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

Massimo Poesio♣, Eduard Barbu♠, Claudio Giuliano♣ and Lorenza Romano♣

♣Center for Mind / Brain Sciences, Università di Trento, Italy
♠Fondazione Bruno Kessler, Trento, Italy

Abstract. Most work on ontology learning from text relies on unsupervised methods for relation extraction inspired by Hearst’s work, and attempts to extract relations identified in work in formal linguistics and ontology. In this paper we present work aiming at extracting from text the set of concept attributes actually associated to concepts according to psychological research, and using state-of-the art supervised relation extraction techniques.

1 INTRODUCTION

Most work on concept acquisition in computational linguistics is based on the distributional hypothesis: that the meaning of a word is specified by the linguistic contexts in which the word occurs—i.e., by the words it co-occurs with [28, 19, 18]. Such models, in which concept properties have proven surprisingly successful, e.g., at modelling synonymy, as shown by their results on the TOEFL test [30], but contrast with the view of concepts held in Artificial Intelligence, Linguistics, and Psychology, in which concepts are seen as complex mental objects having a number of attributes which are semantic in nature—even in those theories derived from the work of Wittgenstein and [27] that do not subscribe to the view that concepts can be ‘defined’ in the Aristotelian sense.

As a result, work on ontology learning from text (see, e.g., [8, 1], and the papers in [5]) has been primarily inspired by the work of [17] on extracting semantic relations from corpora rather than by work based on the distributional hypothesis. Hearst’s proposed pattern-based unsupervised methods, but also one of the first semi-supervised methodologies for extracting further patterns; this approach was further pursued, e.g., by [22].

Rather surprisingly, in the literature on ontology learning one can find little use of supervised methods for relation extraction [34, 11] even though such methods have been shown to have vastly superior performance at relation extraction in competitions such as ACE. To our knowledge, the only such work was carried out by [23], who, however, used a very simple classification scheme and a very simple classifier. We are also not aware of any work attempting to identify the types of relations identified in the psychological literature as characteristic of the concept descriptions provided by humans [33, 20].

This paper further pursues the line of research begun by Almuhareb and Poesio by using the state of the art relation extraction techniques developed by [13, 14, 12] and by using as a target the classification scheme for attributes proposed by Wu and Barsalou [33] on the basis of the analysis of the list of concept attributes specified by human subjects. In this paper we only concentrate on relation extraction.

The structure of the paper is as follows. After a short background section discussing relation extraction for ontology learning and the Wu and Barsalou classification, we discuss our relation extraction methods and how we used them for extracting attributes about a set of concepts; we then discuss the results.

2 BACKGROUND

2.1 Ontology learning

Most work on ontology learning from text is based on the seminal work by Hearst [17] who proposed methods for extracting hyponymy relations (i.e., is-a links) by searching in corpora for instances of patterns such as

\[
NP_1, NP_2 or other NP
\]

as in, e.g., bruises, broken bones and other INJURIES. Hearst’s work on hyponymy was continued and refined in [7, 26], whereas others used such techniques for extracting, for instance, part-of relations [4, 24, 11]. Unsupervised methods for relation extraction have been used for concept clustering and ontology learning by, among others, [8, 2, 3, 32]. In all of this work, patterns for the chosen sets of relations are hand-defined. In other work, a semi-supervised methodology was used, in which patterns were automatically discovered by starting from pairs of instances of the chosen relation (‘seeds’) and then finding in corpora instances of these pairs [17, 22].

Even high-quality patterns tend to extract from corpora a lot of spurious information and/or information about semantic relations other than concept attributes. In [23], we proposed to filter this information by identifying a set of semantic relations that specified concept attributes and training a classifier to recognize these relations. Building on work by Pustejovsky [25] and Guarino [16] we specified a set of relations often used to define concepts that included quality, part-of, related-object, activity, and related-agent; we collected from the Web candidate instances of these relations for a dataset of 400 concepts using (very general) patterns; hand-annotated these data; and used these annotations to train and evaluate two decision tree classifiers—a 5-way classifier and a binary classifier (attribute / not attribute). These classifiers used a great number of features, including:

- morphological features, extracted through heuristics, such as the information that a particular noun might be derived from an adjective or a verb, which is useful to identify qualities and activities respectively;
question patterns, that is, the frequencies obtained by querying the Web with questions of the form “What is the ATTr of CONCEPT” or “When is the ATTr of CONCEPT”;

features of features, i.e., the top attributes of these potential attributes, extracted using the same patterns that we used to extract attributes; and finally

feature use, that is, information about the respective frequency of the use of these nouns as attributes (“the ATTr of the *”) or concepts (“the * of the CONCEPT”).

