Generalizing Syntactic Collocates for Creative Language Generation

David Hardcastle
www.davidhardcastle.net
Overview

- **What?**
  - Collocational Semantic Lexicon
- **Why?**
  - Background and Motivation
- **How?**
  - Implementation and Evaluation
  - Research Findings
Collocational Semantic Lexicon

- Shallow semantic information to support lexical choice over a given syntactic relation.
- Defines domain of R(a,b) for some syntactic relation R:
  - amod(green,x) → amod⁻¹(x,rose)
  - dobj(cook,x) → dobj⁻¹(x,lamb)
  - subj(drive,x) → subj⁻¹(x,author)
Motivation

- **ENIGMA** – a system that generates cryptic crossword clues
- Layered text with two readings
- **NLC** – Natural Language Creation
  - SemRep input reflects puzzle reading
  - Many lexicalizations
  - Some of these map to valid SemReps for surface reading
  - SemRep-SemRep ‘translation’ via text
Motivation

- Part of the NLC process involves applying semantic constraints to syntactic relations.
- We don’t have a semantic representation for the surface reading.
- Need to make lexical choices based on collocational semantic information that is:
  - compositional
  - fine-grained
  - efficient
Mind awkward bairn (5)
Mind awkward bairn (5)
Mind awkward bairn (5)
Generating Clues by NLC

- Start with ‘clue plan’ – semantic representation of the puzzle
- Bottom-up lexicalization
- Apply syntactic constraints and semantic constraints
- Construct successively larger ‘chunks’ representing nodes of the clue plan
- Explore solution space efficiently
Puzzle Reading Constraints

- Anagram(bairn)
- Any \(ab\) or \(ba\) where:
  - \(a\) in \{bairn\}
  - \(b\) in \{out, off, broken, lost, wander, grilled, strangely, stagger, odd, cook, confuse, addled, unusual, awkward, wild, reorder, train, stew, organise, tumble, vary, swirl, \ldots\}
Syntactic Constraints

- **amod**: Adj Noun
  - {broken, lost, grilled, translated, odd, awkward, unusual, wild, …} **bairn**

- **dobj**: Verb Noun
  - {cook, confuse, reorder, train, stew, organise, …} **bairn**

- **subj**: Noun Verb
  - **bairn** {wanders, roams, tumbles, varies, swirls…}
Semantic Constraints

- **amod**: Adj Noun
  - {broken, lost, grilled, translated, odd, **awkward**, unusual, wild, …} bairn

- **dobj**: Verb Noun
  - {cook, confuse, reorder, **train**, stew, organise, …} bairn

- **subj**: Noun Verb
  - bairn {wanders, **roams**, tumbles, varies, swirls…}
Sample Generated Clues

- Strangely tiny scale drawing (5)
- Entertained dicky dames around bend (6)
- Unusually ripe and fresh soul is dangerous (8)
- Tape loud squeal behind ford (5)
- Official measure is falling in essential merits (8)
- Feed store (4)
- A falling look in tune (4)
- Welcome bad receipt before back number (9)
- Mark chopped capers (6)
- Pouring fresh gin after back play (7)
- Everlasting branch is falling in rough seas (7)
- Mind awkward bairn (5)
Approach

- Extract syntactic collocates from BNC using a parser (Stanford)
- Generalize the result data using WordNet
- Organize this data into a lexicon
Extraction

- Used Stanford (although trained on WSJ)
- Also explored regular expressions
- 1,000 hours machine time to parse BNC
- Perhaps better to use MiniPar, RASP or CCG parser for this task?
Generalization: Strategy

- Need to tackle data sparsity – everything has distributional context, not so for relational context
- Generalize using WordNet
- Can’t generalize if insufficient data
- Can’t generalize collocations
- Some sense disambiguation, but can only do this where we can generalize
Generalization: Aim

- **beloved**: wife, daughter, girl, father, grandson, homeland, mother, child

- any woman

- any relative

- beloved-homeland
Generalization: Algorithm

- **Construct Sub-trees**
  - 1. spread cooccurrence frequency across all senses
  - 2. collate by lexnum (stand-in domain)
  - 3. reallocate to sense with highest lexnum weight
  - 4. repeat from 2 until no movement

- **Filter Sub-trees**
  - locate sub-trees using arc-distance
  - test each sub-tree for coverage using all senses
  - record sub-tree roots in lexicon
  - retain singletons with high log-likelihood
  - mark top slice as candidate collocates
Grilled: Initial Allocations

- act
- animal
- artifact
- attribute
- body
- cognition
- communication
- event
- food
- location
- object
- person
- phenomenon
- plant
- possession
- quantity
- relation
- shape
- state
- substance
- time
Grilled: Filtered Allocations

- act
- animal
- artifact
- body
- event
- food
Grilled: Lexicon Entry

- any hyponym of: saltwater_fish, meat
- any of: mullet, sandwich, lobster, prawn, tomato, cob, cheese
- discarded: kidney, square, affair
Entry for Rough

Any hyponym of: fabric, way, indicator, interpretation, representation, surface, structure, explanation, utterance, horizontal_surface, drawing, facility, product, time

Snapshots – a side effect

**brave** person (56%), **act** (30%), communication (8%)

**political** cognition (45%), **act** (21%), **group** (9%), **person** (9%)

**broken** artefact (75%), **body** (8%), communication (7%)

**fallow** animal (32%) time (53%)
Oh nuts!

- ‘scrambled egg’

- WordNet senses:
  - egg food
  - ovum body
  - testis body

- scrambled body-part
Evaluation

- No gold standard – question of judgement
- Some disagreement between human judges – gold standard not possible
- Hard to rate performance (51/60 choices match human choice for 30 subjects, p<0.01)
- Out-performed statistical measure of word association (11/60) – too focused on typical usage
Research Findings

- Statistical methods return typical or prototypical usage, but we want to know about correct usage.
- Granularity is crucial to informing lexical choice in NLG – and existing resources (e.g. FrameNet, PropBank, VerbNet) are very coarse-grained.
- Mapping these resources to (e.g.) WordNet increases vocabulary but not granularity.
- Corpus-based approach can produce a large-scale dataset with fine granularity.
Research Findings (2)

- Generalization relies on isomorphism between feature sets that drive domains and structure of WordNet

- Some examples:
  - meronymy
  - synecdoche
  - domain vocabulary
  - ranged data
  - WordNet topology
  - figurative speech
  - predicate polysemy
Research Findings (3)

- Cautious parameterisation ensures a large number of small generalizations are made, reducing the impact of the problems listed above on data quality.
- So, fine granularity to some extent mitigates the problems raised by using a static classification system.
- Really need a classification system that is multi-dimensional – reflecting the different feature-sets that underpin domains.