Lexical Perspective on Wordnet to Wordnet Mapping

Ewa Rudnicka,[♡] Francis Bond,[♠]
Łukasz Grabowski,[♣] Maciej Piasecki[♡]
Tadeusz Piotrowski[♢]

[♡]Wrocław University of Technology, Poland
[♠]Nanyang Technological University, Singapore
[♣]University of Opole, Poland
[♢]University of Wrocław, Poland
{ewa.rudnicka,maciej.piasecki}@pwr.edu.pl, bond@ieee.org,
lukasz@uni.opole.pl, tadeusz.piotrowski@uwr.edu.pl

GWC-2018: 2018-01-10

Outline

Finer linking

- Many wordnets link synsets through PWN
- Increasingly, many are linked through CILI
- For some pairs (such as Polish-English) there is a richer linking, covering synonymy, hyponymy, meronymy, register, ...
- But all of these links are at the synset level, and many synsets have multiple lexical units (LUs) — however the strength of linking may not be the same for all LUs

Motivation

During the plWordNet and Princeton WordNet synset mapping we observed the potential for finer sense mappings:

- {***złoto***$_{n:3}$, ***Au***$_{n:1}$}^{PL} `I-syn` {***gold***$_{n:3}$, ***Au***$_{n:1}$, ***atomic number 79***$_{n:1}$}^{EN}
- ***złoto***$_{n:3}{}^{PL}$ and ***gold***$_{n:3}{}^{EN}$
- ***Au***$_{n:1}{}^{PL}$ and ***Au***$_{n:1}{}^{EN}$
- Closer to bilingual lexicography

Goals

- We want to link at the LU level
- We distinguish **strong**, **regular** and **weak** equivalence links
- We created a procedure for deciding the strength
- We are now mapping LUs (pl-en), nouns first

Such finer sense mapping will be beneficial for translators and of great use for bilingual WSD

Prerequisites

- sense mapping builds on synset mapping
- sense links considered for pairs of Polish-English LUs from synsets linked by:
  + `I-synonymy`
  + `I-partial synonymy`
  + `I-hyponymy`
- nouns mapped, other POS being mapped

Equivalence Features

Our goal is to operationalize the equivalence so that we can reliably determine its strength using various features.

- **Formal**: number, countability and gender, ...
- **Semantic** and **Pragmatic**: sense, lexicalisation (of concepts), register, collocations, co-text and context
- **Translatability**: based on dictionary listing and translation equivalences extracted from the Polish-English parallel corpus: *Paralela*

Formal features

- **part of speech** (given)
- **gender** (if lexicalised)
- **number** (except for pluralia and singularia tantum)
- **countability** (except for mass/count contrasts in lexicalisation)

Semantic features

- **sense** (going beyond truth-conditions)
- **lexicalisation of concepts** (comparing denotations)
- **register**
- **collocations** (fixed phrases, from dictionaries)
- **co-text** (immediate sentence environment, from parallel corpus)
- **context** (situational and world knowledge)

Translatability

- **dictionary listing**
  - \+ frequency of occurrence in multiple dictionaries
  - \+ rank of the translated term
- **translation probabilities**
  - \+ extracted from the Polish-English parallel corpus *Paralela*

Equivalence types

These are used to link individual lexical units (senses) between the two wordnets.

- **Strong**
- **Regular**
- **Weak** (implied)

Strong Equivalence features

- identity in sense
- similarity in lexicalisation of concepts
- compatibility in register
- a shared set of typical co-texts
- dictionary listing (as the first equivalent)
- bidirectionality (but not uniqueness) of translation
- frequent parallel corpora hits, preferably

Strong Equivalence - examples

- ***drzwi***$_{n:1}$ `I-syn` ***door***$_{n:1}$
- ***grzmot***$_{n:1}$ `I-syn` ***thunder***$_{n:2}$

- ***narzeczona***$_{n:1}$ `I-syn` ***fiancee***$_{n:1}$
- ***centrala***$_{n:2}$ `I-syn` ***headquarters***$_{n:1}$
- ***gruba ryba***$_{n:1}$ `I-partial-syn` ***big fish***$_{n:1}$

- ***okulary***$_{n:1}{}^{PL}$`I-syn` ***glasses***$_{n:3}{}^{EN}$
  For all, identity in sense and register, frequent (often first) dictionary listing, many parallel corpora hits

Regular equivalence features

- largely similar in sense
- compatibility in register
- dictionary listing
- bidirectionality of translation
- a similar set of typical co-texts
- some parallel corpora hits, preferably
- some differences in lexicalisation of concepts are allowed

Regular equivalence - examples

- ***zabytek***$_{n:1}$ `I-partial-syn` ***monument***$_{n:2}$
  Lexical gap (on the English side)
- ***narzeczona***$_{n:1}$ `I-syn` ***bride-to-be***$_{n:1}$
  Additional (temporal) sense specification on the English side; few parallel corpora hits
- ***centrala***$_{n:2}$ `I-syn` ***central office***$_{n:1}$
  Few parallel corpora hits for this pair

Weak equivalence

- All other pairs of LUs from plWordNet and Princeton WordNet synsets linked by `I-synonymy`, `I-partial synonymy` and `I-hypernymy` that do not meet the criteria for strong or regular equivalence
- can be automatically derived from the synset-level links
- often culture specific concepts absent from the second language (cultural gaps) and linked via `I-hyponymy` relation

Weak equivalence - examples

- ***centrala***$_{n:2}$ - ***main office***$_{n:1}$, ***home office***$_{n:2}$, ***home base***$_{n:2}$
  very few or no Paralela hits
- {***stachanowiec***$_{n:1}$, ***przodownik pracy***$_{n:1}$}
  `I-hypo` {***toiler***$_{n:1}$}
  Polish culture specific term, with no direct equivalent: “model worker who greatly exceeds the quota”

Linking procedure

- check features
  - formal
  - semantic
  - pragmatic
- check wordnet info first (sense and synset relations, glosses, register info, examples)
- consult external resources (dictionaries, parallel corpora)
- then assign proper equivalence type: **strong, regular, weak**

Current and Future Work

- the procedure is being verified on a random sample of lexical unit pairs
- extracted from synsets linked by `I-synonymy`, `I-partial synonymy`, `I-hyponymy`
- proportionally for each relation and link type (1-1, 1-many, many-1, many-many)
- extracted 100 random sets with 10 instances for each of the 12 classes: one checked so far

Challenges for estimating translation probability

- polysemous lemmas
- no sense tagged bilingual corpora

$\Rightarrow$ creates difficulty in estimating the number of hits of a specific sense

$\Rightarrow$ manual work and interpretation required

Conclusions

- created a method for finer linking of senses (LUs)
- of great potential for (manual and automatic) translation as well as (bilingual) word sense disambiguation
- adjustable for other language pairs and grammatical categories
- possible to partly automate
  generate prompts for efficient annotation

Acknowledgement

This work was supported by the National Science Centre in Poland under the agreement No. UMO-2015-/18/M/HS2/00100.

Today is the excursion

- 13:00 Bus leaves NEC to Bollywood Vegies
- 15:00 Bus leaves BV to Sungei Buloh (Visitor Center)
  - if too wet we may send one bus back to NEC
- 18:30 Bus leaves Sungei Buloh (Wetland Center) to NEC
- Please dress comfortably
  - comfortable shoes (and hat — mainly in shade)
  - rain-friendly clothes
  - water bottle
- You are free to leave mid-way
  - we will assume you have done so if you are not on the bus