The binary classifier achieved an accuracy of 81.82% as evaluated through cross-validation, which corresponds to an F value of .892 at recognizing attributes and .417 at recognizing non-attributes. The 5-way classifier achieved an accuracy of around 80% at cross-validation, corresponding to an F value over .8 for quality, activity, and part / related object, of .95 for related agent, and .538 for not-an-attribute.

Much of the work on ontology learning just discussed - both unsupervised and supervised - draws its inspiration from work in linguistics (such as Pustejovsky’s) or formal ontology (such as Guarrino’s). We do not know of any work attempting to build concept descriptions on the basis of psychological work on attributes. This is the goal of the present line of research.

2.2 Concept representations in Psychology and the Wu and Barsalou classification of attributes

Psychological theories of concepts [21] also view them as being associated with bundles of ‘features’; ever since Rosch [27] psychologists have been concerned with eliciting such features from human subjects, producing feature databases called feature norms. In the feature generation task human subjects are invited to list what they believe are the most relevant properties for a set of concepts. The experimenter then processes the conceptual descriptions provided by the subjects and registers the final representation in the feature norms. Inspecting feature norms is one of the best ways we have at the moment to identify the most important relations that organize our cognition.

For our experiments in relation extraction we chose the largest feature norm in existence, produced by McRae and colleagues [20]. The norm lists the conceptual descriptions for 541 basic level concepts representing living and not living things. Apart from its size, another advantage of this feature norm is that it is classified using a schema, a modified version of the taxonomy derived by Wu and Barsalou (henceforth: W&B) [33] from studies of human perception. The principles that Wu and Barsalou considered when they built the taxonomy were derived from cognitive psychology: the introspective experience of the subjects when they generate feature norms, the cognitive function relations denoting the function an entity typically fulfills—for instance, “Airplane used for transportations”, the modality specific regions of the brain ontological kinds of Keil and the frame theory of Fillmore.

Wu and Barsalou’s taxonomy has two levels; at the coarsest level the features are classified as taxonomic properties, entity properties, situational properties or introspective properties. The taxonomic properties classify an entity from a taxonomic point of view, entity properties denote general properties of an entity, situational properties are characteristic to situations and introspective properties are properties of subject mental states. At the second level each mentioned category of properties is further subdivided. The modified W&B taxonomy has 27 categories at the second level.

In our experiments, we used the W&B taxonomy as our set of relations, and the pairs concept-feature as instances of these taxonomic relations. We only considered the most representative relations in the taxonomy, characterized as relations that instantiate at least 60 concept-feature pairs. In the experiments reported here we focused on a subset of 6 of these relation types. With the exception of the relation External Surface that holds between a noun and an adjective, all other relations are relations between nominals. Such relations are well studied and the researchers devised both weakly supervised and supervised methods for learning them. We will briefly describe the six relations for which we prepared training sets:

External Surface : External surface properties are those properties of an entity that are perceived on or beyond the entity’s surface, including shape, color, pattern, texture, size, touch, smell, taste. These properties usually denote the qualities of a concrete object, as in, e.g.: “the car is red”. In this case red is a quality of the concept car and the quality resides on the external surface of the car.

Function : Function relations are relations denoting the function an entity typically fulfills—for instance, “Airplane used for transportations”.

Internal Component : These properties denote internal components of an entity and are its hidden parts. (The W&B taxonomy distinguishes between the internal components of an object and the external components of an object; from the point of view of relation learning we regard both of them as simply parts.)

Origin : Origin properties are those properties denoting the origin of an entity, as in “Cigar made in Cuba”.

Participant : Participant properties denote agents who typically use an entity, performs an action on it or interacts with other participants. Example: “desk used by students”.

Superordinate : Superordinate is the well known is-a relation.

3 KERNEL METHODS FOR RELATION EXTRACTION

To extract semantic relations between nominals, we used a supervised approach based on kernel methods that exploits exclusively shallow linguistic information, such as tokenization, sentence splitting, part-of-speech (PoS) tagging and lemmatization [13, 14, 12].

