Domain Driven Technologies for Natural Language Processing

Alfio Massimiliano Gliozzo
ITC-irst
gliozzo@itc.it
http://tcc.itc.it/people/gliozzo/

28/07/06

Abstract

Semantic Domains are a matter of recent interest in Computational Linguistics. Domain Models allow us to represent lexical ambiguity and variability by inducing soft clusters of terms from corpora. Semantic Domains shows many interesting properties: lexical ambiguity inside a domain is sensibly reduced, paradigmatic relations are mainly established among terms in the same domain, semantic domains are stable among languages. These properties have been exploited to develop innovative technologies for a wide range of different Natural Language processing tasks, such as Text Categorization, Word Sense Disambiguation, Cross Language Text Categorization and Ontology Learning from Texts. The Domain Driven systems we developed are minimally supervised and can be easily ported across domains and languages as they rely on semi-supervised learning techniques based on Kernel Methods, reporting the state of the art performances in many cases.
Outline

- Semantic Domains and Linguistic Theory
  - Ambiguity and Variability
  - Semantic Fields
  - Semantic Domains and their properties
- Domain Models
  - Definition
  - Acquisition from Corpora
  - The Domain Kernel
  - Multilingual Domain Models
- Domain Driven Technologies
  - Text Categorization
  - Word Sense Disambiguation
  - Cross Language Text Categorization
  - Ontology Learning from Texts
Ambiguity and Variability

- **Ambiguity**: The same term refers to different concepts
- **Variability**: The same concept can be expressed by different terms
- e.g. the WordNet model (structural)
  - [floor, level, storey, story]
  - [narrative, narration, story, tale]
- **Lexical Semantics**
  - Predictive models (theories) for ambiguity and variability
  - Structuralist view: linguistic value (meaning) is determined by relations among words

The Theory of Semantic Fields (Trier,31)

- A **semantic field** is a set of concepts that covers a whole spectrum of phenomena, a domain (Structuralist paradigm)
- e.g. Medicine: hospital, doctor, virus, HIV, ...
- Terms in the field are highly **paradigmatically related** (e.g. HIV is_a virus)
- Terms in different fields are **unrelated** (e.g. HIV ? algorithm)
- Terms in more than one field are **ambiguous** (e.g. virus[Medicine] vs. virus[CS])
- **Multilinguality**: terms in different languages tend to be structured into the same fields (e.g. computer, laptop, stampante, monitor [CS])
**Semantic Domains**

- **Semantic Domains** are:
  - clusters of very closely related **texts**
  - clusters of very closely related **terms**
- Semantic Domains are Semantic Fields characterized by **Lexical Coherence**
- Long standing tradition at ITC-irst
  - WordNet Domains (Magnini and Cavaglià, 2000)
  - Domain Driven Disambiguation (Magnini et al., 2001)
  - Domain Kernels (Gliozzo et al., 2005)

**WordNet Domains**

An extension of the MultiWordNet
150 Domain Labels from the DDC
**Lexical coherence**

- Most of the terms in a text belong to a specific domain.
- Allows us to induce Semantic Domains from texts.

<table>
<thead>
<tr>
<th>Word class</th>
<th>Nouns</th>
<th>Verbs</th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRD words</td>
<td>18732 (34.5%)</td>
<td>2416 (8.7%)</td>
<td>1992 (9.6%)</td>
<td>436 (3.7%)</td>
<td>21%</td>
</tr>
<tr>
<td>Polysemy</td>
<td>3.90</td>
<td>9.55</td>
<td>4.17</td>
<td>1.62</td>
<td>4.46</td>
</tr>
<tr>
<td>TUD words</td>
<td>13768 (25.3%)</td>
<td>2224 (8.1%)</td>
<td>815 (3.5%)</td>
<td>300 (2.5%)</td>
<td>11%</td>
</tr>
<tr>
<td>Polysemy</td>
<td>4.02</td>
<td>7.88</td>
<td>4.32</td>
<td>1.02</td>
<td>4.49</td>
</tr>
<tr>
<td>TUG words</td>
<td>21902 (40.2%)</td>
<td>22933 (83.2%)</td>
<td>17987 (86.7%)</td>
<td>11131 (53.8%)</td>
<td>64%</td>
</tr>
<tr>
<td>Polysemy</td>
<td>5.03</td>
<td>10.89</td>
<td>4.55</td>
<td>2.78</td>
<td>6.39</td>
</tr>
</tbody>
</table>

