Computational approaches to argument structure

- The automatic acquisition of lexical information from corpora is a longstanding research avenue in computational linguistics
  - subcategorization frames (Korhonen 2002, Schulte im Walde 2009, etc.)
  - selectional preferences (Resnik 1993, Light & Greiff 2002, Erk et al. 2010, etc.)
  - verb classes (Merlo & Stevenson 2001, Schulte im Walde 2006, Kipper-Schuler et al. 2008, etc.)
- Corpus-based information has been used to build lexical resources
  - cf. VALEX for English (Korhonen et al. 2006), LexSchem for French (Messiant et al. 2008), etc.

LexIt: a computational lexical resource for Italian

**LexIt** is a computational framework for the automatic acquisition and exploration of corpus-based distributional profiles of Italian verbs, nouns, and adjectives

- **LexIt** is publicly available through a web interface:
  - [http://sesia.humnet.unipi.it/lexit/](http://sesia.humnet.unipi.it/lexit/)
- **LexIt** is the first large-scale resource of such type for Italian, aiming at characterizing the valence properties of predicates fully on distributional ground.
The LexIt Distributional Profiles

The distributional profile for a word \( w \) is an array of statistical information extracted from a corpus to characterize the distributional behavior of \( w \).

The LexIt distributional profiles include:

- **syntactic profiles**, specifying the syntactic slots (subject, complements, modifiers, etc.) and syntactic frames with which predicates co-occur.
- **semantic profiles**, composed by:
  - the lexical sets with the most prototypical fillers realizing the syntactic slots;
  - the semantic classes characterizing the selectional preferences of syntactic slots.

Association Measures

“A simple association measure interprets co-occurrence frequency \( O \) by comparison with the expected frequency \( E \), and calculates and association score as a quantitative measure for the attraction between two words” (Evert, 2008:18)

**Local Mutual Information (Evert, 2008)**

\[
LMI = O \times \log_2 \frac{O}{E}
\]

Key properties of LMI:

- Downgrades the risk of overestimating the significance of low frequency events.
- Is a two-sided measure: quantifies both attraction and repulsion.

The LexIt framework

- **LexIt** is an open and parametrizable framework.
  - Source corpora.
  - Part of speech to be profiled.
  - Definition of subcategorization frames.
  - Statistical indexes.
  - Semantic classes for selectional preferences, etc.

Today we focus on the acquisition of distributional profiles for Italian verbs.
Introducing LexIt

Building Distributional Profiles

Ongoing Work

Conclusions

Pre-processing

Subcategorization Frames

Lexical sets

Selectional preferences

Building distributional profiles

- Pre-processing: linguistic analysis with automatic tools
- Extraction of subcategorization frames from parsed text
- Assignment of lexical sets to argument slots
- Selectional preferences: from lexical sets to semantic classes

Pre-processing

- Tokenization, Lemmatization, Part-of-speech tagging
  - TANL (Text Analytics and Natural Language), a suite of modules for Italian Natural Language Processing developed by the University of Pisa and ILC-CNR
- Dependency Parsing
  - DeSR, a stochastic dependency parser (Attardi & Dell’Orletta 2009)
  - dependency trees are constructed without relying on any subcategorization lexicon

Subcategorization frames

Subcategorization Frame (SCF):

- represents a pattern of syntactic dependencies headed by the target lemma
- is formed by an unordered set of slots, representing argument positions (i.e., subject, object, etc.)
- is identified by a synthetic label

- Verb SCFs also include:
  - the zero argument construction
    - *Gianni è arrivato* “John arrived” ⇒ subj#0
  - the reflexive pronoun *si*
    - *Il vaso si è rotto* “The vase si-broke” ⇒ subj#si#0
Introducing LexIt

Building Distributional Profiles

Ongoing Work

Conclusions

Pre-processing

Subcategorization Frames

Lexical sets

Selectional preferences

Subcategorization frames

- No formal distinction is made between arguments and adjuncts
  - *abitare al mare* (“to live at the sea”) ⇒ subj#comp-a
  - *mangiare al mare* (“to eat at the sea”) ⇒ subj#comp-a
  - Information between argument-adjuncts is not explicitly encoded in the parser
  - Arguments and adjuncts are notoriously hard to discriminate
- For each frame, the LexIt profiles also specify the most prototypical:
  - Verbal modifiers
    - *entrare correndo* (“to run into”)
  - Adverbial modifiers
    - *correre velocemente* (“to run fast”)

Extracting subcategorization frames

- 104 SCFs were selected among the most frequent syntactic dependency combinations in the parsed corpus
- The joint frequency between each verb and the SCFs was computed from the verb dependency patterns automatically extracted from the parsed corpus
- The statistical salience of each SCFs with the target word was estimated with LMI
Building distributional profiles