The basic idea behind kernel methods [29] is first to embed the input data in a suitable feature space, and then use a linear algorithm to discover nonlinear pattern in the input space. Typically, the mapping is performed implicitly by a so-called kernel function. The kernel function is a similarity measure between the input data that depends exclusively on the specific data type and domain. Characterizing the similarity of the inputs plays a crucial role in determining the success or failure of the learning algorithm, and it is one of the central questions in the field of machine learning. Kernel methods allow designing a modular system, in which the kernel functions operate as an interface between the data and the learning algorithm. Thus the kernel function is the only domain specific module of the system, while the learning algorithm is a general purpose component. Potentially any kernel function can work with any kernel-based algorithm. In our approach we use support vector machines (SVMs) [31, 9].

In order to implement the approach based on shallow linguistic information we employed a linear combination of kernels. Different works on NLP [10, 15, 35] empirically demonstrate the effectiveness of combining kernels in this way, showing that the combined kernel always improves the performance of the individual ones.

Here, we combine two families of kernel functions: global context kernels and local context kernels, where the single kernels are
explicitly calculated as follows

\[ K(x_1, x_2) = \frac{\phi(x_1) \cdot \phi(x_2)}{\|\phi(x_1)\| \cdot \|\phi(x_2)\|}, \tag{1} \]

where \(\phi(\cdot)\) is the embedding vector and \(\|\cdot\|\) is the 2-norm.

3.1 Global Context Kernel

[6], [13] and [14] have shown the validity of the assumption that relations between entities are generally expressed using only words that appear simultaneously in one of the following three contexts.

Fore-Between Tokens before and between the two entities, e.g. "the head of [ORG], Dr. [PER]".

Between Only tokens between the two entities, e.g. "[ORG] [PER]".

Between-After Tokens between and after the two entities, e.g. "[PER], a [ORG] professor".

Successively, it has been shown that this assumption is also correct for semantic relations between nominals [12]. The global context kernel operates on the contexts defined above, where each context is modeled using a bag-of-words kernel. More formally, given a relation example \(R\), we represent a context \(C\) as a row vector

\[ \phi_C(R) = (t f(t_1, C), t f(t_2, C), \ldots, t f(t_l, C)) \in \mathbb{R}^l, \tag{2} \]

where the function \(t f(t_i, C)\) records how many times a particular token \(t_i\) is used in \(C\). Note that this approach differs from the standard bag-of-words as punctuation and stop words are included in \(\phi_C\), while the nominals are not. The classification performance can be further improved by extending \(\phi_C\) to embed n-grams of (contiguous) tokens up to \(n = 3\). By substituting \(\phi_C\) into Equation 1, we obtain the n-gram kernel \(K_n\), which counts uni-grams, bi-grams, ..., n-grams that two patterns have in common. The global context kernel \(K_{GC}(R_1, R_2)\) is then defined as

\[ K_{GC}(R_1, R_2) = K_{FB}(R_1, R_2) + K_{B}(R_1, R_2) + K_{BA}(R_1, R_2), \tag{3} \]

where \(K_{FB}, K_{B}\) and \(K_{BA}\) are respectively the n-gram kernels that operate on the Fore-Between, Between and Between-After patterns and \(R_1\) and \(R_2\) are the examples compared by the kernel function.

3.2 Local Context Kernel

The local contexts of the candidate entities/nominals can provide useful clues for determining their role within the relation. As typically done in entity recognition, we represent each local context by using the following basic features: the token itself, the lemma, the PoS tag, the stem, and orthographic features, within a text window.

Formally, given a relation example \(R\), a local context \(L = t_{-w}, \ldots, t_{-1}, t_0, t_1, \ldots, t_w\) is represented as a row vector

\[ \psi_L(R) = (f_1(L), f_2(L), \ldots, f_m(L)) \in \{0, 1\}^m, \tag{4} \]

where \(f_i\) is a feature function that returns 1 if it is active in the specified position of \(L\), 0 otherwise. For example, the orthographic function \(CAP(t_0)\) is the \(jth\) component of the vector \(\psi_L(R)\) and it is 1

\[ \text{if and only if the token } t_0 \text{ in the local context } L \text{ of } R \text{ is capitalized.} \]

The local context kernel is defined as follows

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

where \(K_{LC}\) and \(K_{right}\) are defined by substituting the embedding of the left and right local context into Equation 1 respectively. Notice that \(K_{LC}\) differs substantially from \(K_{GC}\) as it considers the ordering of the tokens and the feature space is enriched with PoS, lemma and orthographic features.

