Sweetening Ontologies cont'd

Elisabetta Jezek
Università di Pavia

JSSP 2013
Fondazione Bruno Kessler
Nov 20-22, 2013

Outline of the talk

- Background.
- Goal of the work: ontology alignment.
- Introduce the experiment.
- Preliminary results.
- Ongoing and Future.

Background

- Repository of typed predicate-argument structures (T-PAS) for Italian.
- Under development at the Dept. of Humanities of the University of Pavia, in collaboration with the Human Language Technology group of Fondazione Bruno Kessler (FBK), Trento.
- Technical support of the Faculty of Informatics at Masaryk University in Brno (CZ).
- It currently consists of 755 analyzed "average polysemy" verbs (including pronominal forms) (dd. Nov 18, 2013) and about 3000 patterns.
- Manually annotated resource.
- Linguistic research and NLP applications (details in Jezek 2012).

What counts as a typed predicate-argument structure? 1/2

A typed predicate-argument structure (T-PAS) is a corpus-derived argument structure with the specification of the expected semantic type for each argument position, populated by lexical sets (Hanks 1986), i.e. the statistically relevant list of collocates that typically fill each position.

[[Persona]-subj] partecipa [[Evento]-iobj_a]

- Lexical set [[Event]] = {gara, riunione, selezione, manifestazione, seduta, cerimonia, conferenza, votazione, elezione, celebrazione, esequia, competizione, maratona, discussione, messa, festa, marcia, fiaccolata, trattativa, missione, commemorazione, incontro, concorso, convegno, raduno, iniziativa, stage, evento, seminario, torneo, attività, corso, asta, dibattito, progetto, festival...}

Resource architecture

- The resource consists of three components:
- A repository of T-PAS linked to verb senses expressed in the form of implicatures.
- A "shallow" list of semantic type labels (HUMAN, ARTEFACT, EVENT, ecc.).
- A corpus of sentences that represent instantiations of T-PAS.

Corpus Pattern Analysis procedure (CPA, Hanks 2004)

- Choose a target verb and create a sample concordance of 250 actual uses.
- Identify the relevant structure (typical syntagmatic patterns).
- Associate a typing constraint to each argument position in the pattern.
- Assign the instances of the sample to one of the patterns.
- Store the pattern (with the associated corpus instances) in the resource.
- Associate each pattern with at least one sense, expressed in the form of an implicature linked to the typing constrains specified in the pattern.
- [[Human]-subj] essere presente a [[Event]-iobj_a].

Type mismatches

- The paradigmatic sets of words that populate specific argument slots within the same verb sense do not map neatly onto the "expected" type (selected by V) (Pustejovsky and Jezek 2008).
- Mismatches between "pattern" type (expected by V) and "instance" type (inherent in N) within the same grammatical relation.

Aspectual verbs

[[Human]-subj] interrompe [[Event]-obj]

- Arriva Mirko e interrompe la conversazione. 'Mirko arrives and interrupts the conversation' (matching)
- Il presidente interrompe l'oratore. 'The president interrupts the speaker' (Human as Event)

Communication verbs

[[Human]-subj] annuncia [[Event]-obj]

- Lo speaker annuncia la partenza. 'The speaker announces the departure' (matching)
- Il maggiordomo annuncia gli invitati. 'The butler announces the guests' (Human as Event)
- L'altoparlante annunciava l'arrivo del treno. 'The loudspeaker announces the arrival of the train' (Artifact as Human)
- Una telefonata anonima avvisa la polizia. 'An anonymous telephone call alerted the police' (Event as Human)

Directed motion verbs

[[Human]-subj] raggiunge [[Location]-obj]

- Abbiamo raggiunto l'isola alle 5. 'We reached the island at 5' (matching)
- Ho raggiunto il semaforo e ho svoltato a destra. 'I reached the traffic light and turned right' (Artifact as Location)

Shimmering Lexical Sets

- Lexical sets populating a node in the ontology (i.e. a semantic type) tend to "shimmer" (Jezek and Hanks 2010) – that is, the membership of the lexical set tends to vary when one moves from verb to verb: some words drop out while other come in, just as predicated by Wittgenstein (*family resemblances*).
- Different verbs select different prototypical members of a semantic type even if the rest of the set remains the same.

