**The Domain Restriction Hypothesis: Relating Term Similarity and Semantic Consistency**

Alfio Massimiliano Gliozzo  
ITC-irst  
Trento, Italy  
gliozzo@itc.it

Marco Pennacchiotti  
University of Rome Tor Vergata  
Rome, Italy  
pennacchiotti@info.uniroma2.it

Patrick Pantel  
USC, Information Sciences Institute  
Marina del Rey, CA  
pantel@isi.edu

**Abstract**

In this paper, we empirically demonstrate what we call the domain restriction hypothesis, claiming that semantically related terms extracted from a corpus tend to be semantically coherent. We apply this hypothesis to define a post-processing module for the output of Espresso, a state of the art relation extraction system, showing that irrelevant and erroneous relations can be filtered out by our module, increasing the precision of the final output. Results are confirmed by both quantitative and qualitative analyses, showing that very high precision can be reached.

**1 Introduction**

Relation extraction is a fundamental step in many natural language processing applications such as learning ontologies from texts (Buitelaar et al., 2005) and Question Answering (Pasca and Harabagiu, 2001).

The most common approach for acquiring concepts, instances and relations is to harvest semantic knowledge from texts. These techniques have been largely explored and today they achieve reasonable accuracy. Harvested lexical resources, such as concept lists (Pantel and Lin, 2002), facts (Etzioni et al., 2002) and semantic relations (Pantel and Pennacchiotti, 2006) could be then successfully used in different frameworks and applications.

The state of the art technology for relation extraction primarily relies on pattern-based approaches (Snow et al., 2006). These techniques are based on the recognition of the typical patterns that express a particular relation in text (e.g. “X such as Y” usually expresses an is-a relation). Yet, text-based algorithms for relation extraction, in particular pattern-based algorithms, still suffer from a number of limitations due to complexities of natural language, some of which we describe below.

**Irrelevant relations.** These are valid relations that are not of interest in the domain at hand. For example, in a political domain, “Condoleezza Rice is a football fan” is not as relevant as “Condoleezza Rice is the Secretary of State of the United States”. Irrelevant relations are ubiquitous, and affect ontology reliability, if used to populate it, as the relation drives the wrong type of ontological knowledge.

**Erroneous or false relations.** These are particularly harmful, since they directly affect algorithm precision. A pattern-based relation extraction algorithm is particularly likely to extract erroneous relations if it uses generic patterns, which are defined in (Pantel and Pennacchiotti, 2006) as broad coverage, noisy patterns with high recall and low precision (e.g. “X of Y” for part-of relation). Harvesting algorithms either ignore generic patterns (Hearst, 1992) (affecting system recall) or use manually supervised filtering approaches (Girju et al., 2006) or use completely unsupervised Web-filtering methods (Pantel and Pennacchiotti, 2006). Yet, these methods still do not sufficiently mitigate the problem of erroneous relations.

**Background knowledge.** Another aspect that makes relation harvesting difficult is related to the
nature of semantic relations: relations among enti-

ties are mostly paradigmatic (de Saussure, 1922),
and are usually established in absentia (i.e., they are
not made explicit in text). According to Eco’s posi-
tion (Eco, 1979), the background knowledge (e.g.
“persons are humans”) is often assumed by the
writer, and thus is not explicitly mentioned in text.
In some cases, such widely-known relations can be
captured by distributional similarity techniques but
not by pattern-based approaches.

Metaphorical language. Even when paradigmatic
relations are explicitly expressed in texts, it can
be very difficult to distinguish between facts and
metaphoric usage (e.g. the expression “My mind is
a pearl” occurs 17 times on the Web, but it is clear
that mind is not a pearl, at least from an ontological
perspective).

The considerations above outline some of the dif-
ficulties of taking a purely lexico-syntactic approach
to relation extraction. Pragmatic issues (background
knowledge and metaphorical language) and onto-
logical issues (irrelevant relation) can not be solved
at the syntactic level. Also, erroneous relations can
always arise. These considerations lead us to the
intuition that extraction can benefit from imposing
some additional constraints.