3.3 Shallow Linguistic Kernel

The shallow linguistic kernel \(K_{SL}(R_1, R_2)\) is defined as

\[ K_{SL}(R_1, R_2) = K_{LC}(R_1, R_2). \tag{6} \]

It follows directly from the explicit construction of the feature space and from closure properties of kernels that \(K_{SL}\) is a valid kernel.

3.4 Bag-of-Words Kernel

The bag-of-words kernel \(K_{BoW}(R_1, R_2)\) is defined as the global context kernel but it operates on the whole sentence. We defined this kernel for comparison only; notice that it is not used in the shallow linguistic kernel.

4 EXPERIMENTAL SETUP

4.1 Data Collection

We gathered instances from the McRae feature norm for each of the six semantic relations discuss in the previous sections. In the process of feature processing we normalized, lemmatized and part of speech tagged all the relevant features. For example the database concept-feature pair (dog is_animal) will become (dog, animal (noun)). The lemmatization and part of speech tagging was performed using TreeTagger, a language independent part of speech tagger. In the next step we extracted sentences that contain the seeds from a very large web corpus, UKWAC [7], which contains about 2 billions tokens. In all extracted sentences the distance between seeds is at most 5 words including the punctuation. The sentence extraction was performed using CQP interface. We made sure that training set does not contain the same sentence twice and that each instance is present in approximately the same number of sentences. For each relation we collected a set of 500 sentences. Each sentence was annotated as a positive, negative or don’t know example. A sentence is a positive example if the relation between the concept and the feature is expressed in the sentence and negative otherwise. The don’t know option is used when the annotator is unsure if the relation holds or not. Two annotators annotated the set of sentences for each relation. The annotation was performed with GUI based and annotation tool developed specially for the task. Table 1 gives the intra-annotator agreement for each of the six semantic relations using the Kappa statistic [7].

4.2 Preparations for processing

Sentences have been tokenized, lemmatized, and POS tagged with TextPro. We considered each relation as a different binary classification task, and each sentence in the data set is a positive or negative example for the relation. All the experiments were performed using jSRE. The results were obtained by 10-fold cross-validation.

5 RESULTS

Table 2 shows the performance of the shallow linguistic kernel for 6 relations and the micro-average result. Majority and all true are the baselines that return the majority class and always true, respectively. BoW is the bag-of-words kernel, it provides a stronger baseline.

<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>Function</td>
<td>0.62</td>
</tr>
<tr>
<td>Internal Component</td>
<td>0.78</td>
</tr>
<tr>
<td>Origin</td>
<td>0.63</td>
</tr>
<tr>
<td>Participant</td>
<td>0.5</td>
</tr>
<tr>
<td>Superordinate</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 1. Inter-annotator agreement on the six Wu-Barsalou relations

The main point to notice at this stage is that these results are comparable in quality to the results obtained in our previous work, yet were obtained only using the linguistic context in which the pairs occurred, as opposed to the numerous features used in our previous work. Finally, the shallow linguistic kernel significantly outperforms the 3 baselines. This result, together with the results obtained extracting relations between biomedical entities [13], named entities [12], and nominals [14], shows the effectiveness of this simple kernel for relation extraction.

6 CONCLUSIONS

These preliminary results suggest that the relations we chose can be successfully extracted from text with minimum training material and that the supervised classification scheme used in these experiments does not need much in terms of features to achieve these results. In subsequent work we will use these attributes for concept clustering and experiment with extracting more relations from the W&K taxonomy.

ACKNOWLEDGEMENTS

Claudio Giuliano and Lorenza Romano are supported by the X-Media project (http://www.x-media-project.org), sponsored by the European Commission as part of the Information Society Technologies (IST) program under EC grant number IST-FP6-026978. We would like to thank the members of the CIMEC Concepts Working Group, and in particular Marco Baroni and Brian Murphy, for help and suggestions.

REFERENCES


