Eric Kafe

Sense keys

Mapping

Results

Discussion

Summary

References

Mapping Wordnets on the fly

with permanent sense keys

Eric Kafe
kafe@megadoc.net

MegaDoc

January 2023
EHU/UPV, San Sebastian
Spain

Sense Keys

- **Sense keys represent one word sense**:
  Permanent since PWN 1.5SC (1994)
  OEWN since Edition 2021
- **Examples**:
  sequoia%1:20:00::
  obtrusive%5:00:00:noticeable:00
- **Sense key encoding** (PWN Manual , 2010, Senseidx):
  lemma%ss_type:lex_filenum:lex_id:head_word:head_id

Almost all wordnets are mapped at the synset level
But the sense keys can reveal splits and merges

Merge

Eric Kafe

Sense keys
Mapping
Results
Discussion
Summary
References

Sense Key	$PWN_{3.0}$	$CILI_{3.0}$	$CILI_{3.1}$	$PWN_{3.1}$
baseball%1:04:00::	00471613-n:	i37881	i37882	00472688-n
baseball_game%1:04:00::	00471613-n:	i37881	i37882	00472688-n
ball%1:04:01::	00474568-n	i37882	i37882	00472688-n

Merges concern whole synsets, so should not be a problem
But ILI-to-offset mappings lose merged synsets (with *Wn*):

- pwn30.synsets(ili="i37881")[0].translate("oewn")
  → []
- oewn.synsets(ili="i37882")[0].lemmas()
  → [*ball*, *baseball*, *baseball_game*]
- oewn.synsets(ili="i37882")[0].translate("omw-en")[0].lemmas()
  → [*ball*]

Would need a mapping between ILI versions:
$i37881_{3.0} \leftrightarrow i37882_{3.1}$

Split and Merge

French	$PWN_{3.0}$	$CILI_{3.0}$	Sense Key	$CILI_{3.1}$	$PWN_{3.1}$
Aides	09570298-n	i86957	aides%1:18:00::	i86957	09593427-n
∄	09570298-n	i86957	aidoneus%1:18:00::	i86957	09593427-n
Hadès	09570298-n	i86957	hades%1:18:00::	i86957	09593427-n
Pluton	09570298-n	i86957	pluto%1:18:00::	i86958	09593643-n
Dis	09570522-n	i86958	dis%1:18:00::	i86958	09593643-n
∄	09570522-n	i86958	orcus%1:18:00::	i86958	09593643-n
∄	∄	∄	dis_pater%1:18:00::	i86958	09593643-n

- Pluto moved from Greek gods to Roman gods
- CILI mapping works only for English:
  For ex. French *Pluton* doesn't move
- Would need a mapping for particular senses:
  $Pluton_{French:3.0:i86957} \leftrightarrow Pluton_{French:3.1:i86958}$

Algorithm

In NLTK since v. 3.7 (2021):
Load OMW with any PWN or OEWN version

$\text{SENSE\_INDEX}_{source} \leftarrow \{\forall \text{ sense} \in source : \text{sense}_{key} \rightarrow \text{synset\_id}_{source}\}$
$\text{SENSE\_INDEX}_{target} \leftarrow \{\forall \text{ sense} \in target : \text{sense}_{key} \rightarrow \text{synset\_id}_{target}\}$
$\text{MAP\_TO\_MANY} \leftarrow \{\forall \text{ synset\_id}_{source} \in \text{values}(\text{SENSE\_INDEX}_{source}) : \text{synset\_id}_{source} \rightarrow \emptyset\}$
for $\text{sense}_{key} \in \text{SENSE\_INDEX}_{source} \cap \text{SENSE\_INDEX}_{target}$ **do**
    $\text{MAP\_TO\_MANY}[\text{synset\_id}_{source}].\text{append}(\text{synset\_id}_{target})$
end for
$\text{MAP\_TO\_ONE} \leftarrow \{\forall \text{ synset\_id}_{source} \in \text{MAP\_TO\_MANY} : \text{synset\_id}_{source} \rightarrow \text{argmax}(\text{count}(\text{synset\_id}_{target}))\}$
return MAP_TO_ONE

- Map to most frequent target synset
- Runs in linear time
- Works with any synset id. (f. ex. offsets or ILI)

No Splits

Sense Key	$\text{PWN}_{3.0}$	$\text{CILI}_{3.0}$	$\text{CILI}_{3.1}$	$\text{PWN}_{3.1}$
aides%1:18:00::	09570298-n	i86957	i86957	09593427-n
aidoneus%1:18:00::	09570298-n	i86957	i86957	09593427-n
hades%1:18:00::	09570298-n	i86957	i86957	09593427-n
pluto%1:18:00::	09570298-n	i86957	i86957	09593427-n
dis%1:18:00::	09570522-n	i86958	i86958	09593643-n
orcus%1:18:00::	09570522-n	i86958	i86958	09593643-n
dis_pater%1:18:00::	∄	∄	i86958	09593643-n

- Synset mappings can't handle splits
- Synsets are not split (Pluto doesn't move)
- All senses accurate, except one false positive (*fp*)

Results

Table: The 15 biggest Wordnets in OMW 1.4, mapped to OEWN 2021, using synset offsets vs. CILI 1.0

