Where is Princeton WordNet headed?

Christiane Fellbaum

Four major shortcomings

We have addressed some of them, but only partly

Others have not been addressed at all

Your views, advice (and possible actions in and beyond this workshop) are appreciated!

Three major shortcomings

Sparse connections

Senses (synsets) cannot always be clearly discriminated by automatic systems

Few cross-part of speech pointers

WordNet is basically four distinct networks

Users rely on “glosses” for syntagmatic relations

Few cross-part of speech pointers

Noun attributes are linked to adjective "dumbbells"

"Morphosemantic" relation

Domain labels link some verbs, adjectives to noun "domains"

But: inconsistent granularity of domains

Need structured domain ontology

Glosses

Manual annotation of glosses has addressed the x-POS problem

But many more glosses need to tagged!

On the to-do list: standardize syntax of glosses make them easier to parse

No root nodes

Only noun trees have a shared root (“entity”)

Adjective “dumbbells” are disconnected

Verb “bushes” have no shared root

Problem for edge-counting algorithms!

General disregard of verbs, adjectives

Arcs are not weighted

But intuitively, the semantic distance differs

e.g.

white elephant$_1$ ISA possession

white elephant$_2$ ISA elephant

Semantic distance between hyponym and hypernym in the first case is much larger

Work with Boyd-Graber, Osherson, Shapire has addressed these shortcomings (too few connections, few x-POS pointers, no weights)

But only on a small scale

Scale up automatically?

Some recent improvements

Database Format

- Database moved from text-based files to SQLite
- Important advantage: allows addition of new semantic relations among words
- Tables are there, can be queried internally

Needed: web interface to allow queries from outside users

Editing

Needed: New editing interface for adding, removing, changing entries

- Could make it available so that people can customize and create their own WN versions

Is this a good idea?

- Potential problem: need *one* reference standard for bake-offs, training and evaluating systems

Recent Enhancements

New Part of Speech

Three kinds of phrases:

--Idioms/proverbs

--Internet acronyms

--light/support verb constructions

Idioms

A rolling stone gathers no moss

(people who always move will not prosper, will not form attachments)

beat a dead horse (discuss a question that is no longer relevant)

fall off the wagon (start drinking too much alcohol again after a period of moderation*)*

Lend/give someone a hand (help)

Idioms

Most are semantically non-compositional

Some are not plausible (have no possible literal reading)

Most idioms express complex concepts, cannot be paraphrased with a single word/verb

Hence do not fit readily into WordNet's current structure (as nouns or verbs)

Lexicon and grammar meet!

Idioms

Challenge for NLP: identification of idioms in text

Contrary to claims and traditional representation in resources: not fixed, invariable strings!

Hence, lexical entry treating idioms as long strings would miss many tokens

“beat a dead horse”

Web examples*:*

This dead horse was beaten into oblivion/weeks ago/senselessly (passive)

CBers love to beat dead horses (plural*)*

Not trying to beat a half-dead horse any more dead (lexical modification*)*

Other idioms have free variable that is filled differently according to context

(lend *someone* a hand)

Obligatory negation can be expressed differently

He does*n't* give a damn

He *never* gives a damn

Nobody gives a damn

Idioms

Many idioms have components to which speakers assign a meaning

Can be modified like simple words

Can be used in different POS, different contexts

Need to be identified, semantically interpreted by automatic systems

Simple solution

Add entire idiom to WN

Identify meaningful units within the idiom

Link it to corresponding synsets with lexemes that are used outside the idioms in free distribution

Simple solution

Beat a dead horse

beat - belabor%2:41:00:: to work at or to absurd length

dead - dead%3:00:00:noncurrent:00 no longer having force or relevance

horse - argument%1:10:00:: a discussion in which reasons are advanced for and against some proposition or proposal

Links between phrasal entry and synsets with simple lexeme limits interpretation of idiom components to idiomatic context

Allows for interpretation (not necessarily generation) of idioms where synonym is used

Beat/flog a dead horse

Limitations

Not all idioms are decomposable

Kick the bucket/buy the farm = die

Components do not refer

Other phrases: Internet language

Very frequent in certain text genres

Create synsets with synonyms, examples

{AATK, always_at_the_keyboard, "Sure, message me when you get back. You know me, I'm AATK”}

{GNE1, good_night_everyone, "I'm heading to bed now. GNE1”}

Polysemy:

{LOL, lots_of_love,..}

{LOL, laughing_out_loud,…}

Light verb constructions

Make an exception (~except)

Take a photograph (~photograph)

Have a drink (~drink)

LVCs are very picky about choice of verb

**take/*have an exception*

**make a photograph*

**take a drink*

Important for tasks requiring generation

By far, the most frequent us of verbs like *make, take, have, get* etc. is in LVCs (Hanks)

Like idioms, LVCs admit of (grammatical) variation

Made no/many exceptions

Pictures were taken

Etc.

Better to avoid entering entire constructions

LVCs

Separate entry for light verb use of *make, take*,..

Link nouns to corresponding light verb

Future POS addition?

