Making WordNet Data (Retro)Fit for a Multilingual Dictionary

Problems with WordNet as Lexicography

Inadequate Definitions

- English definitions are indicative, not definitive
- One English meaning ascribed to multiple terms in multiple languages
- Mixed quality, e.g. *elevator car*: where passengers ride up and down
- Not accurate for all members, e.g. eat, feed: take in food; used of animals only
- Tautologies, e.g. visit: pay a brief visit
- Outright errors, e.g. *law practice*:

Translation Term Problems

- Incorrect glosses, often matching the wrong sense of homographs
- "Kitchen sink" collections of somewhat related terms
- Semantic drift correct glosses of English terms that are a bit off from synset definition
- Meanings mostly assumed from English definitions
- No information beyond lemmatic spelling – what is usually known is that one form of a term is a near equivalent of something in English

Solutions: WordNet as Seed Data

Crowd Review of Existing Data

- Competition for best-written definitions per concept/spelling entity
 - If PWN definition is good, it will win
- Validation/ rejection of bilingual matches by bilingual speakers
 - English <-> Wordnet X
 - Wordnet X <-> Wordnet Y



Data in Own Languages

- Definitions of terms in their own language
- *dedo* in Portuguese is different from the English elicitation term
- Own-language definitions can be translated to English or other languages
- Usage examples from own-language sources, e.g. blogs and tweets
- Additional lexical data
 - Inflected forms
 - Pronunciations

the practice of law



Erratic Coverage

- Missing senses, e.g. *light*: traffic signal
- Missing terms, e.g. *lightsaber*
- Spotty relationships, e.g. no pair between *boat* and *ship*, but a tie for *jalopy* and *bus*
- Random named entities, often figures of US or UK cultural significance from a by-gone era
- Cultural focus on US and UK concepts, e.g. *shortstop* (inherent to any English-based elicitation list)

Licenses and Limitations

- Many Wordnets are not copyrightavailable for further use or modification
- Many Wordnets are done/ dormant
- no changes planned or possible
- Active Wordnets are opaque about how public or professionals can contribute new data or change existing data

More Data for English Terms

- Descriptions of differences between synset members, e.g. snuggle vs. nestle
- User-curated usage examples, video links, images
- Geo-tagged pronunciations
- Geo-tagged usage sightings
- Lexicalized etymologies for historical and comparative linguistics
- BabelNet and other linked data
- Etc... (bringing in data from wheels that have already been invented,
- working with partners on new ways to enhance English data)

• Etc, etc, etc...

kamusi.org

More Terms, More Languages

- Compare WordNet to other sources to find omissions
- Terms from bilingual dictionaries can address cultural bias
- Candidates for WordNet inclusion could be selected based on popularity (search logs, number of languages that choose to translate)
- When Kamusi processes produce entries linked to WordNet in languages that do not already contain them, WordNets for those languages are created or expanded
 - Data merged from existing sources is directly matched to synset senses

Lumping and Splitting

Synsets: From Lumps to Lemmas

- All English members of a synset are ascribed the same meaning and usage examples
- Most other languages are ascribed the same meaning

200+ terms lumped together for *rag*:

Catalan *renyar*, *reprendre*

ar , reprendre ώνω , κατσαδιάζω , βάζω τις φωνές , επιπλήττω

- English $\begin{array}{l} rebuke_3 (\underline{\triangleright} \ \underline{\flat}), scold_3 (\underline{e} \ \underline{\flat}), scold_3 (\underline{b} \ \underline{b}), scold_3 (\underline{b} \ \underline{b} \ \underline{b}), scold_3 (\underline{b} \ \underline{b}), scold_3 (\underline{b} \ \underline{b}),$
- reproof (_) , jaw , call down , bawl out , remonstrate , chew out , chew up Basque demanda egin , agirakatu , agira egin , errieta egin , larderiatu , ateleka
- Bosque dernanda egin , agirakatu , agira egin , emeta egin , laraeriatu , ateleka Asum دادن , گله کردن از , زخم زبان زدن , تازیانه زدن , خطابه گفتن , سرزنش کردن , پرچانگی کردن از , ا
- Finnish kritisoida , ojentaa , moittia ankarasti , protestoida , moittia , sättiä , läksyttää , nuhdella , haukkua , torua , soimata French réprimander , reprendre , aronder , reprocher , enqueuler
- French réprimander , reprendre , gronde נתו-לו-עַל-הַרֹאשׁ , נַזֶּר , שַׁטַף-אֶת , גָּעַר
- נתן-לו-עַכ-הָראָש , שָׁן , אָשָן-אָד , שָׁשַן אָז , אָשָן Croatian osuđivati , koriti , kuditi , prekoravati , pokuditi , ukoriti , ukoravati , karati , prekoriti , grditi , osuditi , izgrditi Indonesian kutuk , memaki , merojuki , marah , menggebos , mencercai , mencela , menyegak , menegur , menengking , mencerca , mengkritik menggurui , mengaibkan , menghardik , mendamprat , menyemprot , menembis , cela , menyentil , menjentik , memarahi , memekis Italian rabuffare , redaravire , rampognare , rimproverare , riprendere , rabbuffare , rimbottare , saidare
- Italian rabuffare, redarguire, rampognare, rimproverare, riprendere, rabbuffare, rimbrottare, sgridare Japanese 叱飛ばす,怒る,叱する,叱咤+する,譴責+する,叱責+する,叱り付ける,怒りつける,叱責,叱付ける,譴責,叱り飛ばす,叱りとばす,叱咤, I る,叱りつける,怒り付ける
- Dutch aanwrijven , berispen , schelden , imputeren , op het matje roepen , nadragen , verwijten , aanrekenen , inwrijven , aantijgen Portuguese reprimir , censurar , condenar , repreender , incriminar , repreensão , reprovar Chinese
- Chinese (traditional) *訓言*
- Romanian *mustra , dojeni , moraliza , dăscăli* Slovak nadával bromžil wěltal brožil nanov
- Slovak nadávať, hromžiť, vyčítať, hrešiť, napomínať, Slovene ukoriti, pograjati, hruliti, okregati, pokregati, k
- Slovene ukoriti, pograjati, hruliti, okregati, pokregati, kregati, ošteti, karati, grajati, ozmerjati, pokarati Spanish discutir, amonestar, poner verde, reconvenir, reprochar, reprender a gritos, censurar, regañar, rechazar, increpar, repr tener unas palabras, echar una reprimenda, vociferar, criticar, llamar la stensión, daruna paliza
- Swedish tillrättavisa
- Thai ด่ำ , ด่า , ดูด่า , ดุ , ผ
- kutuk , memaki , merajuki , marah , menggebos , mencercai , mencela , menyegak , menegur , memberi aib , menengking , men Malaysian , mengkritik , menghamun , mengaibkan , hentam , berkhutbah , memekis , mendamprat , menyemprot , menembis , cela , men menjentik , memarahi

OMW: One Big Concept, Separated per language

PWN: Vertical Focus

Linked

- Synsets conceived as clusters for ontological relationships
- Meronyms/ holonyms
- Hypernyms/ hyponyms, etc
- Objective was understanding about psychology, not linguistics
- "Definitions" intended for group identification, not lexical precision
- Translations attached post hoc, not by design
- Naïve bilingual model of raw horizontal equivalence

Kamusi: Separate Concepts,

Kamusl.org

Similarity vs. Sameness

Synsets Are Not (necessarily) Synonyms

- Within a language, subtle differences exist for important reasons, e.g. nuance among {*approximate, estimate, gauge, guess, judge*}
- Larger English synsets inspire very large translation synsets
- Translation introduces semantic driftespecially notable in larger synsets
- 1/(#English terms)² odds that a term in one translation language will equate with a given term in another language

Synsets *Are* Topical Relationships

- Synset definition describes the semantic relationship
- Synset relations perform like horizontal ontologies
 - Members share a certain property (topic) but independent essences
- Members should generally have independent definitions and examples in addition to synset topical guides
- Imputed translations among languages should be seen as only topically indicative until human verified

kamusi.org

- Terms from one language Wordnet can be seen together with:
 - "Same" idea terms in that language
 - English synset definition
 - Matching clusters from other languages
- Each term in one cluster is parallel to each term in all other clusters
- 11 Japanese terms = 20 Arabic terms:



GWC 2016: Global WordNet Conference 28 January 2016 - Bucharest, Romania

- Synset members within a language are separate entities with own specific meanings
- Each pair across languages is an independent relationship to be diagrammed and validated
- ~100,000 synsets = ~10,000,000 pairs





Substitutes

- 1. Parallel basically the same idea
- 2. Similar substantial overlap, but noteworthy differences
- Explanatory invented term in one language to fill lexical gap for a concept indigenous to another
- A term can be *parallel* to one synset/ translation set member but *similar* or *explanatory* to another (programming complexity)
- Differences can be elaborated in definition-like field

Degree of *separation* between terms:

- WordNet data has (often faulty) presumption of 1° manual validity for
 - Same-language synset members
- Links to English

Joints

- Relationships via intermediate languages are mapped transitively, i.e.
 A <-> B <-> C <-> D
 - A and D are 3rd generation links (3°)
- For GWN data, "B" is always English, most computed pairs are 2°
- Evaluating predicted bilingual joints is future work via crowd systems





Contact: martin.benjamin@epfl.ch