In this paper, we integrate Espresso with a lex-
cal distribution technique modeling semantic co-
herence through semantic domains (Magnini et al.,
2002). These are defined as common discourse top-
ics which demonstrate lexical coherence, such as
ECONOMICS or POLITICS. We explore whether se-
mantic domains can provide the needed additional
constraints to mitigate the acceptance of erroneous
relations. At the lexical level, semantic domains
identify clusters of (domain) paradigmatically re-
lated terms. We believe that the main advantage of
adopting semantic domains in relation extraction is
that relations are established mainly among terms in
the same Domain, while concepts belonging to dif-
ferent fields are mostly unrelated (Gliozzo, 2005),
as described in Section 2. For example, in a chem-
istry domain, an is-a will tend to relate only terms of
that domain (e.g., nitrogen is-a element), while out-
of-domain relations are likely to be erroneous e.g.,
driver is-a element.

By integrating pattern-based and distributional ap-
proaches we aim to capture the two characteristic
properties of semantic relations:

- **Syntagmatic properties**: if two terms X and
  Y are in a given relation, they tend to co-
  occur in texts, and are mostly connected by spe-
  cific lexical-syntactic patterns (e.g., the patter
  “X is a Y” connects terms in is-a relations).
  This aspect is captured using a pattern-based
  approach.

- **Domain properties**: if a semantic relation
  among two terms X and Y holds, both X
  and Y should belong to the same semantic
domain (i.e. they are semantically coherent),
  where semantic domains are sets of terms
  characterized by very similar distributional
  properties in a (possibly domain specific)
corpus.

In Section 2, we develop the concept of semantic do-
main and an automatic acquisition procedure based
on Latent Semantic Analysis (LSA) and we provide
empirical evidence of the connection between rela-
tion extraction and domain modelling. Section 3 de-
scribes the Espresso system. Section 4 concerns our
integration of semantic domains and Espresso. In
Section 5, we evaluate the impact of our LSA do-
main restriction module on improving a state of the
art relation extraction system. In Section 6 we draw
some interesting research directions opened by our
work.

2 Semantic Domains

Semantic domains are common areas of human
discussion, which demonstrate lexical coherence,
such as ECONOMICS, POLITICS, LAW, SCIENCE,
(Magnini et al., 2002). At the lexical level, se-
mantic domains identify clusters of (domain) related
lexical-concepts, i.e. sets of highly paradigmatically
related words also known as Semantic Fields.

In the literature, semantic domains have been
inferred from corpora by adopting term clustering
methodologies (Gliozzo, 2005), and have been used
for several NLP tasks, such as Text Categorization
and Ontology Learning (Gliozzo, 2006).

Semantic domains can be described by Domain
Models (DMs) (Gliozzo, 2005). A DM is a com-
putational model for semantic domains, that represents domain information at the term level, by defining a set of term clusters. Each cluster represents a Semantic Domain, i.e. a set of terms that often co-occur in texts having similar topics. A DM is represented by a \( k \times k' \) rectangular matrix \( D \), containing the domain relevance for each term with respect to each domain, as illustrated in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>MEDICINE</th>
<th>COMPUTER_SCIENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AIDS</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>virus</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>laptop</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Example of a Domain Model

DMs can be acquired from texts in a completely unsupervised way by exploiting a lexical coherence assumption. To this end, term clustering algorithms can be used with each cluster representing a Semantic Domain. The degree of association among terms and clusters, estimated by the learning algorithm, provides a domain relevance function. For our experiments we adopted a clustering strategy based on LSA (Deerwester et al., 1990), following the methodology described in (Gliozzo, 2005). The input of the LSA process is a term-by-document matrix \( T \) reporting the term frequencies in the whole corpus for each term. The matrix is decomposed by means of a Singular Value Decomposition (SVD), identifying the principal components of \( T \). This operation is done off-line, and can be efficiently performed on large corpora. SVD decomposes \( T \) into three matrixes \( T \approx \mathbf{V} \Sigma \mathbf{U}^T \) where \( \Sigma_k \) is the diagonal \( k \times k \) matrix containing the highest \( k' \ll k \) eigenvalues of \( T \) on the diagonal, and all the remaining elements are 0. The parameter \( k' \) is the dimensionality of the domain and can be fixed in advance\(^1\). Under this setting we define the domain matrix \( D_{\text{LSA}} \) as