---

**Domain Restriction**

- Paradigmatic relations hold mainly among terms in the same domain.
Outline

- Semantic Domains and Linguistic Theory
  - Ambiguity and Variability
  - Semantic Fields
  - Semantic Domains and their properties
- Domain Models
  - Definition
  - Acquisition form Corpora
  - The Domain Kernel
  - Multilingual Domain Models
- Domain Driven Technologies
  - Text Categorization
  - Word Sense Disambiguation
  - Cross Language Text Categorization
  - Ontology Learning from Texts

Domain Model

- It represents the associations among the terms in the vocabulary and a set of Semantic Domains

<table>
<thead>
<tr>
<th>HIV</th>
<th>Medicine</th>
<th>Computer Science</th>
</tr>
</thead>
<tbody>
<tr>
<td></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>

- (domain-)Ambiguity: terms belong to many domains
- (domain-)Variability: Terms in the same domain refer to very closely related concepts
Corpus based acquisition of Domain Models

LSA for term clustering 1/2

Term Vectors are mapped into a lower dimensional Latent Semantic Space
**LSA for term clustering 2/2**

```
(('payment#n".0.29346246) ("property#n".0.28544194) ("pension#n".0.28260407) ("pay#v".0.2661988) ...)
(('vehicle#n".0.32334787) ("engine#n".0.3180726) ("car#n".0.3062577) ("hotel#n".0.29830772) ("driver#n".0.28349584) ("transport#n".0.27458063) ...)
(('"painting#n".0.24917021) ("artist#n".0.24005291) ("gallery#n".0.23767446) ("exhibition#n".0.23528978) ("art#n".0.21457596) ("arts#n".0.21291412) ...)
```

28/07/06

---

**The Domain Space**

Terms and texts are projected into a common Space

\[
D(\theta) = \frac{\sum_{\theta \in \theta} \sum_{i} P(w_i, t \theta)^2}{\sqrt{\sum_{\theta \in \theta} \sum_{i} P(w_i, t \theta)^2}}
\]

28/07/06
The Domain Kernel

\[ K_D(t_i, t_j) = \frac{\langle D(t_i), D(t_j) \rangle}{\sqrt{\langle D(t_j), D(t_j) \rangle \langle D(t_i), D(t_i) \rangle}} \]

- A: he is affected by AIDS
- B: HIV is a virus
- C: the laptop has been infected by a virus
- \( \text{sim-bow}(A, B) < \text{sim-bow}(B, C) \) (?)
- \( \text{sim-DOM}(A, B) > \text{sim-DOM}(B, C) \) (!)

Multilingual Domain Space

Texts in different languages are mapped in a common Multilingual Domain Space (Gliozzo and Strapparava, 2005) by the Multilingual Domain Kernel.
Multilingual Domain Kernel

- Explicit feature mapping from the VSMs to the Domain Space

\[ D(t_j) = t_j(I^{\text{IDF}}D) = t_j^D \]

\[ D = \]

- \( D \) is acquired from comparable corpora and bilingual dictionaries

Acquisition from bilingual dictionaries
**Acquisition from Comparable Corpora**

- Proper nouns are often expressed in different languages by using the same terms (e.g. Microsoft, IBM)
- Comparable Corpora: Common lemmata 14% of which 97% are nouns

