How and when to add new concepts and how to define them

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 & maintain

--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

Coverage of systematically related senses is incomplete

Many existing entries and definitions need updating

No uniform definition format

Candidates for addition

Two types of additions

Words, senses that have entered the language

Arguable accidental gaps

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

New words and meanings

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)

[Cave trademark lawyers!]

Additions?

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

--replace or add current/accepted terms to synsets

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 Politically incorrect and outdated words will always show up in (historical) corpora. Keep and tag them?

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/meanings may have a very short half-life

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

Use, e.g., Google book corpus

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?

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 (not the case currently)
- Esp. Multi-Word-Units

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 (Osherson & Fellbaum 2010)

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

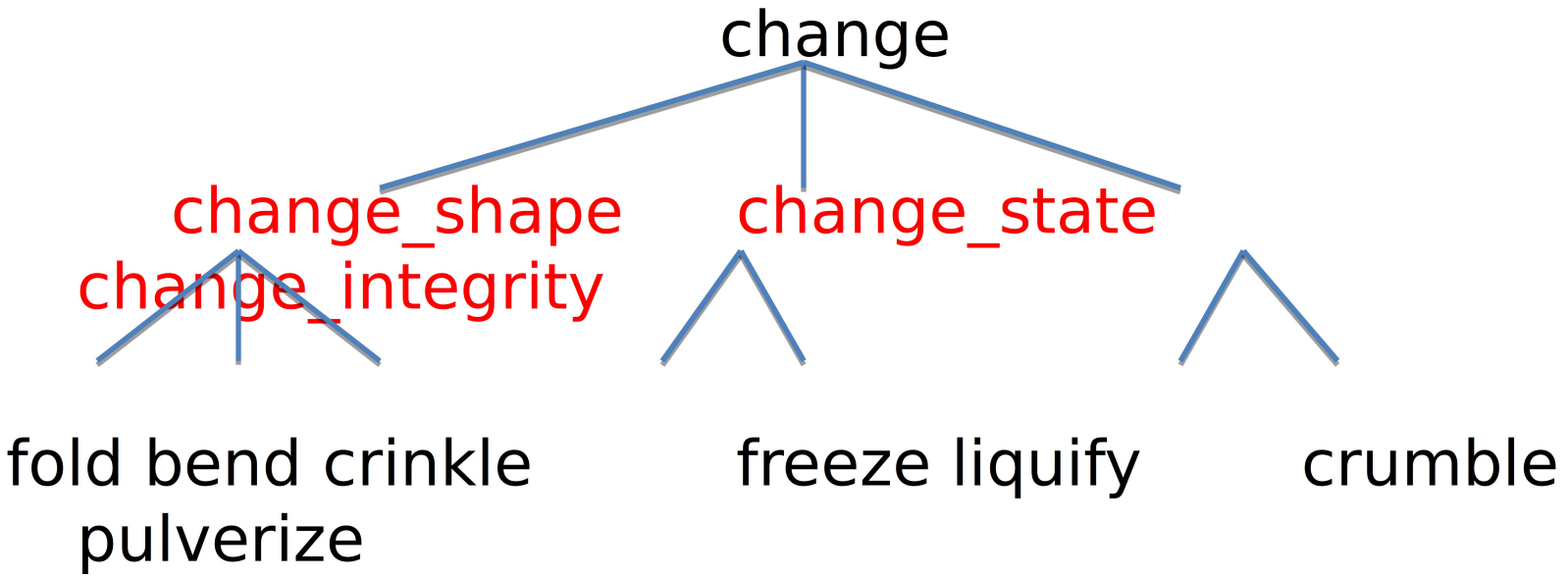
Motivated by the need to distinguish sub-categories

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

Can be semi-automatically discovered through clustering of arguments

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

Lexical gaps?