Shimmering Lexical Sets

lavare [[Body Part]-obj]

- Lexical set [[Body Part]] = {denti, mano, piede, viso, faccia, schiena, testa, orecchio, volto ...}

amputare [[Body Part]-obj]

- Lexical set [[Body Part]] = {arto, gamba, braccio, dito, orecchio, mano, piede ...}

CPA type list

- By applying the CPA procedure to the analysis of concordances for ca 1500 English, Italian and Spanish verbs we compiled a list of about 230 semantic types obtained from manual clustering and generalization over sets of lexical items found in the argument positions in the corpus.

Ontological categories vs. linguistic classes

- These types look very much like conceptual / ontological categories for nouns but should instead be conceived as semantic classes, as they are induced by the analysis of selectional properties of verbs.
- They are language-driven, and reflect how we talk about entities in the world.
- Despite the obvious correlations, they differ from categories of entities defined on the basis of ontological axioms, such as those of DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering, cf. Masolo, Borgo, Gangemi, Guarino, Oltramari 2003).

Research questions/goals

- How do semantic classes obtained through pattern-based corpus analysis differ from categories which are defined on the basis of axiomatization?
- How do we organize the list into a structure for purposes of NLP applications?

Experiment

- Aligning the type inventory to the categories of DOLCE.
- Enhance the taxonomic structuring of CPA list using the OntoClean methodology (Guarino and Welty, 2002, 2009) which was exploited to built DOLCE.

Why DOLCE?

- DOLCE does not commit to a strictly referentialist metaphysics and aims at capturing the ontological categories underlying natural language and human commonsense (Gangemi et al. 2002).
- It is not based on empirical evidence, but it has a formal structure defined on ontological principles and axioms that we do not possess.
- Mutual benefit of the experiment.
- The top-level of WordNet has been aligned to DOLCE, in order to obtain an ontologically adequate lexical resource, meant to be conceptually more rigorous, cognitively transparent, and efficiently exploitable in several applications (Gangemi et al. 2002).
- As a result, CPA classes will be also indirectly linked to wordnet synsets through DOLCE.

Characteristics of DOLCE

- Built according to the OntoClean Methodology.
- The method is based on checking meta-properties (Essence and Rigidity; Unity; Identity), which impose constraints on the taxonomic structure of an ontology.
- They can be used to either validate the ontological consistency of existing taxonomic links, or to create "clean" taxonomic links.

Rigidity 1/2

- A property is *rigid* if it is essential to all its possible instances.
- An instance of a rigid property cannot stop being an instance of that property in a different situation.
- Test: "Can x cease to be y?". If x can cease to be y, y is not a rigid property of x.
- Example: *being a person* is a rigid property, while *being a student* is anti-rigid.

Rigidity 2/2

- Anti-rigid properties cannot subsume rigid properties.
- For example, the property of *being a student* cannot subsume *being a person* if the former is anti-rigid and the latter is rigid.

Unity

- An instance of a class characterized by *Unity* is a whole.
- Test: Can x be arbitrarily scattered? If so, then it lacks Unity.
- For example, *water* does not represent a whole object, while *ocean* does.
- A [-U] property cannot subsume a [+U] property.