	Synsets	**Map**$_{Offset}$			**Map**$_{CILI}$		
Language	*PWN 3.0*	*OEWN 2021*	*Lost*	*%*	*OEWN 2021*	*Lost*	*%*
English	117659	117454	**205**	0.17	117427	232	0.20
Finnish	116763	116562	**201**	0.17	116535	228	0.20
Thai	73350	73240	**110**	0.15	73223	127	0.17
French	59091	59015	**76**	0.13	59005	86	0.15
Japanese	57184	57086	**98**	0.17	57080	104	0.18
Romanian	56026	55941	**85**	0.15	55931	95	0.17
Catalan	45826	45773	**53**	0.12	45769	57	0.12
Portuguese	43895	43844	**51**	0.12	43840	55	0.13
Slovenian	42583	42520	**63**	0.15	42513	70	0.16
Mandarin Chinese	42300	42249	**51**	0.12	42240	60	0.14
Spanish	38512	38431	**81**	0.21	38418	94	0.24
Indonesian	38085	38018	**67**	0.18	38011	74	0.19
Standard Malay	36911	36843	**68**	0.18	36836	75	0.20
Italian	35001	34964	**37**	0.11	34960	41	0.12
Polish	33826	33798	**28**	0.08	33794	32	0.09
Average	55800.80	55715.87	**84.93**	0.15	55705.47	95.33	0.16

- Synset offsets always better than CILI 1.0
- 99.85% true positives in average

Performance between $PWN_{3.0}$ and $OEWN_{2021}$

Synsets	**Mapped**	**Not Mapped**
True	$PWN_{3.0} \cap OEWN_{2021}$	$\emptyset$
	$tp = 117454$	$tn = 0$
False	*Splits*	$\complement^{PWN_{3.0}}_{OEWN_{2021}}$
	$fp = 44$	$fn = 205$

Almost perfect performance results:

$$precision = \frac{tp}{tp + fp} = 0.9996 \quad (1)$$

$$recall = \frac{tp}{tp + fn} = 0.9983 \quad (2)$$

$$f1 = \frac{2 * precision * recall}{precision + recall} = 0.9989 \quad (3)$$

Discussion

Strengths :

- map in linear time
- almost perfect accuracy
- no problem with merges

Limitations :

- needs stable sense keys
- cannot handle splits

Opportunities :

- very few problems, easy manual review
- link multilingual wordnets directly to OEWN, instead of going through PWN 3.0

Changed Sense Keys

Eric Kafe

Sense keys
Mapping
Results
Discussion
Summary
References

Sense Key	PWN 3.0	OEWN 2021
sequoia%1:20:00::	*either of two huge coniferous California trees that reach a height of 300 feet; sometimes placed in the Taxodiaceae*	∄
sequoia%1:20:01::	∄	*either of two huge coniferous California trees that reach a height of 300 feet; sometimes placed in the Taxodiaceae*

More Changed Sense Keys

Sense Key	PWN 3.0	OEWN 2021
stub_out%2:30:00::	*extinguish by crushing*	∄
stub_out%2:35:01::	∄	*extinguish by crushing*
obtrusive%3:00:00::	*undesirably noticeable*	∄
obtrusive%5:00:00-:noticeable:00	∄	*undesirably noticeable*
newfangled%5:00:00-:original:00	*(of a new kind or fashion) gratuitously new*	∄
newfangled%5:00:00-:new:00	∄	*(of a new kind or fashion) gratuitously new*

Summary

- Wordnets need mapping, even with ILI
- Baseline algorithm extremely fast and accurate
- Only review a few splits and losses
- Perspective: link directly to OEWN

Software

- NLTK https://www.nltk.org
  https://github.com/nltk/nltk
- Wn (Goodman and Bond, 2021)
  https://github.com/goodmami/wn

Databases

- OEWN https:
  //github.com/globalwordnet/english-wordnet
- *wndb* https://github.com/x-englishwordnet/wndb
- Collaborative Interlingual index (CILI)
  https://github.com/globalwordnet/cili
- Sense Key Index (SKI) https://github.com/ekaf/ski

References I

Francis Bond, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Adam Pease, and Piek Vossen. 2014. A multilingual lexico-semantic database and ontology. In *Towards the Multilingual Semantic Web*, pages 243–258. Springer.

Francis Bond, Luis Morgado da Costa, Michael Wayne Goodman, John Philip McCrae, and Ahti Lohk. 2020. Some issues with building a multilingual Wordnet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3189–3197, Marseille, France. European Language Resources Association.

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016. CILI: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57, Bucharest, Romania. Global Wordnet Association.

References II

Christiane Fellbaum. 1998. *WordNet, An Electronic Lexical Database*. MIT Press, Cambridge.

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The wn python library for wordnets. In *Proceedings of the 11th Global Wordnet Conference*, pages 100–107, University of South Africa (UNISA). Global Wordnet Association.

Eric Kafe. 2018. Persistent semantic identity in wordnet. *Cognitive Studies | Études cognitives*, 18.

References III

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France. The European Language Resources Association (ELRA).

Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual GlobalWordnet grid. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 424–431, Bucharest, Romania. Global Wordnet Association.

WordNet-team. 2010. Wordnet 3.0 reference manual. In *WordNet Documentation*. Princeton University, `https://wordnet.princeton.edu/documentation`.