Senses or usages? 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}

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

Alternations

- Encoding usages/senses in separate entries increases polysemy
- Seen as undesirable by many NLP researchers because it degrades performance
- But crucial for good processing, applications such as translation
- Need to interface with parser

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

Definitions

Now: much of the semantic burden is carried by definitions

Definitions clarify meaning of the synset members distinguish it from similar meanings

Words in definitions contribute to “bag of words” associated with a given concept

The “Gloss Corpus”

Replaces SemCor

Content words (senses) in definitions (“glosses”) are manually linked to the appropriate synsets

Corpus (= the set of all definitions) is useful for

--training and testing WSD systems

--informing learners about the meaning of the headword(s)

The task ahead

Currently, not all content words/senses are represented in synsets

These can be automatically identified

This requires manual inspection, linking

And/or: bootstrap additional links?

Improving definitions

WordNet creators often could not come up with a definition

Used a synonym instead

{***run*** *(****carry out*** *(an errand*))

This may not cause no harm but it’s bad form

Bad definition

Defined word should not appear in the definition

run (stretch out over a distance, space, time, or scope; **run** or extend between two points or beyond a certain point)

Improving definitions: Maintenance

Update definitions to reflect changes in meaning

(*book, phone*)

Some concerns

Don’t replicate world knowledge in definitions

Links to external knowledge sources exist (e.g., Wikipedia)

kat, **khat**, qat, quat, cat, Arabian tea, African tea (the leaves of the shrub Catha edulis which are chewed like tobacco or used to make tea; has the effect of a euphoric stimulant)

Good WordNet definitions

Follow a standard format

Different for different POS

Avoid boolean expressions, esp. *or*

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

Definition guidelines

A definition should be a phrase that

--belongs to the same POS as the headword

e.g., the definition of a noun should be a phrase headed by a noun

--a relative/subordinate clause specifying the difference with the superordinate (i.e., the differentiating properties that distinguish it from its parent and its sisters)

Definition guidelines

NP head of a definition for a noun, modifier

synonym NP:

ambition (a cherished desire)

NP with relative clause:

ski (narrow wood or metal or plastic runners used in pairs for gliding over snow)

Definitions

Verb Phrases

V+Adv

leap *(move or jump suddenly)*

V+PP

dance *(move in a graceful and rhythmical way)*

V (+NP)

jump *(start (a car engine…))*

Definitions--Syntax

- Mostly incomplete sentences
- Parsers need to be tuned
- Phrases (like NP, VP) are categories that parsers can recognize

Bad definitions

- Identify and correct
- Identification can be done automatically in many cases
- Check for duplicate of headword, definition
- Specific syntactic structures

Bad definitions
Some examples

Verbs with noun arguments that are not syntactically distinguished from synonyms

operate, **run** (direct or control; projects, businesses, etc.) "*She is running a relief operation in the Sudan*"

Bad definitions
Alternative view

- Words referring to typical arguments of verbs provide context, should be included but identified as arguments
- Semantically similar words can be identified via links

Bad definitions

Possibly hiding multiple senses:
“or/such as/including/etc./mainly..”

Can be automatically identified

Bad definition?

book (a number of sheets (ticket or stamps etc.) bound together on one edge) "*he bought a book of stamps*"

Bad definitions?

stand (a small table for holding **articles of various kind**s) "*a bedside stand*"

Not helpful—too high-level category

Literal and metaphoric meaning

stand (occupy a place **o**r location, **also metaphorically**) "*We stand on common ground*"

Split into multiple synsets or shove into different part of processing

Definitions or Bag of Words? A re-think

Alternative proposal:

Don’t worry too much about format of definitions

Consider them primarily bags of words that supply context and support WSD

Unlike in paper dictionaries, related words need not be part of the definition

Rethink definitions?

The head noun/verb need not be the superordinate

motorcycle, bike (a **motor vehicle** with two wheels and a strong frame)

Superordinate (hypernym) is given independently/part of network structure

Rethink definitions?

Meronyms don’t need to (re)appear in the definitions:

car, auto, automobile, machine, motorcar (a motor vehicle with four **wheels;** usually propelled by an internal **combustion engine**)

Rethink definitions?

Instead, add information such as the **use** of an artifact:

sawhorse, ***horse***, *sawbuck*, *buck* (*a framework for holding wood that is being sawed*)

Or characteristic **arguments** or **adjuncts** of a verb:

slap (hit with something flat, like a **paddle** or the **open hand**)

Conclusion

Move from framework inherited from classical lexicography towards perspective of automatic semantic processing