Linguistic Kernel** is a combination between the local and the global kernel (it is the kernel used in our experiments)

\[ K_{GC}(R_1, R_2) + K_{LC}(R_1, R_2). \]

- **Bag of Words Kernel** (a kernel used as one of the baselines).

  - A sentence is represented by the words it contains (without word ordering)
  - Defined like the global kernel but it operates on the whole sentence
Experimental setup

- We gathered instances for each of the six semantic relations.

- We extracted sentences containing the instances from a very large corpus (UKWAC is a web corpus that has 2 billion words)

- For each relation we gathered 500 sentences. Each sentence was annotated by 2 annotators as either positive, negative or don’t know example.

  - Positive Example. Dogs can eat fruits such as apples the doctor said.

  - Negative Example. I saw many unknown animals running and a dog barking.
An in-house sentence annotation tool

1. A Fulbright research scholar in 1993, she has uncovered much early repertoire for the harp in European music archives and this has formed the basis of her own ensemble, Musica Fabula.

Situation-> Function
Function Properties are a subclass of Situation Properties. Situation properties denote property of a situations, where a situation typically includes one or more participants, at some place and time, engaging in an event, with one or more entities (e.g., picnic, conversation, vacation, meal). Function Properties denote a typical goal or role that an entity serves for an agent
Examples:
<Airplane> used for <transportation>.
<Apron> worn around <neck>.
<Flute> used in <orchestras>.
<table>
<thead>
<tr>
<th>Relation</th>
<th>Kappa Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Surface</td>
<td>0.79</td>
</tr>
<tr>
<td>Internal Component</td>
<td>0.78</td>
</tr>
<tr>
<td>Superordinate</td>
<td>0.73</td>
</tr>
<tr>
<td>Origin</td>
<td>0.63</td>
</tr>
<tr>
<td>Function</td>
<td>0.62</td>
</tr>
<tr>
<td>Participant</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Data pre-processing

- Sentences have been tokenized, lemmatized, and POS tagged.
- Each relation is considered as a distinct classification task and each sentence as a positive or negative example for the relation.
- The results were obtained by using 10-fold cross-validation. The performance of the shallow linguistic kernel, and the baseline is given in the following tables.
<table>
<thead>
<tr>
<th>Relation</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Surface</td>
<td>84.0</td>
<td>86.4</td>
<td>85.2</td>
</tr>
<tr>
<td>Function</td>
<td>87.2</td>
<td>81.5</td>
<td>84.2</td>
</tr>
<tr>
<td>Internal Component</td>
<td>70.4</td>
<td>75.0</td>
<td>72.6</td>
</tr>
<tr>
<td>Origin</td>
<td>72.2</td>
<td>73.2</td>
<td>72.7</td>
</tr>
<tr>
<td>Participant</td>
<td>85.1</td>
<td>84.0</td>
<td>84.5</td>
</tr>
<tr>
<td>Superordinate</td>
<td>79.0</td>
<td>86.8</td>
<td>82.7</td>
</tr>
<tr>
<td>Relation</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------</td>
<td>--------</td>
<td>-----</td>
</tr>
<tr>
<td>External Surface</td>
<td>84.0</td>
<td>86.4</td>
<td>85.2</td>
</tr>
<tr>
<td>Function</td>
<td><strong>87.2</strong></td>
<td>81.5</td>
<td>84.2</td>
</tr>
<tr>
<td>Internal Component</td>
<td>70.4</td>
<td>75.0</td>
<td>72.6</td>
</tr>
<tr>
<td>Origin</td>
<td>72.2</td>
<td>73.2</td>
<td>72.7</td>
</tr>
<tr>
<td>Participant</td>
<td>85.1</td>
<td>84.0</td>
<td>84.5</td>
</tr>
<tr>
<td>Superordinate</td>
<td>79.0</td>
<td><strong>86.8</strong></td>
<td>82.7</td>
</tr>
</tbody>
</table>