Prepositions

Can arguably be structured with antonymy relations (D. Katunar)

Current work

Represent gradability

E.g., many adjectives are scalar: express different degrees of intensity of a shared attribute

Can be represented as points on a scale

Supplement current “dumbbell” structure, which does not differentiate among “similar” adjectives (Sheinman et al., 2013)

Scales

AdjScale method

Sheinman & Tokunaga (2009a, 2009b)

Induce lexical-semantic patterns from corpus

Patterns identify stronger/weaker adjectives

Apply patterns to pre-classified pairs of adjectives

Patterns

One example*:*

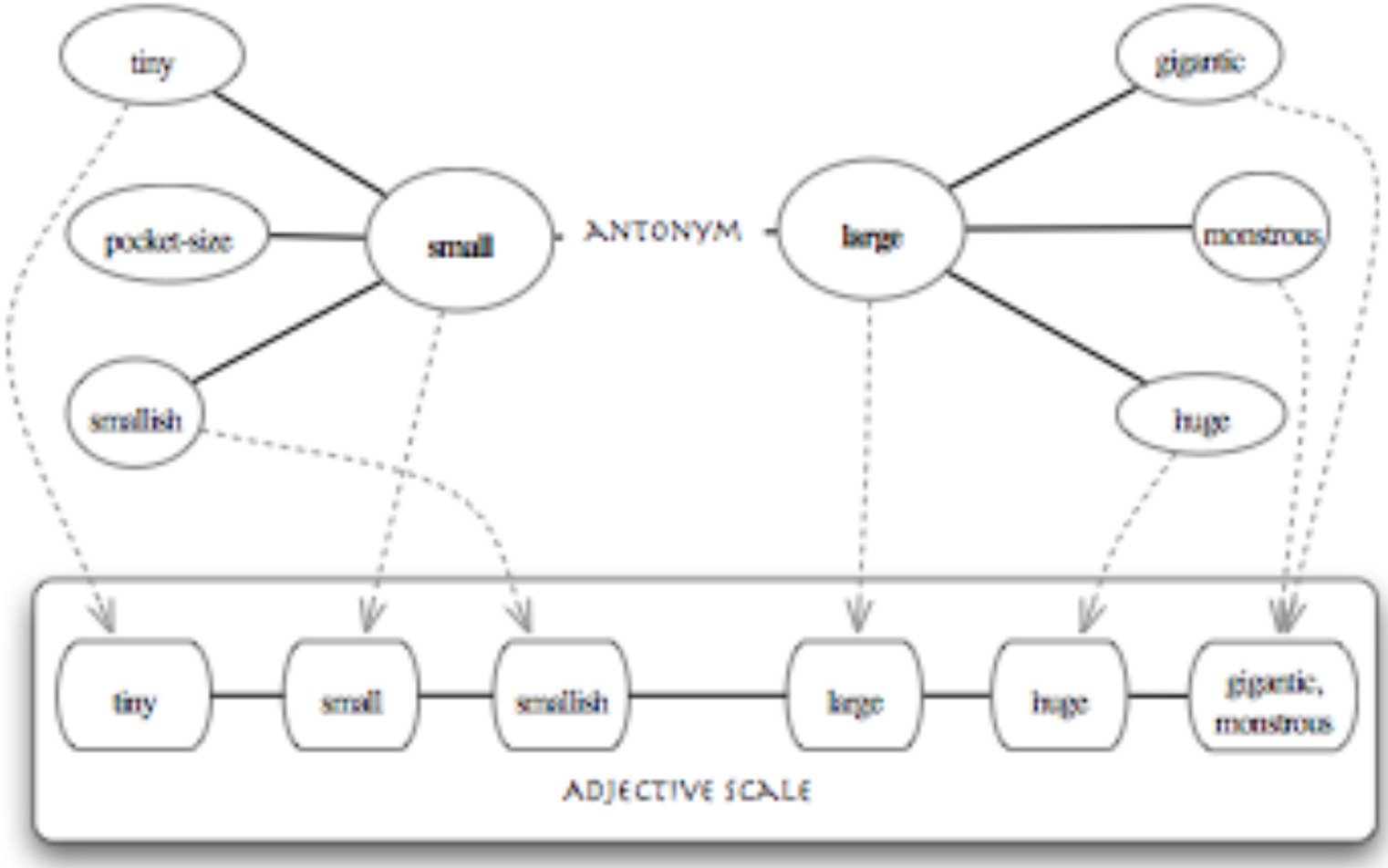
X if not Y

X even Y

X < Y

Represent placement of adjectives on a given scale in addition to the current “dumbbell”

AdjScales in WordNet



Current work

Noun compounds

shoe sale/summer sale/baby sale/fire sale

Very productive (unless idiomatic)

Noun compounds

Represent meaning of compound members as well as that of the entire compound

Allow for exploitation of WN structure to account for new compounds

Noun compounds

We are currently collecting judgments from "Turkers" to see

(a) How they interpret attested compounds (identify compounds whose members are related in the same way)

(b) How far Turkers go to interpret randomly generated unseen compounds

Thanks for any and all questions & comments!