\[
D_{\text{LSA}} = \mathbf{I}^N \sqrt{\Sigma_k'}
\]

where \( \mathbf{I}^N \) is a diagonal matrix such that \( i_{i,i} = 1/\sqrt{w_i'^2} \) and \( w_i' \) is the \( i^{th} \) row of the matrix \( \mathbf{V} \sqrt{\Sigma_k'} \).

\(^1\)It is not clear how to choose the right dimensionality. In our experiments we used 100 dimensions.

\(^2\)Details of this operation can be found in (Gliozzo, 2005).

Once a DM has been defined by the matrix \( D \), the Domain Space is a \( k' \) dimensional space, in which both texts and terms are associated to Domain Vectors (DVs), i.e. vectors representing their domain relevancies with respect to each domain. The DV \( t_i' \) for the term \( t_i \in \mathcal{V} \) is the \( i^{th} \) row of \( D \), where \( \mathcal{V} = \{ t_1, t_2, \ldots, t_k \} \) is the vocabulary of the corpus. The domain similarity \( \phi_d(t_i, t_j) \) among terms is then estimated by the cosine among their corresponding DVs in the Domain Space, defined as follows:

\[
\phi_d(t_i, t_j) = \frac{\langle t_i', t_j' \rangle}{\sqrt{\langle t_i', t_i' \rangle \langle t_j', t_j' \rangle}}
\]

Figure 1: Probability of finding paradigmatic relations

The main advantage of adopting semantic domains for relation extraction is that they allow us to impose a domain restriction on the set of candidate pairs of related terms. In fact, semantic relations can be established mainly among terms in the same Semantic Domain, while concepts belonging to different fields are mostly unrelated.

To show the validity of the domain restriction we conducted a preliminary experiment, contrasting the probability for two words to be related in WordNet (Magnini and Cavaglià, 2000) with their domain similarity, measured in the Domain Space induced from the British National Corpus. In particular, for each couple of words, we estimated the domain similarity, and we collected word pairs in sets characterized by different ranges of similarity (e.g. all the pairs between 0.8 and 0.9). Then we estimated the
probability of each couple of words in different sets to be linked by a semantic relation in WordNet, such as synonymy, hyperonymy, co-hyponymy and domain in WordNet Domains (Magnini et al., 2002). Results in Figure 1 show a monotonic crescent relation between these two quantities. In particular the probability for two words to be related tends to 0 when their similarity is negative (i.e., they are not domain related), supporting the basic hypothesis of this work. In Section 4 we will show that this property can be used to improve the overall performances of the relation extraction algorithm.

3 The pattern-based Espresso system

Espresso (Pantel and Pennacchiotti, 2006) is a corpus-based general purpose, broad, and accurate relation extraction algorithm requiring minimal supervision, whose core is based on the framework adopted in (Hearst, 1992). Espresso introduces two main innovations that guarantee high performance: (i) a principled measure for estimating the reliability of relational patterns and instances; (ii) an algorithm for exploiting generic patterns. Generic patterns are broad coverage noisy patterns (high recall and low precision), e.g. “X of Y” for the part-of relation. As underlined in the introduction, previous algorithms either required significant manual work to make use of generic patterns, or simply ignore them. Espresso exploits an unsupervised Web-based filtering method to detect generic patterns and to distinguish their correct and incorrect instances.

Given a specific relation (e.g. is-a) and a POS-tagged corpus, Espresso takes as input few seed instances (e.g. nitrogen is-a element) or seed surface patterns (e.g. X/NN such/JJ as/N Y/NN). It then incrementally learns new patterns and instances by iterating on the following three phases, until a specific stop condition is met (i.e., new patterns are below a pre-defined threshold of reliability).

