Whither WordNet?

Christiane Fellbaum

George A. Miller

Princeton University

WordNet was made possible by...

Many collaborators, among them

Katherine Miller, Derek Gross, Randee Tengi, Brian Gustafson, Robert Thomas, Shari Landes, Claudia Leacock, Martin Chodorow, Richard Beckwith, Ben Haskell, Susanne R. Wolff, Suzyn Berger, Pam Wakefield,....many, many students

Somewhat fewer sponsors

ARI, ONR, (D)ARPA, McDonnell Foundation, LDC, Mellon Foundation, ARDA/AQUAINT, NSF

A bit of history

- 1986: George Miller plans WordNet to test current theories of human semantic memory (Collins and Quillian, *inter alia*)
- 1987: verbs are added to WordNet
- 1991: first release of WordNet version 1.0
- 1998 EuroWordNet (Piek Vossen)

...

- 2002: WordNet goes Global
- 2006: approx. 8,000 downloads daily

  WordNets in some 40 languages

The Good...

- WordNet is freely available; Princeton provides user support
- WordNet is customizable
- Princeton releases serve as standards for the NLP community
- WordNet is large: coverage and average polysemy are the same as those of standard collegiate dictionaries

...the Not So Good...

- No (sufficiently large) corpus was available when WordNet was built. Entries are largely created by lexicographers
- WordNet was an experiment. There was no prior model and no plan to build an NLP tool. Add-ons rather than re-design
- Sparsity of relations and links was not an issue. Evidence for syntagmatic associations (Fillenbaum and Jones, *inter alia*) was ignored
- Duplicate, overlapping senses? Excessive polysemy? Not a problem if you consider WordNet as a thesaurus (as we did early on)

...and ome desiderata...

- Users articulate ideas and needs for specific improvements
- Sharing of resources and tools that can be folded into WordNet or speed up enhancements
- Merging and alignment of resources (e.g., FrameNet-WN)
- Communication, collaboration, division of labor among research teams and users rather than competition and duplication of efforts
- Maintain balance of (psycho)linguistic/symbolic and statistical perspectives

Create **few** resources with **many** kinds of annotations, incl.

--word senses

--subjectivity (Wiebe)

--temporal relations (Pustejovsky)

--frames (Berkeley FrameNet)

etc.

Greatest Challenge: WSD

- People do it effortlessly--but how?
- Implicit assumption: Dictionary model of word sense representation
- When the dictionary user encounters a sense that fits the context, he can close the dictionary
- Other senses may fit as well, but redundancy is not a problem
- But automatic systems must select one sense over others

Greatest Current Challenge: WSD

- Early experiments with semantic tagging (Kilgarriff 1991, Princeton SemCor) showed that people often have trouble selecting the dictionary sense of a polysemous word that is appropriate to a given context
- One solution: sense clustering, underspecification
- But clustering often involves mutually exclusive criteria (semantics, syntax, frames, domains)
- “forced choice”? Offer only few sense alternatives to taggers

Current Work: Gloss Annotation

(Work sponsored by ARDA/AQUAINT)

- Nouns, verbs, adjectives, adverbs in the definitions (glosses) of WN synsets are manually linked to the context-appropriate synsets
- Closed system--WN database is in synch with the annotated glosses

Gloss Annotation

- Annotators can choose pre-defined sense clusters or any combination of multiple senses
- Combinations of senses suggest new clusters
- Never-used senses: redundant?
- Targeted tagging (all tokens associated with a given string)
- Database editing proceeds in parallel based on feedback from annotators
- Hope: tagged corpus of glosses will be helpful for automatic WSD

Current Work:WordNetPlus

(with Jordan Boyd-Graber, Daniel Osherson, Moses Charikar and Robert Schapire)

Work supported by the NSF

WordNetPlus

Motivation: WSD would be easier if WN were more densely connected

But how to overcome sparseness?

WordNetPlus

- Current WN relations are few, mostly “classical”, mostly paradigmatic
- Why not others? Word association norms show that WN relations account for at most half of the responses given. Major lack: cross-POS, syntagmatic relations
- There are many dimensions of meaning similarity
- Maybe we lack imagination or cannot articulate or label many kinds of semantic similarities?