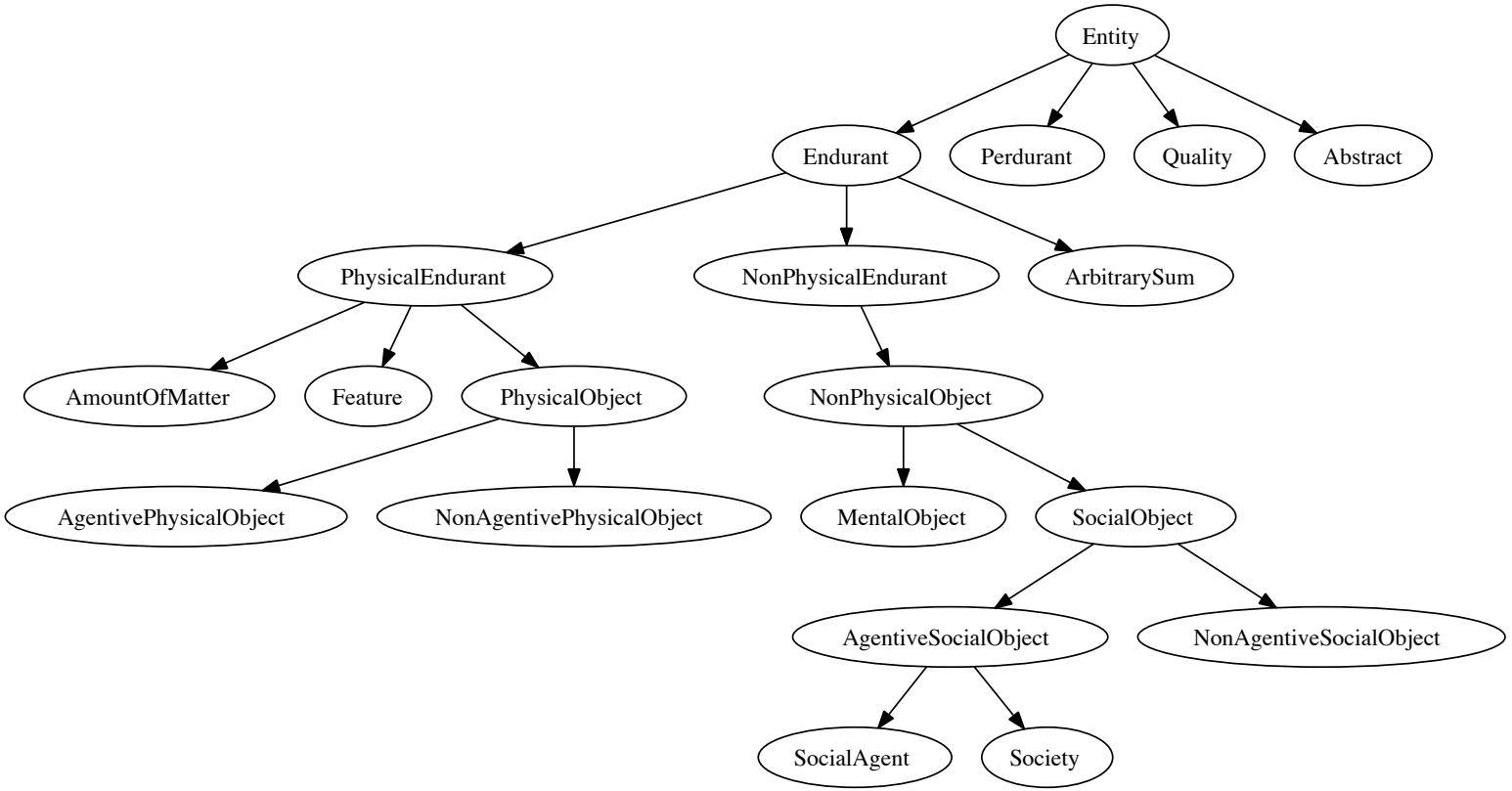
Identity

- To be identical two entities must share the same essential properties.
- For example, a statue is not the clay it is made of, because the statue has the essential property of *having a certain shape*.
- The relationship is not Subsumption but Constitution: statues are constituted by clay, but they are more than clay.

Backbone taxonomy

- The backbone taxonomy is the structure that results from the sum of "clean" subsumption relations.
- It helps in focusing on the more important classes for understanding the invariant, essential aspects of a domain, whereas other relations help in organizing the instances.

