Modeling Subcategorization Through Co-occurrence

A Computational Lexical Resource for Italian Verbs

Gabriella Lapesa[1], Alessandro Lenci[2]

[1]University of Osnabrück, Institute of Cognitive Science
[2]University of Pisa, Department of Linguistics

Explorations in Syntactic Government and Subcategorisation
University of Cambridge, 2nd September 2011

Outline

1. Introducing LexIt
   - The Project
   - Distributional Profiles
2. Building Distributional Profiles
   - Pre-processing
   - Subcategorization Frames
   - Lexical sets
   - Selectional preferences
3. Ongoing Work
4. Conclusions

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 (Korohnen *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/
- *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

Modeling subcategorization through co-occurrence

- Distributional profiles in *LexIt* are automatically extracted from large corpora with computational linguistics tools (without any manual revision)
- The *LexIt* profiles contain statistical indexes to identify the most salient and prototypical distributional features of predicates:
  - co-occurrence frequency
  - association measures
- Corpus-derived statistics are used to model the association between verbs and syntactic constructions, lexical fillers and semantic classes as a gradient preference instead of categorical selection

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)

$$LMI = O \times log_2 \frac{O}{E} \tag{1}$$

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

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

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

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**

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")

Subcategorization frames

Frame: SUBJ#OBJ#COMP-A
Target: *dare* ("to give"), freq 336731; Frame-verb: freq 107388, LMI 327656

- *Gianni ha dato il libro a Maria* "Gianni gave the book to Mary"
- *Gianni ha dato a Maria il libro* "Gianni gave Mary the book"
- *Gianni ha generosamente dato a Maria il libro* "Gianni gave Mary the book generously"
- *(Lui) ha dato il libro a sua madre piangendo* "(He) gave the book to his mother crying"

Subcategorization frames

SUBJ#SI#0

- Target: *rompere* ("to break") freq 52537; Frame-verb: freq 1980, LMI 3293.
  - *Il vetro si è rotto* "The glass broke"
  - *Il vetro si rompe facilmente* "Glass breakes easily"
- Target: *fermare* ("to stop") freq 52537; Frame-verb : freq 10967, LMI 28864.
  - *La macchina si è fermata frenando* "The car stopped braking"

Extracting subcategorization frames

1. 104 SCFs were selected among the most frequent syntactic dependency combinations in the parsed corpus
2. The joint frequency between each verb and the SCFs was computed from the verb dependency patterns automatically extracted from the parsed corpus
3. 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

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

Lexical sets

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

Filler	Frequency	LMI
libro (“book”)	1617	9907
giornale (“magazine”)	1511	9939
testo (“text”)	567	2951
articolo (“article”)	435	2172
lettera (“letter”)	476	2157
dichiarazione (“declaration”)	432	2013
romanzo (“novel”)	303	1661
sceneggiatura (“plot”)	236	1601
pagina (“page”)	338	1588
comunicato (“announcement”)	237	1053

Lexical sets

Example: *comunicare* (“to communicate”), SCF: subj#obj#comp-a, slot: obj

Filler	Frequency	LMI
decisione (“decision”)	126	719
notizia (“news”)	90	505
intenzione (“intention”)	34	211
nome (“name”)	28	97
variazione (“variation”)	11	68
esito (“outcome”)	13	66
disponibilità (“availability”)	13	64
esistenza (“existence”)	12	54
risultato (“esult”)	18	53
informazione (“information”)	13	52

Lexical sets

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

Filler	Frequency	LMI
troppo ("too much")	181	1470
molto ("a lot")	92	546
dietro ("behind")	53	360
via ("away")	62	354
tanto ("a lot")	57	347
avanti ("forward")	52	258
sempre ("always")	60	257
insieme ("together")	47	225
bene ("well")	46	169
velocemente ("quickly")	20	155

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

Extracting selectional preferences

- The selectional preferences of a slot are obtained through an inductive generalization from the slot lexical sets:
  1. the slot-filler joint frequency was uniformly divided among the different senses assigned to the filler in MultiWordNet
  2. the slot-class joint frequency was obtained by propagating the sense frequency up to the 24 top-nodes
  3. the LMI association score between the slot and each semantic class was computed using the slot-class joint frequency
  4. the semantic classes with LMI $>$ 0 were selected to represent the selectional preferences of the slot

Selectional preferences

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

Semantic Class	Association Strength
Communication	16452
Artifact	2151
Substance	149
Time	12

Lexical set

libro (“book”), *giornale* (“magazine”), *testo* (“text”), *articolo* (“article”), *lettera* (“letter”), *dichiarazione* (“declaration”), *romanzo* (“novel”), *sceneggiatura* (“plot”), *pagina* (“page”), *comunicato* (“announcement”).

Selectional preferences

Example: *comunicare* (“to communicate”), SCF: subj#obj#comp-*a*, slot: obj

Semantic Class	Association Strength
Knowledge	187
Act	110
Feeling	93
Attribute	71
Communication	65
State	57

Lexical Set

decisione (*decision*), notizia (*news*), intenzione (*intention*), nome (*name*), variazione (*variation*), esito (*outcome*), disponibilità (*availability*), esistenza (*existence*), risultato (*result*).

