Adding synsets to WordNet: Why, when, how?

Christiane Fellbaum

Princeton University

Looking back

WordNet was not built for NLP

Developed before the community discovered it

Modifications, additions were done piecemeal, often determined by a particular funder

e.g., Navy grant motivated entries like

{head (nautical) a toilet on a boat or ship}

New wordnets offer the chance to do it better/ right

Looking ahead

Make WN and all wordnets a better tool for language processing, linguistic research

Human and machine use

--Provide coverage in targeted ways

--Update

--Improve quality of meaning representation

--Ensure consistency

--Ensure compatibility and potential interfacing with other resources-

What can be learned from development of wordnets in other languages?

What is missing or amiss in PWN3.1?

Missing: the odd entry in the middle level

Entries for new concepts

Many existing entries and definitions need updating

Coverage of systematically related senses is incomplete

Wishlist--Main Points

Update PWN

Proceed systematically

Align with other wordnets

The lexicon is dynamic

Regular processes that add to the lexicon

Verbification of nouns (*to google, to skateboard*)

Meanings of verbs cannot be derived in a regular fashion from those of the nouns! (Osherson et al.)

Morphosemantic noun-verb links must be added in each case with definition

Updates
The lexicon is dynamic

Words and meanings come and go

New concepts: *smombie, vape, selfie (stick), blog, emoji, (un-)friend, go_viral, meme, hoverboard, tweet, twitter, book-book,…*

Fashionable foods (mostly loanwords): *farro, edamame* (cf. Jurafsky's book about fancy restaurant menus)

Words (incl. loans) tied to recent events: *tsunami, bird_flu, Ebola, Hiroshima*

Proper nouns => common nouns, verbs: google, facebook (verb)

[Caveat trademark lawyers!]

Ins and Outs

Delete or mark as such (with date range?) obsolete entries:

cassette player, betamax, *rotary_dial, percolator, debutante, millenium_bug*
(need to wait another 984 yrs!)

Groovy

Politically incorrect/insensitive words (*dwarf, retarded*)

--replace or add current/accepted terms

secretary=>administrative assistant

janitor,cleaner=>custodian

Negro=>Black, African-American

This raises the descriptive vs. prescriptive question. Politically incorrect and outdated words will always show up in (historical) corpora.

What to add?

Sublanguages (e.g., youth language: *sweet, beast, hookup*)

Emoji--arguably meaning-carrying “ideographs”?

What and when to add: Criteria

Cannot include everything. How to select?

Frequency in open-domain corpora?

This covers forms only, not (necessarily) meanings

Some word/meaningss may have a very short half-life

Define a frequency/time metric as a threshold before creating a new synset (member)?

What to add?

Acronyms

Internet/texting language (much of current communication is in this form)

LOL, OMG, KWIM, YGWYD,…

Can be polysemous (*LOL*: laughing out loud/lots of love)

ACL: Association for Computational Linguistics/ anterior cruciate ligament)

What to add?

Proper Nouns

Potentially unlimited

Names of countries:

--disappear (*German Democratic Republic)*

--change (*Zaire=> Democratic Republic of Congo*

--change superordinate in political re-organization (cf. breakup of Yugoslavia)

What to add?

Proper Nouns

What is part of cultural knowledge?

--people (historical figure; real and fictional)

--events (wars, institutions, works of art,…)

A few lexicographers cannot capture shared popular culture

Crowdsource?

A word about definitions

Originally not part of the “net”

Meaning representation in terms of relations only was found to be insufficient for NLP—not enough discrimination among senses

Now: much of the semantic burden is carried by definitions

A word about definitions

Follow a standard format

Probably different for different POS

Parsers need to be tuned

Avoid boolean expressions, esp. *or*

Examine definitions that may cover sub-classes
(*such as*, *including*,…should raise red flag

A word about definitions

Ensure that all content words/senses are represented in synsets

Link content words (senses) to synsets ("Gloss corpus")

Don't replicate Wikipedia's world knowledge

Update (*book, phone*)

Adding leaves

Terminology (medical, biological, legal,…}

Can’t possibly do it for all domains and don’t have competency

Train experts!

Where possible, include both expert and lay terms in a synset

{patella, knee_cap}

{*chimpanzee, chimp, pan_troglodytes*}

New entries: systematic coverage

Many kind of lexemes must be entered consistently and systematically

Esp. MWUs

New entries: systematic coverage

Phrasal verbs

--there are many

--they are often polysemous (e.g., *break_down*)

--meaning is often non-compositional

Can be hard for POS taggers, parsers to detect and identify as a single lexical unit

Systematic additions
MWE

"Fixed" expressions

Idioms (*hit the ceiling, rock the boat*)

Not as fixed morphologically, syntactically, lexically as often claimed!

