Cross-Linguistic Discovery of Semantic Regularity

Ben Wing

CSC 620

Introduction

***What is** metonymy?*

- "A non-literal figure of speech in which the name of one thing is substituted for that of another related to it."

Classic examples from journalism:

- **"The colonies revolted against the crown",** where *crown* refers to the English monarchy. [**"symbol of"** relationship]
- **"The White House promised a thorough investigation",** where *White House* refers to the Office of the President. [**"housed in"** relationship]
- **"The latest Security Council Resolution received praise from London, but Washington threatened to veto it."** where *Washington* and *London* refer to the U.S. and British Governments. Same for other capital cities. [**"located in"** relationship]

- Other examples:

- **"There is a furor in the Roman Catholic Church!"** (so said a preacher in front of the student union, awhile ago). *church* here refers to an institution, not a building. [**"housed in"** relationship]
- **"I prepared five cups of tea."** *cup* is a measurement as well as a container. This can be extended to all sorts of containers: *box, bag, kettle, basket* [**"amount contained in"** relationship] and even *ball* [**"amount contained in volume occupied by"** relationship].
- *wool, fleece,* etc. as materials but also coverings (**"Jason sought the Golden Fleece"**) [**"made out of"** relationship]
- abstract to concrete movement: *institution* meaning "the act of instituting" but also "the result of instituting"; similar for *congregation, building*, *cutting,* etc. [**"entity resulting from an act of"** relationship] Note also *ontology* ["**entity that implements a theory of?"**], *government* ["**entity whose purpose is carrying out an act of?"**], etc.

***What is** polysemy?*

- Very simple -- a single word has two or more meanings.
- The above examples in fact are both metonymic and polysemic; this is called **metonymic polysemy.** [Potentially, the semantic shift could be associated with a shift in form as well.]
- **Regular polysemy** is when the same kind of metonymic relationship between two senses of a word applies to many different words -- i.e. there are systematic connections between different sense of the same word. [examples above]

Aim

- **Question of this paper**: Cross-linguistically, how universal are the patterns of regular polysemy?

- **Prior research**: Mostly small-scale investigations.

- (Kamei 1992) investigated various metonymic relationships in Chinese, Japanese, English (25 test sentences).
  - Result: Often two group up against one.

- (Seto 1996) investigated container-content relationships, as for *kettle* above, in Japanese, Korean, Mongolian, Javanese, Turkish, Italian, Germanic, and English.
  - Result: This particular metonymic relationship seems universal.

- (Peters 2000) identified a certain number of fairly universal relations, e.g. container/content and producer/product.

- But ... all small-scale.
- ***WordNet to the rescue!***
- EuroWordNet covers eight languages. It structures all of them like WordNet, and, crucially, identifies synsets across languages.
- **Methodology:**
  - analyse WordNet 1.6 to obtain English candidates for regular polysemic patterns
  - Process by "lexical triangulation" [since three languages used!]
  - manually evaluate results

Automatic Candidate Selection

- Look for all cases where a word [only noun] has two different senses whose hypernyms are the same as two senses of another word. Group by hypernym pairs.

- **Example [cf. above];**
  - **fabric** (something made by weaving or felting or knitting or crocheting natural or synthetic fibers)
  - **covering** (a natural object that covers or envelops)
- **Words with senses under both hypernyms:** *fleece, hair, tapa, wool*

- **Result: 8062 English nouns.**

Lexical Triangulation

1. **Compare three languages, English, Dutch and Spanish**

   (chosen because they wanted to get different families represented but needed fairly comple wordnets [hence not Estonian?])

2. **Look for English words with two senses in different synsets where the corresponding synsets in both Dutch and Spanish also have a word in common.**
   - Example: *church*, *iglesia* and *kerk* all refer both to a physical building for prayer and a religious institution.
   - **Result:** 920 English nouns.

3. **Intersect the resulting words with the words from the previous section (Automatic Candidate Selection).**
   - **Result:** 404 English nouns (5% of initial 8062).

3. **This was too many for manual evaluation**, so go back to the previous section, throw out groups of exactly two nouns, do the intersection again.
   – Result: 394 English nouns. (?? very little reduction!)

4. **Pick 177 at random for manual evaluation. For each group represented, verify that it's valid.** Examine the two hypernyms that define the group and make sure **[a] they are reasonably specific; [b] there is "semantic homogeneity“** i.e. there is actually a semantic relation, e.g. "is housed in" [cf. examples above] that applies to the majority of words in the group.
   – Result: 109 of the words (62%) displayed valid polysemic patterns, 68 (38%) did not.