- Pre-processing: linguistic analysis with automatic tools ✓
- Extraction of subcategorization frames from parsed text ✓
- Assignment of lexical sets to argument slots
- Selectional preferences: from lexical sets to semantic classes

### Lexical sets

**Example:** leggere ("to read"), SCF: subj#obj, slot: obj

<table>
<thead>
<tr>
<th>Filler</th>
<th>Frequency</th>
<th>LMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>libro (&quot;book&quot;)</td>
<td>1617</td>
<td>9907</td>
</tr>
<tr>
<td>giornale (&quot;magazine&quot;)</td>
<td>1511</td>
<td>9939</td>
</tr>
<tr>
<td>testo (&quot;text&quot;)</td>
<td>567</td>
<td>2951</td>
</tr>
<tr>
<td>articolo (&quot;article&quot;)</td>
<td>435</td>
<td>2172</td>
</tr>
<tr>
<td>lettera (&quot;letter&quot;)</td>
<td>476</td>
<td>2157</td>
</tr>
<tr>
<td>dichiarazione (&quot;declaration&quot;)</td>
<td>432</td>
<td>2013</td>
</tr>
<tr>
<td>romanzo (&quot;novel&quot;)</td>
<td>303</td>
<td>1661</td>
</tr>
<tr>
<td>sceneggiatura (&quot;plot&quot;)</td>
<td>236</td>
<td>1601</td>
</tr>
<tr>
<td>pagina (&quot;page&quot;)</td>
<td>338</td>
<td>1588</td>
</tr>
<tr>
<td>comunicato (&quot;announcement&quot;)</td>
<td>237</td>
<td>1053</td>
</tr>
</tbody>
</table>

**Lexical set (Hanks 1996; Hanks and Pustejovsky 2005)**

The set of the words that typically occur with a target verb in a given syntactic position, ranked by their degree of prototypicality

- For each slot in a SCF, the slot-filler association strength was computed with LMI
- The slot lexical set is formed by the lexical fillers with LMI > 0

<table>
<thead>
<tr>
<th>Filler</th>
<th>Frequency</th>
<th>LMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>decisione (&quot;decision&quot;)</td>
<td>126</td>
<td>719</td>
</tr>
<tr>
<td>notizia (&quot;news&quot;)</td>
<td>90</td>
<td>505</td>
</tr>
<tr>
<td>intenzione (&quot;intention&quot;)</td>
<td>34</td>
<td>211</td>
</tr>
<tr>
<td>nome (&quot;name&quot;)</td>
<td>28</td>
<td>97</td>
</tr>
<tr>
<td>variazione (&quot;variation&quot;)</td>
<td>11</td>
<td>68</td>
</tr>
<tr>
<td>esito (&quot;outcome&quot;)</td>
<td>13</td>
<td>66</td>
</tr>
<tr>
<td>disponibilità (&quot;availability&quot;)</td>
<td>13</td>
<td>64</td>
</tr>
<tr>
<td>esistenza (&quot;existence&quot;)</td>
<td>12</td>
<td>54</td>
</tr>
<tr>
<td>risultato (&quot;result&quot;)</td>
<td>18</td>
<td>53</td>
</tr>
<tr>
<td>informazione (&quot;information&quot;)</td>
<td>13</td>
<td>52</td>
</tr>
</tbody>
</table>
Introducing LexIt
Building Distributional Profiles
Ongoing Work
Conclusions

Lexical sets

Example: correre ("to run"), SCF: subj#0, slot: adverbial modifier

<table>
<thead>
<tr>
<th>Filler</th>
<th>Frequency</th>
<th>LMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>troppo</td>
<td>181</td>
<td>1470</td>
</tr>
<tr>
<td>molto</td>
<td>92</td>
<td>546</td>
</tr>
<tr>
<td>dietro</td>
<td>53</td>
<td>360</td>
</tr>
<tr>
<td>via</td>
<td>62</td>
<td>354</td>
</tr>
<tr>
<td>tanto</td>
<td>57</td>
<td>347</td>
</tr>
<tr>
<td>avanti</td>
<td>52</td>
<td>258</td>
</tr>
<tr>
<td>sempre</td>
<td>60</td>
<td>257</td>
</tr>
<tr>
<td>insieme</td>
<td>47</td>
<td>225</td>
</tr>
<tr>
<td>bene</td>
<td>46</td>
<td>169</td>
</tr>
<tr>
<td>velocemente</td>
<td>20</td>
<td>155</td>
</tr>
</tbody>
</table>

Building distributional profiles

- Pre-processing: linguistic analysis with automatic tools
- Extraction of subcategorization frames from parsed text
- Assignment of lexical sets to argument slots
- Selectional preferences: from lexical sets to semantic classes