Taxonomy of DOLCE basic categories (excerpt)



DOLCE basic categories (excerpt)

- DOLCE top level distinguishes between Endurant, Perdurant, Quality and Abstract.
- An Endurant participates in a Perdurant: for example a *person* (Endurant) may participate in a *discussion* (Perdurant).
- Qualities inhere to entities; every entity comes with certain qualities (color, smell, size, weight etc.), which exist as long as the entity exist.
- Abstracts are entities with no spatial nor temporal qualities.

Further Distinctions in DOLCE 1/3

- Within Endurant, DOLCE distinguishes between Physical and Non-physical (according to whether they have direct spatial qualities).
- Within Physical, a distinction is drawn between between Amount of Matter, Object, and Feature, based on the notion of Unity and the relation of Dependence.
- Object are Endurants with Unity, Amounts of Matter are Endurants with no Unity (none of them is an essential whole).

Further Distinctions in DOLCE 2/3

- Objects and Amounts of Matter are not dependent on other objects, while Features are dependent on another object, their "host".
- Examples of Features are Relevant Parts such as a *bump*, and Places such as *a hole in a piece of cheese*, the *underneath of a table* etc.
- Physical Objects are divided into Agentive and Non-agentive according to whether or not they have intentions.
- Agentive Objects are constituted by Non-agentive Objects: for example, a *person* is constituted by an *organism*.

Further Distinctions in DOLCE 3/3

- Non-physical Objects ("abstracts" in common parlance) are divided into Social Objects and Mental Objects according to whether or not they are are generically dependent a community of agents.
- Social Objects are further divided into Agentive and Non-agentive.
- Agentive Social Objects are for example Societies such as *Sony*.
- Non-agentive Social Objects are *laws*, *norms*, *peace treaties* ecc., which are generically dependent on Societies.

Characteristics of CPA list 1/2

- Classes are identified according to a pattern-based distributional bottom-up analysis.
- No claim of robustness against the state of the art in scientific knowledge (i.e. [[Horse]], [[Dog]] vs. [[Mammal]]).
- The list is linguistically justified; classes reflect the combinatorial preferences of lexical items.
- A class may be motivated by a single verb, i.e. [[Furniture]] for *arredare* "furnish".
- Anthropocentricity.

Characteristics of CPA list 2/2

- Taxonomic structure is mostly based on *prima facie* decisions reflecting our intuition about the meaning ascribed to the terms used and by comparing the lexical sets of different classes.
- Nodes in the structure are classes themselves, i.e. they are identified from a lexical set by observing verb pattern selection.
- Taxonomic structure is highly relevant because the aim is to identify the level of specificity of the selectional properties of V.

Mapping (excerpt)