==> This automatic filtering method has a 62% success rate for identifying "valid regular polysemic patterns".

Examples:

Hypernymic Pair: Person (a human being) - **Quality** (an essential and distinguishing attribute of something or someone)

English RP class (11 total): *attraction, authority, beauty, ...*

Dutch RP class (1 total): *schoonheid*

Spanish RP class (4 total): *belleza, atracción, autoridad, imagen*

Word intersection between all three languages: 9% of set derived from WordNet

Hypernymic Pair: Control (the activity of managing or exerting control over something) - **Trait** (a distinguishing feature of one's personal nature)

English RP class (7 total): *abstinence, sobriety, inhibition, restraint, self-control, self-denial, self-discipline*

Dutch RP class (2 total): *zelfcontrole, onthouding*

Spanish RP class (3 total): *abstinencia, abnegación, inhibición*

Word intersection between all three languages: 36% of set derived from WordNet

Examples (cont’d)

Hypernymic Pair: Plant (a living organism lacking the power of locomotion) - **Edible Fruit** (edible reproductive body of a seed plant especially one having sweet flesh)

English RP class (159 total): *apple, boxberry, blackcurrant, banana, fig...*

Dutch RP class (9 total): *banaan, vijg, persimoen, meloen...*

Spanish RP class (20 total): *banana, plátano, melón, caqui, hijo...*

Word intersection between all three languages: 2.5% of set derived from WordNet

Hypernymic Pair: Occupation (the principal activity in your life) - **Discipline** (a branch of knowledge)

English RP class (6 total): *architecture, literature, politics, law, theology, interior design*

Dutch RP class (1 total): *architectuur*

Spanish RP class (2 total): *arquitectura, teología*

Word intersection between all three languages: 16% of set derived from WordNet

Universality of Regular Polysemy

Conclusions:

1. Potentially indicative of the cross-linguistic validity of these particular relationships
2. But what about the low coverage? Dutch and Spanish wordnets are less complete (66025 synsets for English, 28352 for Dutch, 24073 for Spanish) but the numbers still seem way way low, usually only 2-5%. Possibilities:

- The regular polysemic pattern is not in fact universal across all three languages, or at least not equally productive.
- The pattern is valid but the missing sense just happens to be unattested, and could such an extension would be a valid usage in the language.
- The pattern is valid but the missing sense is blocked by an already existing word with that meaning. (E.g. *club* in English is either an organization or the building housing the organization. The Dutch equivalent *vereniging* can only refer to the former, and *verenigingshuis* is used for the latter.) *** But note English *club house*! Why is there no blocking effect here?
- The missing sense is in fact attested, and the problem is in WordNet. (E.g. Dutch *ambassade* means either an embassy building or the organization inside it that represents a country. English *embassy* has the same two senses but only the first is in WordNet. Oops!)

Coverage and Extendibility

Which of the above possibilities apply? Very small experiment:

1. Choose the following pair:

*** I have a hard time understanding the distinction between the two senses.

Hypernymic Pair: Occupation (the principal activity in your life) - **Discipline** (a branch of knowledge)

English RP class (6 total): *architecture, literature, politics, law, theology, interior design*

Dutch RP class (1 total): *architectuur*

Spanish RP class (2 total): *arquitectura, teología*

Word intersection between all three languages: 16% of set derived from WordNet

2. For each missing word in Dutch or Spanish, ask two native speakers to help sort things out -- does (or can) the word extend to the other sense?

3. Results:
    - Of the 5 missing Dutch words, 3 in fact had both senses lumped together (one meaning, linked to both senses by a near-synonymy relation). Hence, WordNet bug. (For one of the three words, though, the extension from discipline to occupation was not judged acceptable by the native speakers, hence another bug.)
    - Of the remaining 2 Dutch words, 1 could be extended.
    - Of the 4 missing Spanish words, 2 could be extended, 1 could not, and for 1 the informants disagreed.
4. Hence: 50% of words could be successfully extended to follow an automatically-derived regular polysemic pattern.

Conclusions

- The same methods could be applied beyond EuroWordNet, to any multilingual resource with hypernymic relations and correspondences between languages.
- Some regular polysemy patterns are valid across all three languages and appear to have some universality.
- There is potential for (semi-)automatically enhancing the semantic compatibility and consistency of wordnets through meaning extensions based on regular polysemic information (patterns) derivable from other wordnets.