**Pattern Induction.** Given an input set of seed instances $I$, Espresso infers new patterns connecting as many instances as possible in the given corpus. To do so, Espresso uses a slight modification of the state of the art algorithm described in (Ravichandran and Hovy, 2002). For each instance in input, the sentences containing it are first retrieved and then generalized, by replacing term expressions with a terminological label using regular expressions on the POS-tags. This generalization allows to ease the problem of data sparseness in small corpora. Unfortunately, as patterns become more generic, they are more prone to low precision.

**Pattern Ranking and Selection.** Espresso ranks all extracted patterns using a reliability measure $r_\pi$ and discards all but the top-$k$ $P$ patterns, where $k$ is set to the number of patterns from the previous iteration plus one. $r_\pi$ captures the intuition that a reliable pattern is one that is both highly precise and one that extracts many instances. $r_\pi$ is formally defined as the average strength of association between a pattern $p$ and each input instance $i$ in $I$, weighted by the reliability $r_i$ of the instance $i$ (described later):

$$r_\pi(p) = \frac{\sum_{i \in I} \left( \frac{\text{pmi}(i,p)}{\max \text{pmi}} \ast r_i(i) \right)}{|I|}$$

where $\text{pmi}(i,p)$ is the pointwise mutual information (pmi) between $i$ and $p$ (estimated with Maximum Likelihood Estimation), and $\max \text{pmi}$ is the maximum pmi between all patterns and all instances.

**Instance Extraction, Ranking, Selection.** Espresso extracts from the corpus the set of instances $I$ matching the patterns in $P$. In this phase generic patterns are detected, and their instances are filtered, using a technique described in detail in (Pantel and Pennacchiotti, 2006). Instances are then ranked using a reliability measure $r_i$, similar to that adopted for patterns. A reliable instance should be highly associated with as many reliable patterns as possible:

$$r_i(i) = \frac{\sum_{p \in P} \left( \frac{\text{pmi}(i,p)}{\max \text{pmi}} \ast r_\pi(i) \right)}{|P|}$$

Finally, the best scoring instances are selected for the following iteration. If the number of extracted instances is too low (as often happens in small corpora) Espresso enters an expansion phase, in which instances are expanded by using web based and syntactic techniques.
The output Espresso is a list of instances \( i = (X, Y) \in I \), ranked according to \( r_i(i) \). This score accounts for the syntagmatic similarity between \( X \) and \( Y \), i.e., how strong is the co-occurrence of \( X \) and \( Y \) in texts with a given pattern \( p \).

A key role in the Espresso algorithm is played by the reliability measures. The accuracy of the whole extraction process is in fact highly sensitive to the ranking of patterns and instances because, at each iteration, only the best scoring entities are retained. For instance, if an erroneous instance is selected after the first iteration, it could in theory affect the following pattern extraction phase and cause drift in consequent iterations. This issue is critical for generic patterns (where precision is still a problem, even with Web-based filtering), and could sometimes also affect non-generic patterns.

It would be then useful to integrate Espresso with a technique able to retain only very precise instances, without compromising recall. As syntagmatic strategies are already in place, another strategy is needed. In the next Section, we show how this can be achieved using instance domain information.

4 Integrating syntagmatic and domain information

The strategy of integrating syntagmatic and domain information has demonstrated to be fruitful in many NLP tasks, such as Word Sense Disambiguation and open domain Ontology Learning (Gliozzo, 2006). According to the structural view (de Saussure, 1922), both aspects contribute to determine the linguistic value (i.e. the meaning) of words: the meaning of lexical constituents is determined by a complex network of semantic relations among words. This suggests that relation extraction can benefit from accounting for both syntagmatic and domain aspects at the same time.

To demonstrate the validity of this claim we can explore many different integration schemata. For example we can restrict the search space (i.e. the set of candidate instances) to the set of all those terms belonging to the same domain. Another possibility is to exploit a similarity metric for domain relatedness to re-rank the output instances \( I \) of Espresso, hoping that the top ranked ones will mostly be those which are correct. One advantage of this latter methodology is that it can be applied to the output of any relation extraction system without any modification to the system itself. In addition, this methodology can be evaluated by adopting standard Information Retrieval (IR) measures, such as mean average precision (see Section 5). Because of these advantages, we decided to adopt the re-ranking procedure.