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

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

Evaluation of the SCF module

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

Ongoing Work

Argument Polysemy

- 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?

The multiple facets of *LexIt*

- **A valence lexicon** *Combinatory dictionary of Contemporary Spanish* (Bosque, 2004), *Wörterbuch der Italianischen Verben* (Blumenthal and Rovere, 1998)
- **A dictionary of collocations** *Oxford Collocation Dictionary* (Deuter and Venning, 2002)
- **A corpus-based electronic dictionary** *Collins Cobuild English Dictionary* (Sinclair, 1996)

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

- Attardi, Giuseppe & Felice Dell'Orletta. 2009. "Reverse Revision and Linear Tree Combination for Dependency Parsing". *Proceedings of NAACL-HLT 2009*, Boulder, Col.261-264.
- Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston & Marco Mazzoleni. 2004. "Introducing the *La Repubblica* Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian". *Proceedings of LREC 2004*. Lisboa. 1771- 1774.
- Blumenthal, Peter & Giovanni Rovere. 1998. *Wörterbuch der italienischen Verben*. Ernest Klettverlag, Stuttgart.
- Bosque, Ignacio. 2004. *REDES: diccionario combinatorio del español contemporáneo*. SM Ediciones, Madrid.
- Cruse, David Alan. 1986. *Lexical semantics*. Cambridge University Press.

References

- Deuter, Margaret & Harry Venning. 2002. *Oxford Collocations dictionary for students of English*. Oxford University Press.
- Erk, Katrin, Sebastian Padó & Ulrike Padó . 2010. "Corpus-driven Model of Regular and Inverse Selectional Preferences". *Computational Linguistics* 36(4), 723–763.
- Evert, Stefan. 2008. "Corpora and Collocations". In *Corpus Linguistics. An International Handbook* ed. by Anke Lüdeling & Merja Kytö, 1212-1248. Berlin: Mouton de Gruyter.
- Fellbaum, Christiane, ed. 1998. *WordNet An Electronic Lexical Database*. Cambridge, Mass: MIT Press.
- Firth, John Rupert. 1957. "A Synopsis of Linguistic Theory, 1930-1955". in J.R. Firth et al. *Studies in Linguistic Analysis*. Special volume of the Philological Society. Oxford: Blackwell.
- Hanks, Patrick. 1996. "Contextual Dependency and Lexical Sets". *International Journal of Corpus Linguistics* 1:1.75-98.

References

- Hanks, Patrick & James Pustejovsky. 2005. "A pattern dictionary for natural language processing". *Revue Française de linguistique appliquée*. 63-82.
- Harris, Zellig S. 1954. "Distributional Structure". *Word*, 10:2-3.146-62 [reprinted in Harris, Zellig S., 1970. *Papers in Structural and Transformational Linguistics*, Dordrecht: Reidel.775-794].
- Kipper-Schuler, Karin, Anna Korhonen, Neville Ryant & Martha Palmer. 2008. "A Large-Scale Classification of English Verbs". *Journal of Language Resources and Evaluation* 42:1.21-40.
- Korhonen, Anna. 2009. "Automatic Lexical Classification - Balancing between Machine Learning and Linguistics". *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, Hong Kong.
- Korhonen, Anna, Yuval Krymolowski & Ted Briscoe. 2006. "A Large Subcategorization Lexicon for Natural Language Processing Applications" *Proceedings of the Language Resources and Evaluation Conference* (LREC), Genova.

References

- Light, Mark & Warren Greiff. 2002. “Statistical Models for the Induction and Use of Selectional Preferences”. *Cognitive Science*: 26.269281.
- Merlo, Paola & Suzanne Stevenson. 2001. “Automatic Verb Classification Based on Statistical Distributions of Argument Structure”. *Computational Linguistics* 27:3.373-408.
- Messiant, Cedric, Anna Korhonen & Thierry Poibeau. 2008. “LexSchem: A Large Subcategorization Lexicon for French Verbs”. *Proceedings of the Language Resources and Evaluation Conference* (LREC), Marrakech.
- Miller, George A. & Walther Charles. 1991. “Contextual Correlates of Semantic Similarity”. *Language and Cognitive Processes* 6.1-28.
- Pianta, Emanuele, Luisa Bentivogli & Christian Girardi. 2002. “MultiWordNet: Developing an Aligned Multilingual Database”. *Proceedings of the 1st Global WordNet Conference.* Mysore.
- Pustejovsky, James. 1995. *The Generative Lexicon*, Cambridge, Mass.: MIT Press.

References

- Resnik, Philip. 1993. *Selection and information: a class-based approach to lexical relationships.* Phd dissertation, University of Pennsylvania.
- Schulte im Walde, Sabine. 2006. “Experiments on the Automatic Induction of German Semantic Verb Classes”. *Computational Linguistics* 32:2.159-194.
- Schulte im Walde, Sabine. 2009. “The Induction of Verb Frames and Verb Classes from Corpora”. *Corpus Linguistics. An International Handbook* ed. by Anke Lüdeling & Merja Kytö, 952-972. Berlin: Mouton de Gruyter.
- Sinclair, John. 1996. *Collins Cobuild English Dictionary*. Harper Collins Publishers, London.