<table>
<thead>
<tr>
<th></th>
<th>English documents</th>
<th>Italian documents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( d_0 ) ( \ldots ) ( d_{n-1} ) ( d_n )</td>
<td>( d_0 ) ( \ldots ) ( d_{n-1} ) ( d_n )</td>
</tr>
<tr>
<td><strong>English</strong></td>
<td>( w_1^e ) 0 1 ( \ldots ) 0 1</td>
<td>( w_1^i ) 0 0 ( \ldots ) 0 1</td>
</tr>
<tr>
<td><strong>Lexicon</strong></td>
<td>1 1 ( \ldots ) 1 0</td>
<td>0 0 ( \ldots ) 0 1</td>
</tr>
<tr>
<td></td>
<td>( w_{n-2}^e ) ( \ldots ) ( w_n^e )</td>
<td>( w_{n-2}^i ) ( \ldots ) ( w_n^i )</td>
</tr>
<tr>
<td></td>
<td>( w_1^c ) 0 1 ( \ldots ) 0 0</td>
<td>( w_1^c ) 0 0 ( \ldots ) 0 0</td>
</tr>
<tr>
<td></td>
<td>0 1 ( \ldots ) 0 0</td>
<td>0 0 ( \ldots ) 0 0</td>
</tr>
</tbody>
</table>

**Lexical Overlapping**

**Outline**

- Semantic Domains and Linguistic Theory
  - Ambiguity and Variability
  - Semantic Fields
  - Semantic Domains and their properties
- Domain Models
  - Definition
  - Acquisition form Corpora
  - The Domain Space
  - Multilingual Domain Models
- Domain Driven Technologies
  - Text Categorization
  - Word Sense Disambiguation
  - Cross Language Text Categorization
  - Ontology Learning from Texts
Support Vector Machines

Support Vector Machines

find optimal separating hyperplane

Find support vectors

Similarity with support vectors is evaluated by the kernel function

\[ f(x) = \sum_{i=1}^{T} \lambda_i F(x_i) \cdot F(x) + \lambda_0 = \sum_{i=1}^{T} \lambda_i K(x_i, x) + \lambda_0 \]

external knowledge can be "plugged"

The Kernel Function is the only domain specific component

Kernel Functions

"External" Knowledge can be used to define the kernel function (semi-supervised learning)

Different kernels can be composed (e.g. if \( K_1 \) and \( K_2 \) are kernels any linear combination \( K \) of them is still a kernel)

\[ K_\sigma(x_i, x_j) = \sum_{i=1}^{n} \frac{K_1(x_i, x_j)}{\sqrt{K_1(x_i, x_j)K_1(x_i, x_i)}} \]
**Text Categorization evaluation**

### Table 3: Number of training examples needed by $K_D$ and $K_{BAV}$ to reach the same micro-F1 on the Reuters task

<table>
<thead>
<tr>
<th>F1</th>
<th>Domain Kernel</th>
<th>Bow Kernel</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>14</td>
<td>267</td>
<td>5%</td>
</tr>
<tr>
<td>84</td>
<td>146</td>
<td>1380</td>
<td>10%</td>
</tr>
<tr>
<td>90</td>
<td>668</td>
<td>6680</td>
<td>10%</td>
</tr>
</tbody>
</table>

### Table 4: Number of training examples needed by $K_D$ and $K_{BAV}$ to reach the same micro-F1 on the 20Newsgroups task

<table>
<thead>
<tr>
<th>F1</th>
<th>Domain Kernel</th>
<th>Bow Kernel</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>30</td>
<td>500</td>
<td>6%</td>
</tr>
<tr>
<td>70</td>
<td>98</td>
<td>1182</td>
<td>8%</td>
</tr>
<tr>
<td>85</td>
<td>2272</td>
<td>7879</td>
<td>29%</td>
</tr>
</tbody>
</table>

---

**Intensional Learning**

- Categories are described by means of discriminative words instead of labeled texts. E.g. [CHEMISTRY] chemistry, atom, ...
- Bootstrap: (i) retrieve a preliminary set of labeled documents, (ii) train a supervised classifier on them.
- DM are used in the first step (term-text similarity).
Discussion: SVD, compression and learning theory

\[ m > \frac{1}{\epsilon} \left( 4 \log_2 \frac{2}{\delta} + 8V(P) \log_2 \frac{13}{\epsilon} \right) \]

Domain Driven Disambiguation

Choose the sense whose domain maximize the similarity with the domain of the context

- The laptop has been infected by a virus [CS]
- HIV is a virus [BIO]
- High precision for “domain words”
- No Syntagmatic Distinctions
- Domain Grained Disambiguation