The procedure is defined as follows: each instance extracted by Espresso is assigned a Domain Similarity score \( \phi_d(X, Y) \) estimated in the domain space according to Equation 2; a higher score is then assigned to the instances that tend to co-occur in the same documents in the corpus. For example, the candidate instances ethanol is-a nonaromatic alcohol has a higher score than ethanol is-a something, as ethanol and alcohol are both from the chemistry domain, while something is a generic term and is thus not associated to any domain.

Instances are then re-ranked according to \( \phi_d(X, Y) \), which is used as the new index of reliability instead of the original reliability scores of Espresso. In Subsection 5.2 we will show that the re-ranking technique improves the original reliability scores of Espresso.

5 Evaluation

In this Section we evaluate the benefits of applying the domain information to relation extraction (ESP-LSA), by measuring the improvements of Espresso due to domain based re-ranking.

5.1 Experimental Settings

As a baseline system, we used the ESP- implementation of Espresso described in (Pantel and Pennacchiotti, 2006). ESP- is a fully functioning Espresso system, without the generic pattern filtering module (ESP+). We decided to use ESP- for two main reasons. First, the manual evaluation process would have been too time consuming, as ESP+ extracts thousands of relations. Also, the small scale experiment for EXP- allows us to better analyse and compare the results.

To perform the re-ranking operation, we acquired a Domain Model from the input corpus itself. To this aim we performed a SVD of the term by document matrix \( T \) describing the input corpus, indexing all the candidate terms recognized by Espresso.
As an evaluation benchmark, we adopted the same instance sets extracted by ESP- in the experiment described in (Pantel and Pennacchiotti, 2006). We used an input corpus of 313,590 words, a college chemistry textbook (Brown et al. 2003), pre-processed using the Alembic Workbench POS-tagger (Day et al. 1997). We considered the following relations: **is-a**, **part-of**, **reaction** (a relation of chemical reaction among chemical entities) and **production** (a process or chemical element/object producing a result). ESP- extracted 200 *is-a*, 111 *part-of*, 40 *reaction* and 196 *production* instances.

### 5.2 Quantitative Analysis

The experimental evaluation compared the accuracy of the ranked set of instances extracted by ESP- with the re-ranking produced on these instances by ESP-LSA. By analogy to IR, we are interested in extracting positive instances (i.e. semantically related words). Accordingly, we utilize the standard definitions of precision and recall typically used in IR. Table 2 reports the Mean Average Precision obtained by both ESP- and ESP-LSA on the extracted relations, showing the substantial improvements on all the relations due to domain based re-ranking.

<table>
<thead>
<tr>
<th></th>
<th>ESP-</th>
<th>ESP-LSA</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>is-a</em></td>
<td>0.54</td>
<td>0.75</td>
<td>+0.21</td>
</tr>
<tr>
<td><em>part-of</em></td>
<td>0.65</td>
<td>0.82</td>
<td>+0.17</td>
</tr>
<tr>
<td><em>react</em></td>
<td>0.75</td>
<td>0.82</td>
<td>+0.07</td>
</tr>
<tr>
<td><em>produce</em></td>
<td>0.55</td>
<td>0.62</td>
<td>+0.07</td>
</tr>
</tbody>
</table>

Table 2: Mean Average Precision reported by ESP- and ESP-LSA

Figures 2, 3, 4 and 5 report the precision/recall curves obtained for each relation, estimated by measuring the precision / recall at each point of the ranked list. Results show that precision is very high especially for the top ranked relations extracted by ESP-LSA. Precision reaches the upper bound for the top ranked part of the **part-of** relation, while it is close to 0.9 for the **is-a** relation. In all cases, the precision reported by the ESP-LSA system surpass those of the ESP- system at all recall points.