- Endurant *live in time (and can therefore exhibit changes) by participating in a Perdurant* ->Entity (current) -> Participant
  - Physical Endurant *have direct spatial qualities*
    - Amount of matter *endurants with no unity, none of them is an essential whole, change identity when they change parts (mereologically invariant)* -> Stuff
      - Solid
        - Material
          - Glass
          - Metal
          - Wood
          - Cloth
      - Fluid
        - Vapour
          - Gas
          - Smell
          - Air
        - Liquid
          - Water
          - Beverage [Artifact, Liquid]
            - Water [Beverage, Liquid]
            - Alcoholic Drink
              - Wine
    - Physical Object *endurants with unity, mereologically variant, non dependent on other objects*
      - Agentive *endurants with intentions, constituted by non-Agentive Physical Objects (spatially co-localized with them)* -> Animate
        - Human
          - Fetus [Human, Animal]
        - Animal
          - Horse
          - Primate
          - Cat
          - Fetus [Human, Animal]
        - Bird
        - Cetacean
        - Fish
        - Insect
        - Snake
        - Spider
      - Non-Agentive *endurants without intentions*
        - Inanimate
          - Artifact
            - Artwork *includes video*
              - Movie
              - Picture
            - Beverage [Artifact, Liquid]
              - Water [Beverage, Liquid]
              - Alcoholic Drink
                - Wine
            - Building [Artifact, Location]
              - Cinema
              - Theater
            - Device
              - Software
            - Document [Artifact, Information]
            - Food
              - Meat
            - Garment
              - Footwear
            - Machine
              - Vehicle
                - Road Vehicle
                  - Bicycle
                  - Car *escludes trucks, buses, motorbikes, and cycles*
                  - Motorbike
                  - Truck
                - Water Vehicle
                  - Boat
                  - Ship
                - Plane
                - Train
              - Computer
            - Weapon
              - Bomb
              - Firearm
              - Projectile
            - Container
            - Drug
            - Engine
            - Flag
            - Furniture
            - Image
            - Medium, e.g. radio, TV, the Press
            - Musical Instrument
          - Plant
            - Tree
          - Location (missing in DOLCE?)
            - Natural Landscape Feature
              - Watercourse *includes lakes and the sea as well as rivers and streams*
                - Waterway *canals, also navigable rivers*
              - Hill
              - Land
            - Route e.g. *roads, railways*
              - Waterway
            - Geographical Area *e.g. states*
            - Building [Artifact, Location]
              - Cinema
              - Theater
        - Feature *parasitic entities constantly dependent on physical objects - their hosts (not spatially co-localized with them)*
          - Relevant Part e.g. *bump, damage*
            - Blemish
          - Place e.g. *crack, hole, opening, window, doorway*
            - Aperture
  - Non-Physical Endurant *have no direct spatial qualities* -> AbstractEntity (*different from DOLCE Abstract*)
    - Non-Physical Object *endurant with unity, mereologically variant, non dependent on other objects*
      - Mental Object *non dependent on a human society* -> Concept
      - Social Object *endurants dependent on a community of agents e.g. by means of linguistic acts (not constituted by agentive physical objects, they depend on them)*
        - Agentive
          - Social Agent
          - Society -> Institution
        - Non-Agentive
    - ... Other types of Abstract Entities *such as abstract masses*
  - Arbitrary Sum

Endurant vs. Entity

- DOLCE Endurant is a structuring node which fits very well in the CPA organization.
- DOLCE Endurant links to [[Entity]] in CPA.
- In point of fact an [[Entity]] in CPA is a [[Participant]] in an [[Eventuality]].

Endurants and the Object/Stuff distinction

- DOLCE Physical Endurant does not map onto CPA [[Physical Object]].
- Amount of Matter is a sister node of Physical Endurant in DOLCE, while in CPA [[Stuff]] is a hyponym of [[Physical Object]] ([[Inanimate Physical Object]]).
- It seems reasonable to move [[Stuff]] (and its hyponyms) higher in the taxonomy.

Abstracts and the tangible/non tangible distinction

- [[Abstract Entities]] in CPA are entities without spatial qualities.
- Maps to both DOLCE Abstracts (entities without temporal qualities, such as mathematical objects) and Non-physical Endurants.

Agency and the Animate/Inanimate Distinction

- The label Agent is used in DOLCE to express a potential Agent, i.e. a living being endowed with intentions.
- Physical Objects that have intentionality (i.e. the capability of heading for/dealing with objects or states of the world, cf. Searle) are called Agentive, those which do not are called Non-agentive.
- In CPA [[Agent]] it is not present, as it is considered a role.
- DOLCE Agentive/Non-agentive Physical Objects distinction has no direct equivalent in CPA.
- Agentive Physical Object in DOLCE may be mapped to [[Animate]] in CPA.
- [[Animate]] in CPA excludes [[Plant]] but includes the animal kingdom taxonomy - organized differently from the Lynnean one.

Parasitic Entities

- DOLCE has a node Feature for parasitic entities that are constantly dependent on physical objects (their hosts).
- in DOLCE, Feature subsumes Place and Relevant Part.
- CPA [[Aperture]] links to DOLCE Place and [[Blemish]] links to DOLCE Relevant Part.