Lexical entry should support automatic identification and interpretation, even in non-canonical form

PWN has linked many idiom constituents to appropriate synsets

Systematic additions
MWU

Light/support verb expressions

V+NP (*commit a crime, take a break*)

V+PP (*come to a decision, get to the point*)

There are many…

Many are synset mates of simplex verbs

Nonlexicalized synsets

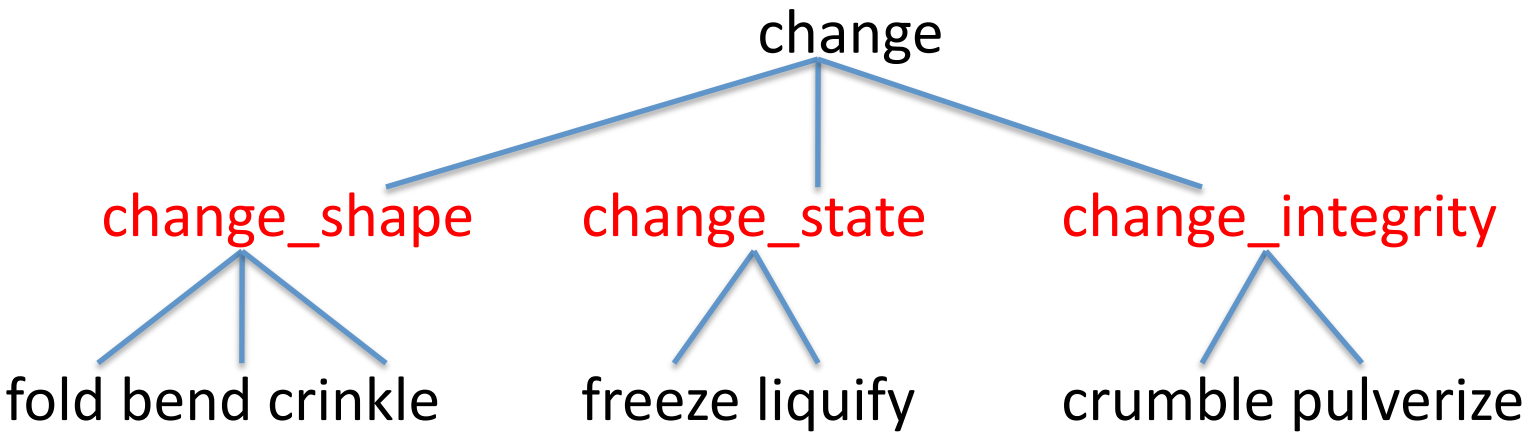
PWN includes many

Motivated by the need to distinguish sub-categories

Intuitive but backed up by corpus data (e.g., classes of verb arguments)

Expect some/many to be lexicalized in other languages (accidental lexical gaps in English?)

Lexical gaps?



Some unresolved issues

Tags: lexeme-specific

Register (slang, formal, youth language,….)

Regional, Dialects (British vs. US English, NY vs. TX,…)

How many, which?

How to avoid overlap?

Domain tags

Domain tags (nautical, medicine, math) apply to entire synset

Link tag lexeme back to appropriate synsets

Design domain ontology at the right level of generality that complements noun trees

Attach tags at the highest possible level

Children of tagged synset inherit that tag/are part of that domain

Major challenge

Where does the lexicon end and grammar begin?

How much grammatical information should be encoded in the lexicon?

Can the lexicon interface successfully with grammar rules that apply to specific lexical items/classes?

Lexicon-Grammar

English verb alternations

Regular and productive over large verb classes

Different syntax—different meanings, different hypernyms?

Broader question: should syntax drive semantic distinctions?

Lexicon-Grammar

Unaccusative (causative/inchoative)

John broke/cracked/chipped the cup =>{change, modify, alter, make_different}

The cup broke/cracked/chipped=>{change, become_different}

WN's structure forces sense distinctions via different hypernyms

Pairs are not uniformly encoded/linked in PWN3.1

Lexicon-Grammar

Middle alternation

PWN structure forces distinct senses

Toyota **sells** millions of this model=>{*exchange*}

This model **sells** easily/quickly=>{*be, have_a quality}*

Transitive verbs have many different supreordinates

Intransitives all have superordinate {*be*}

Lexicon-Grammar

Locative alternation

*John **loaded** the wagon with hay => {fill}*

*John **loaded** hay onto the wagon=> {put, place}*

Alternations

Encoding usages/senses in separate entries increases polysemy

Seen as undesirable by many NLP researchers

But crucial for good processing!

Need to interface with parser

Alternative for representing alternations?

If both usages/senses are combined in one synset, what should the definition be?

Traditional dictionaries definitions:

to V or cause to be Ved

May pose problems crosslingual mappings

Other productive lexical processes

***Out*-prefixation**

John **sells**/swims/performs (intransitive; different superordinates)

John **outsells**/outswims/outperforms Bob (transitive; shared superordinate *surpass*)

Productive processes

Human users have grammar that interfaces with the lexicon

Can interpret verbs with *out*-prefixation and alternation-induced meaning changes

Computers cannot—they need lexical entries

Verb hierarchies

Troponymy relation links general (*x)* and more specific (*y*) verbs

To y is to x in some manner

This formula does not apply to many prefixed pairs

Productive processes

pre-prefixation

We are pre-boarding passengers with children

⇒*board (initial/early boarding; pre-select; pre-print, pre-register, pre-pay, pre-process)*

Troponymy?

re- (again): *reheat, repaint, reconstruct*,…

Expand meaning of troponymy or add new type of link for temporal relations among verbs?

Adding new synsets, senses

Generally agreed-upon desideratum:

limit polysemy

Use corpus-based context similarities to determine sense distinctions

Both n-grams and classes of syntactic arguments

(e.g., which verbs select for noun subjects or objects referring to vehicles?)

Conclusions

- Proceed systematically
- Can’t ever be complete
- Use corpora to guide selection of words and senses to include
- Consider outsourcing
- Train experts to encode domain terminology
- Lexicon-grammar interface remains a significant challenge
- Expanded meanings of relations vs. new relations?