### 5.3 Qualitative Analysis

Table 3 shows the best scoring instances for ESP- and ESP-LSA on the evaluated relations. Results show that ESP-LSA tends to assign a much lower score to erroneous instances, as compared to the original Espresso reliability ranking. For example for the **part-of** relation, the ESP- ranks the erroneous instance *geometry part-of ion* in 23th position, while ESP-LSA re-ranks it in 92nd. In this case, a lower score is assigned because *geometry* is not particularly tied to the domain of chemistry. Also, ESP-LSA tends to penalize instances derived from parsing/tokenization errors:
Figure 4: Syntagmatic vs. Domain ranking for the part-of relation

Figure 5: Syntagmatic vs. Domain ranking for the react relation

\emph{Binary hydrogen compounds hydrogen react elements} is 16th for ESP-, while in the last tenth of the ESP-LSA. In addition, out-of-domain relations are successfully interpreted by ESP-LSA. For example, \emph{magnesium metal react elemental oxygen} lies at the end of ESP- rank, as there are not enough syntagmatic evidence (co-occurrences) that let the instance emerge. The domain analysis of ESP-LSA promotes this instance to the 2nd rank position. However, in few cases, the strategy adopted by ESP-LSA tends to promote erroneous instances (e.g. \emph{high voltage produce voltage}). Yet, results show that these are isolated cases.

6 Conclusion and future work

In this paper, we propose the domain restriction hypothesis, claiming that semantically related terms extracted from a corpus tend to be semantically coherent. Applying this hypothesis, we presented a new method to improve the precision of pattern-based relation extraction algorithms, where the integration of domain information allows the system to filter out many irrelevant relations, erroneous candidate pairs and metaphorical language relational expressions, while capturing the assumed knowledge required to discover paradigmatic associations among terms. Experimental evidences supports this claim both qualitatively and quantitatively, opening a promising research direction, that we plan to explore much more in depth. In the future, we plan to compare LSA to other term similarity measures, to train the LSA model on large open domain corpora and to apply our technique to both generic and specific corpora in different domains. We want also to increase the level of integration of the LSA technique in the Espresso algorithm, by using LSA as an alternative reliability measure at each iteration. We will also explore the domain restriction property of semantic domains to develop open domain ontology learning systems, as proposed in (Gliozzo, 2006).

The \emph{domain restriction hypothesis} has potential to greatly impact many applications where matching textual expressions is a primary component. It is our hope that by combining existing ranking strategies in applications such as information retrieval, question answering, information extraction and document classification, with knowledge of the coherence of the underlying text, one will see significant improvements in matching accuracy.
<table>
<thead>
<tr>
<th>Relation</th>
<th>ESP-</th>
<th>ESP - LSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>X is-a Y</td>
<td>Aluminum ; metal nitride; ion ; strong Br heat_flow ; calorimeter complete_ionic_equation ; spectator</td>
<td>F ; electronegative_atoms O ; electronegative_atoms NaCN ; cyanide_salt NaCN ; cyanide_salts</td>
</tr>
<tr>
<td>X part-of Y</td>
<td>elements ; compound composition ; substance blocks ; tripeptide elements ; sodium_chloride</td>
<td>amino_acid_building_blocks ; tripeptide acid_building_blocks ; tripeptide powdered_zinc_meta ; battery building_blocks ; tripeptide</td>
</tr>
<tr>
<td>X react Y</td>
<td>hydrazine ; water magnesium_metal ; hydrochloric_acid magnesium ; oxygen magnesium_metal ; acid</td>
<td>magnesium_meta ; elemental_oxygen nitrogen ; ammonia sodium_meta ; chloride carbon_dioxide ; methane</td>
</tr>
<tr>
<td>X produce Y</td>
<td>bromine ; bromide oxygen ; oxide common_fuels ; dioxide kidneys ; stones</td>
<td>high_voltage ; voltage reactions ; reactions dr_jekyll ; hyde yellow_pigments ; green_pigment</td>
</tr>
</tbody>
</table>

Table 3: Top scoring relations extracted by ESP- and ESP-LSA.

Acknowledgments

Thanks to Roberto Basili for his precious comments, suggestions and support. Alfio Gliozzo was supported by the OntoText project, funded by the Autonomous Province of Trento under the FUP-2004, and the FIRB founded project N.RBIN045PXH.

References