The laptop has been infected by a virus [CS]
HIV is a virus [BIO]
High precision for “domain words”
No Syntagmatic Distinctions
Domain Grained Disambiguation
Kernel Methods for WSD [ACL-05]

- Modeling independently
  - Syntagmatic relations (String Kernel)
  - Domain relations (Domain Kernel)
- Semi-supervised approach: Domain Models are acquired from unlabeled data

<table>
<thead>
<tr>
<th></th>
<th>MF</th>
<th>Agreement</th>
<th>BEST</th>
<th>$K_{std}$</th>
<th>$K_{std}$</th>
<th>DM+</th>
<th>BEST+</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>53.2</td>
<td>67.3</td>
<td>72.9</td>
<td>69.7</td>
<td>73.3</td>
<td>3.6</td>
<td>0.4</td>
</tr>
<tr>
<td>Catalan</td>
<td>66.3</td>
<td>93.1</td>
<td>85.2</td>
<td>83.2</td>
<td>89.0</td>
<td>3.8</td>
<td>3.8</td>
</tr>
<tr>
<td>Italian</td>
<td>18.0</td>
<td>89.0</td>
<td>53.1</td>
<td>53.1</td>
<td>61.3</td>
<td>8.2</td>
<td>8.2</td>
</tr>
<tr>
<td>Spanish</td>
<td>67.7</td>
<td>85.3</td>
<td>84.2</td>
<td>84.2</td>
<td>88.2</td>
<td>4.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Table 4: Comparative evaluation on the lexical sample tasks. Columns report: the Most Frequent baseline, the Inter annotator agreement, the F1 of the best system at Senseval-3, the F1 of $K_{std}$, the F1 of $K_{std}^+$, DM+ (the improvement due to DM, i.e. $K_{std}^+ - K_{std}$), and BEST+ (the improvement on the state-of-the-art, i.e. $K_{std}^+ -$ BEST).

Cross Language Text Categorization

- ADNKRONOS news agencies in two languages (English and Italian)
- Common set of categories is used
- Comparable Corpora: news of the same period of time but no alignment is given

<table>
<thead>
<tr>
<th>Categories</th>
<th>English</th>
<th>Italian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Test</td>
</tr>
<tr>
<td>Quality_of_Life</td>
<td>5759</td>
<td>1989</td>
</tr>
<tr>
<td>Made_in_Italy</td>
<td>5711</td>
<td>1864</td>
</tr>
<tr>
<td>Tourism</td>
<td>5731</td>
<td>1857</td>
</tr>
<tr>
<td>Culture_and_School</td>
<td>3665</td>
<td>1245</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>20866</td>
<td>6955</td>
</tr>
</tbody>
</table>
Cross Language Text Categorization

- Dictionary
- Monosemous
- Comparable

GOD: General Ontology Discovery

- Search as an Ontology Learning Process

Paradigmatic relations hold mainly among words in the same domain
Terminology Extraction

Core Ontology

<table>
<thead>
<tr>
<th>WSD: Conceptual Density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Table 1: WSD Results for the Domain Music

Table 2: Evaluation for the Ontology Population capability of the generated Core Ontologies
Relation Extraction

- Candidate ISA relations identified by syntactic patterns
- Espresso (ACL 2006)
- Domain similarity as a ranking criterion
- Chemistry Domain
- MAP 0.85 vs 0.73

Conclusion

- Semantic Domains shows many interesting properties
  - Lexical coherence
  - Domain ambiguity
  - Domain Restriction
  - Multilinguality
- Domain Models can be automatically acquired from corpora
- Semi supervised learning in the Domain Space
- Improved state of the art in
  - Text Categorization (better learning curve)
  - WSD (best Senseval-3 system)
- CLTC (close to the monolingual settings)
Future Work

- New evaluation tasks
  - Topic detection
  - Summarization
- GOD as a search engine
  - Domain Models from WEB scale corpora
  - Domain Driven Relation Extraction
- Unsupervised learning of morphology and syntax
- Multimedia: AV retrieval