Basic Idea

Connect **all** synsets (within/across POS) by means of **directed, weighted** arcs

WordNetPlus

- Dense network can be exploited to find all related/unrelated words and concepts
- Graded relatedness allows for finer distinctions
- Less training data needed for automatic WSD
- Algorithms relying on dense net structure will yield better results

From WordNet to WordNetPlus

- Cross-POS links (*traffic*, *congested*, *stop*)
- New relations: *Holland-tulip*, *sweater-wool*, *axe-tree*, *buy-shop*, *red-flame*,...
- Relations are not labeled!
- Arcs are directed:*dollar-green/*green-dollar*
- Strength of relation is weighted

From WordNet to WordNetPlus

Arcs capture **evocation**

Evocation:

“How strongly does concept A bring to mind concept B?”

From WordNet to WordNetPlus

Method

Depart from empirical data

Scale up automatically

Multiple Paths to Evocation

- rose - flower (hyponomy)
- banana - kiwi (co-hyponyms)
- egg - bacon (co-occurrence)
- check - money (topic/domain)
- yell - talk (troponymy)
- yell - loud (?)
- yell - voice (~instrument)
- wet - dry (antonymy)
- dry - desert (prototypical property)
- wet - desert (~antonymy)

etc.

From WordNet to WordNetPlus

- We identified 1K “core” synsets:
- Central member of synset is a highly frequent string in the BNC
- Manually determined the most salient synset(s) containing that string
- Distribution across POS reflects that in the lexicon:

642 noun synsets

207 verb synsets

151 adjective synsets

Collecting Evocation Ratings

- Based on synset--not word--pairs
- “How strongly does S_1 bring to mind S_2?”
- Avoid idiosyncratic associations (*grandmother-pudding*)
- Try to guess “average student’s” ratings
- Avoid formal similarity (*rake-fake*)
- Most synset pairs will not be related by evocation

Collecting Human Ratings

- Wrote rating manual
- Designed interface for ratings with sliding bar to indicate strength of association
- Strength of evocation ranged from 0-100
- Five anchor points with verbal label (no/remote/moderate/strong/very strong association)

Experiment cont'd

- Two experimenters rated evocations for two groups of 500 synsets each (gold standards for training and testing)
- Mean correlation was .78
- This was a (pleasant) surprise!

Evocation Ratings: Training and Testing

24 Princeton students rated evocations for one group of 500 synsets (the training set)

After each rating, the gold standard rating appeared as feedback

Students then rated the second group of 500 synsets without feedback (testing)

Calculated Pearson correlation betw. annotators' ratings and gold standard

median .72

lowest .64

avg. correlation between training and testing .70

Collecting Ratings

- Post-training/testing: collected judgments for 120K randomly chosen synset pairs (subset of 1K)
- At least three raters for each synset pair

Example Ratings

code-sip	0
listen-recording	60
pleasure-happy	100

Two thirds of ratings (67%) were 0

WordNetPlus Ratings and Other Similarity Measures

Rank order Spearman Coefficient for similarity measures (cf. WordNet::Similarity, Pedersen & Pathwardhan)

Leacock & Chodorow (similarity based on WordNet structure): 0.130

Lesk (overlap of strings in glosses): 0.008

Peters' Infomap (LSA vectors from BNC): 0.131

WordNetPlus Ratings and Other Similarity Measures

Lack of correlation shows that Evocation is an empirical measure of semantic similarity that is not captured by the other measures

Partial explanations:

WordNet-based measures are within, not across, POS

Leacock & Chodorow do not capture similarity among verbs or adjectives

LSA is strictly string, not meaning-based

Measures are based on symmetric relations, but evocation is not

Scaling Up

- Collection of 120,000 ratings took one year
- To connect all 1,000 synsets, 999,000 ratings are needed
- Too much to do manually!
- Current work: build an annotator “robot”
- Learn to rate evocations like a human

Features for Machine Learning

- WordNet-based features:

Jiang & Conrath

WN Paths

Lesk

Hirst & St.Onge

Leacock & Chodorow

Features for Machine Learning

Context vectors derived from the BNC:

Relative Entropy, Frequency,...

Machine Learning Evocations

- Boosting (Schapire & Singer’s BoosTex)
- Learns to automatically apply labels to examples in a dataset

Preliminary Results

- Algorithm predicted the right distribution of evocations (many 0's)
- For some data points with high (human) evocation ratings, prediction was zero evocation
- For many data points with zero (human) evocation, high evocation was predicted
- Best performance on nouns
- Worst on Adjectives

Work is ongoing...

WordNetPlus will be made freely available to the community

Link WordNetPlus to Global WordNets?

Thank you