Relation extraction Performance
<table>
<thead>
<tr>
<th>Relation</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Voting</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Surface</td>
<td>84.0</td>
<td>86.4</td>
<td>85.2</td>
<td>Micro</td>
<td>82.1</td>
<td>82.8</td>
<td>82.5</td>
</tr>
<tr>
<td>Function</td>
<td>87.2</td>
<td>81.5</td>
<td>84.2</td>
<td>Majority</td>
<td>65.9</td>
<td>73.9</td>
<td>70</td>
</tr>
<tr>
<td>Internal Component</td>
<td>70.4</td>
<td>75.0</td>
<td>72.6</td>
<td>All true</td>
<td>50.9</td>
<td>100.0</td>
<td>67.5</td>
</tr>
<tr>
<td>Origin</td>
<td>72.2</td>
<td>73.2</td>
<td>72.7</td>
<td>BoW</td>
<td>71.8</td>
<td>74.9</td>
<td>73.3</td>
</tr>
<tr>
<td>Participant</td>
<td>85.1</td>
<td>84.0</td>
<td>84.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superordinate</td>
<td>79.0</td>
<td>86.8</td>
<td>82.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Relation extraction Performance**
<table>
<thead>
<tr>
<th>Relation</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Surface</td>
<td>84.0</td>
<td>86.4</td>
<td>85.2</td>
</tr>
<tr>
<td>Function</td>
<td>87.2</td>
<td>81.5</td>
<td>84.2</td>
</tr>
<tr>
<td>Internal Component</td>
<td>70.4</td>
<td>75.0</td>
<td>72.6</td>
</tr>
<tr>
<td>Origin</td>
<td>72.2</td>
<td>73.2</td>
<td>72.7</td>
</tr>
<tr>
<td>Participant</td>
<td>85.1</td>
<td>84.0</td>
<td>84.5</td>
</tr>
<tr>
<td>Superordinate</td>
<td>79.0</td>
<td>86.8</td>
<td>82.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Voting</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro</td>
<td>82.1</td>
<td>82.8</td>
<td>82.5</td>
</tr>
<tr>
<td>Majority</td>
<td>65.9</td>
<td>73.9</td>
<td>70</td>
</tr>
<tr>
<td>All true</td>
<td>50.9</td>
<td>100.0</td>
<td>67.5</td>
</tr>
<tr>
<td>BoW</td>
<td>71.8</td>
<td>74.9</td>
<td>73.3</td>
</tr>
</tbody>
</table>

Comparison with the baseline

Relation extraction Performance
<table>
<thead>
<tr>
<th>Relation</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Voting</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Surface</td>
<td>84.0</td>
<td>86.4</td>
<td>85.2</td>
<td>Micro</td>
<td>82.1</td>
<td>82.8</td>
<td>82.5</td>
</tr>
<tr>
<td>Function</td>
<td>87.2</td>
<td>81.5</td>
<td>84.2</td>
<td>Majority</td>
<td>65.9</td>
<td>73.9</td>
<td>70</td>
</tr>
<tr>
<td>Internal Component</td>
<td>70.4</td>
<td>75.0</td>
<td>72.6</td>
<td>All true</td>
<td>50.9</td>
<td>100.0</td>
<td>67.5</td>
</tr>
<tr>
<td>Origin</td>
<td>72.2</td>
<td>73.2</td>
<td>72.7</td>
<td>BoW</td>
<td>71.8</td>
<td>74.9</td>
<td>73.3</td>
</tr>
<tr>
<td>Participant</td>
<td>85.1</td>
<td>84.0</td>
<td>84.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superordinate</td>
<td>79.0</td>
<td>86.8</td>
<td>82.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Comparison with the baseline**

**Relation extraction Performance**

Micro-Average: We sum tp, fp, and fn for all 6 experiments.
Results and Conclusions

- Results are comparable with the state of the art in relation extraction.
- The shallow linguistic kernel significantly outperforms the 3 baselines.
- We performed experiments mapping the relations in WB taxonomy on a better set of relations (better from the point of view of learning) and used a pattern based approach to learn them.
- In the future we want to use this new set of relations for supervised learning.
The cognitively inspired structure

• The mapping of the relations of WB taxonomy to a set of better from the point of view relations is the following

  • Superordinate

  • Part

  • Stuff: Denotes the stuff the things are made of (bottle made of glass)

  • Location: Typical places where the objects are found (airplanes found in airports)

  • Action: The characteristic actions denoting the behavior of an entity (the cat meow) or the functions the entities typically fulfill (the heart pumps blood)

  • Quality
Weakly supervised approach

- To learn these relations (all experiments are performed on Ukwac) we devised a double strategy
  - For Superordinate, Part, and Stuff relations we use a pattern based approach.
  - Quality and Action properties are learnt quantifying the strength of association between the nouns representing the focal concepts and the adjectives or the verbs modifying them.
    - The best Quality properties for the noun turtle are (green, hard, small)
    - The best Action Properties for the noun turtle are (dive, nest, hatch)
Weakly supervised results

- We then compare the concepts descriptions acquired in this way with the feature norm concepts descriptions.
  - Superordinate: Recall (87%), Precision of (85%)
  - Quality: Recall (60%), Precision (60%)
  - Action: Recall (70%) and a Precision (83%)
THANK-YOU!