Selectional preferences

Selectional preferences for a (noun-selecting) slot $s$

A ranked list of the noun semantic classes (e.g., PERSON, ANIMAL, etc.) that best describe the semantic types of the fillers of $s$, i.e., the semantic constraints of $s$

Semantic classes in LexIt

- ANIMAL, ARTIFACT, ACT, ATTRIBUTE, FOOD, COMMUNICATION, KNOWLEDGE, BODY PART, EVENT, NATURAL PHENOMENON, SHAPE, GROUP, LOCATION, MOTIVATION, NATURAL OBJECT, PERSON, PLANT, POSSESSION, PROCESS, QUANTITY, FEELING, SUBSTANCE, STATE, TIME

LexIt and WordNet

- The LexIt classes are the 24 top-nodes of the Italian section of MultiWordNet (Pianta et al. 2002), a large scale multilingual lexicon based on Princeton’s WordNet (Fellbaum 1998)
  - word senses are represented by synsets (i.e., sets of synonyms)
  - synsets are arranged in a semantic hierarchy

Two points to keep in mind:

- semantically ambiguous words belong to more than one synset
- the top-nodes we selected are mutually exclusive: no subtyping relations hold among the LexIt semantic classes
Introducing LexIt

Building Distributional Profiles

Ongoing Work

Conclusions

Pre-processing

Subcategorization Frames

Lexical sets

Selectional preferences

Extracting selectional preferences

The selectional preferences of a slot are obtained through an 
inductive generalization from the slot lexical sets:

- the slot-filler joint frequency was uniformly divided among the 
different senses assigned to the filler in MultiWordNet
- the slot-class joint frequency was obtained by propagating the 
sense frequency up to the 24 top-nodes
- the LMI association score between the slot and each semantic 
class was computed using the slot-class joint frequency
- the semantic classes with LMI > 0 were selected to represent 
the selectional preferences of the slot

Gabriella Lapesa, Alessandro Lenci

Modeling Subcategorization Through Co-occurrence 26/38

Selecting preferences

Example: leggere (“to read”). SCF: subj#obj, slot: obj

<table>
<thead>
<tr>
<th>Semantic Class</th>
<th>Association Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>16452</td>
</tr>
<tr>
<td>Artifact</td>
<td>2151</td>
</tr>
<tr>
<td>Substance</td>
<td>149</td>
</tr>
<tr>
<td>Time</td>
<td>12</td>
</tr>
</tbody>
</table>

Lexical set


Gabriella Lapesa, Alessandro Lenci

Modeling Subcategorization Through Co-occurrence 27/38

Distributional semantics

Meaning and distribution

The analysis of a relevant number of contexts of a word sheds light 
on key aspects of its meaning (cf. Harris 1954, Firth 1957, Cruse 
1986, Miller & Charles 1991, etc.)

Distributional semantic profiles have both a descriptive and a 
predictive function:

- lexical sets provide a “snapshot” of the most typical fillers of a 
  verb in a certain syntactic position
- selectional preferences generalize from these instances to more 
  abstract semantic properties of the verb arguments, thereby 
  making predictions about previously unseen slot fillers

Gabriella Lapesa, Alessandro Lenci

Modeling Subcategorization Through Co-occurrence 28/38
The current status of LexIt

- **LexIt corpora**
  - *La Repubblica* (ca. 331 millions tokens of newspaper articles)
  - *Wikipedia.it* (ca. 152 millions tokens)
- Distributional profiles for verbs and nouns
  - *La Repubblica*: 3,873 most frequent verbs, and 12,766 most frequent nouns (min. freq. = 100)
  - *Wikipedia.it*: 2,831 most frequent verbs, and 11,056 most frequent nouns (min. freq. = 100)
- Distributional profiles for adjectives are coming soon!

Ongoing Work

- Comparison of syntactic profiles contained in LexIt with a manually developed valence lexicon: *Wörterbuch der Italianischen Verben* (Blumenthal & Rovere 1998)
- Qualitative analysis of the syntactic profiles, to identify the frames wrongly associated to the target verbs

Logical polysemy (Pustejovsky 1995): “the ability of some words to appear in contexts that are contradictory in type specifications”

Relying on the information concerning selectional preferences over single classes we applied association measures to construct corpus-based “polysemic semantic types” possibly associated to frame slots

- e.g., how many words occurring in the direct object position of the verb to read are assigned by MultiWordNet to both Artifact and Communication?
Conclusions

- **LexIt** contains distributional information of Italian words automatically extracted from corpora
  - it is not "noise-free", due to the current limits of computational linguistics tools (e.g., part-of-speech tagging and parsing errors)
- Possible applications
  - induction of distributional verb classes
  - "usage-based" models of the syntax-semantics interface
  - acquisition of frequency data about subcategorization frames for psycholinguistic research

References

References