Locations

- [[Aperture]] is a hyponym of [[Location]] in CPA.
- CPA has [[Location]] while DOLCE has Place.
- However, CPA [[Location]] does not map onto DOLCE Place, because Place is a subtype of Feature in DOLCE.
- What is the category of DOLCE for [[Location]] such as *islands* or *mountains*?

Natural vs. Artifacts distinction

- Neither DOLCE not CPA distinguish between Artifacts and Naturals. CPA has [[Artifact]] but no Natural counterpart.
- The distinction between Natural and Artifact is orthogonal to other classes.
- Amount of Matter may be Natural (*gold*) vs. Artifact (*plastic*).
- [[Location]] may be a Natural (*a mountain*) or a functional location (*park*).
- [[Feature]] may be Artifact or Natural?

Types vs. Roles

- CPA has [[Food]] and [[Beverage]] as hyponym classes of [[Artifact]].
- "Nothing is necessarily food, and just about anything is possibly food". (Guarino and Welty, 2009, 218).
- "Anything that is food can also possibly not be food, so anti-rigid".
- Food is a role an entity may play in an eating event, not a type.
- Roles are anti-ridig properties that characterize the way something participates to a contingent event.
- The link between *apple* and Food is not Subsumption but rather Purpose.

Systematic Polysemy

- Systematic polysemy is currently treated as multiple inheritance in CPA.
- Not accommodated in DOLCE yet.
- Multiple inheritance in CPA currently includes cases of classic systematic polysemy (*lunches*, *books*, *windows*) and other phenomena such as metonymies, coercions etc.
- [[Document]] [Artifact, Information]
- [[Building]] [Artifact, Location]

Conclusions

- Granularity of classes.
- Mutual benefit of the experiment.
- Insights on the language/cognition interface.

Ongoing and future work

- Complete the alignment of Non-physical Endurants, Perdurants, and Qualities.
- Align the results to DOLCE's version used in the ontology component of Senso Comune resource (Oltramari et al. 2013).
- Accommodate systematic polysemy distinguishing it from coercion (Jezek and Vieu in preparation).
- Compare the results of the mapping to DOLCE's backbone taxonomy with IS_A relations automatically extracted from corpora.

Acknowledgments

- I would like to thank the Senso Comune group, particularly Laure Vieu, Guido Vetere, Alessandro Oltramari, for their input to this research.
- I also thank the audience of the *Wolverhamtpton CPA workshop*, University of Wolverhampton, Aug 28-29, 2013, where this research was first presented, for their fruitful comments.

References

- Gangemi, A. Guarino, N., Masolo, C. Oltramari A., Schneider L. et al. (2002). Sweetening Ontologies with DOLCE. In Gómez-Pérez A. and V.R. Benjamins (eds.) *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02), Ontologies and the Semantic Web*, Berlin, Springer-Verlag, 166-181.
- Guarino, N. and C. Welty. 2002. Evaluating ontological decisions with OntoClean. In *Communications of the ACM*, 45(2):61–65.
- Guarino, N. and C. Welty. 2009. An overview of OntoClean. In Staab, S. and R. Studer (eds.) *Handbook on Ontologies* (second edition), Berlin, Springer-Verlag, 201-220.

References

- Hanks, P. 2004. Corpus Pattern Analysis. In Williams, G. and S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, 87-98.
- Jezek, E. 2012. Acquiring typed predicate-argument structures from corpora. In Bunt H. (ed.) *Proceedings of the Eighth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation ISA-8*, Pisa, October 35, 2012, 28-33.
- Jezek, E. Vieu, L. In preparation. Distributional analysis of copredication: towards distinguishing systematic polysemy from coercion. ms. Università di Pavia, IRIT-CNRS Toulouse/LOA-ISTC-CNR Trento.

References

- Oltramari, A. Vetere, G. Chiari, I. Jezek, E. Zanzotto, F.M. Nissim, M. Gangemi, A. 2013. Senso Comune: A collaborative Knowledge Resource for Italian. In Iryna Gurevych and Jungi Kim (eds.) *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, Berlin-Heidelberg, Springer, 45-68.