GeoNames Wordnet (gnwn): extracting wordnets from GeoNames

Francis Bond
Nanyang Technological University
bond@ieee.org

Arthur Bond
United World College of Southeast Asia
artcbond@gmail.com

Abstract

This paper introduces a new multilingual lexicon of geographical place names. The names are based on (and linked to) the GeoNames collection. Each location is treated as a new synset, which is linked by `instance_hypernym` to a small set of supertypes. These supertypes are linked to the collaborative interlingual index, based on mappings from GeoDomainWordnet. If a location is already in the interlingual index, then it is also linked to the entry, using mappings from the Geo-Wordnet. Finally, if GeoNames places the location in a larger location, this is linked using the `mero_location` link. Wordnets can be built for any language in GeoNames, we give results for those wordnets in the Open Multilingual Wordnet. We discuss how it is mapped and the characteristics of the extracted wordnets.

1 Introduction

The aim of this paper is to create a large multilingual lexicon of place names, through the use of the vast open source database GeoNames.[1]

Wordnets generally contain open-class words, with only a few proper names. Some names need to be there, as they are derivationally related to open-class words (such as ***Vratislavian*** "a native of Wrocław"). However the general trend is to leave proper names out, and instead link them through other specialist lexicons (Vossen et al., 2016). The goal is for specialists on names to curate names, with wordnets only having to maintain a smaller collection of links.

Another popular approach is to merge completely, into a vast combined resources such as YAGO (Suchanek et al., 2007) or BabelNet (Navigli and Ponzetto, 2012). This can cause problems when a subsidiary resource updates: propagating the corrections into the merged resource is an unsolved problem. For example, the version of GeoNames used in BabelNet 4.0 is from April 2015, a four year difference.[2]

Apart from general merging, there are two main resources made from merging Wordnet with GeoNames. The first, Geo-Wordnet (Buscaldi and Rosso, 2008) links locations in Princeton Wordnet (PWN: Fellbaum, 1998) to GeoNames (we will call this GWN-link). The second, GeoWordnet (Giunchiglia et al., 2010) links the supertypes in GeoNames to PWN synsets (we will call this GWN-super). These are complimentary mappings, but as far as we know, no one has combined them. GeoWordNetDomains (Frontini et al., 2016) further refines the mappings from GeoWordnet and adds some more internal structure. Both GeoWordnet and GeoWordNetDomains link the synsets to English and Italian (the Multiwordnet (Pianta et al., 2002) and Italwordnet (Toral et al., 2010) respectively), but do not consider other languages. All the resources are described in more detail in Section 2.

In this paper, we introduce a method for creating lexicons of placenames for any language in GeoNames: the Geoname Wordnet. Each location is treated as a new synset, which is linked by `instance_hypernym`[3] to a small set of supertypes, linked to the collaborative interlingual index, based on mappings from GeoDomainWordnet. If a location is already in the interlingual index, then it is also linked to the entry, using mappings from the Geo-Wordnet. Finally, we add some additional structure, if GeoNames places the location in a larger location, this is linked using the `mero_location` link. This is described in Section 3. The code to create the lexicons is available at `https://github.com/fcbond/geonames-wordnet`.

We present some statistics of the resulting word-

[1] https://www.GeoNames.org/

[2] `https://babelnet.org/about` accessed on 2019-05-20.

[3] Links are linked to their definition by the Global Wordnet Association Working Group.

net, along with some examples, in Section 4, and finish with some conclusions and ideas for future work in Section 5.

2 Resources

We give descriptions of the major resources we use here. All of GeoNames' information is downloadable and can be found on their website.

2.1 GeoNames

GeoNames is a geographical database, under a Creative Commons license. It boasts over 25 million geographical names, which ultimately are categorised into one of nine categories, and then into one of 645 sub-categories. GeoNames' search engine allows you to search for the location and its accompanying information. Editing these locations are then open to the public, for anyone to correct any mistakes, or perhaps add a new location.

The Wroclaw Panorama for example, is an instance of Geonames' richness of information and features. We show the online result in Figure 2 and a subset of the information available in Figure 1.[4] It immediately comes up with a top 3 items list. The first item was the correct location. It details the location name, type of physical place, which in this case is categorised into the overarching theme of Spots, Buildings or Farms, (see Table 1), as well as the sub-category Monuments (S.MNMT). It is then classified into five potential administrative divisions: Panorama Racławicka is in Wroclaw City, which is in Wroclaw County, which is in Lower Silesia. Lower Silesia is in Poland, but the country is not part of an administrative division, it is simply a country, rendering the location Panorama Racławicka with just three administrative classes.

For the sake of this paper, information regarding coordinates, elevation, timezone, and modification dates of the data points, which GeoNames also offers, have not been used.

The alternative names shown in Figure 1 include a wide variety of languages; how many are featured for each entry has a great deal of variability. Many of the names (almost 40%) are not associated with a language. In the above instance Panorama Racławicka is in Polish, but GeoNames does not indicate that that is the case. In this case, an extra step needs to be done to deduce the language and it is not a trivial task. Names can also be marked with features:

[4] GeoName ids are linked to the GeoNames website.

ID	GN: 11839964
name	Panorama Racławicka
feature	S.MNMT "Monument"
lat-lon	N 51°06′36″ E 17°02′40″
country	GN: 798544 PL "Poland"
adm1	GN: 3337492 "Lower Silesia"
adm2	GN: 7530801 "Wroclaw County"
adm3	GN: 7531292 "Wrocław"
alt	fr Panorama de Racławice ja パノラマ・ラツワヴィツカ en Racławice Panorama link …/wiki/Racławice_Panorama

Figure 1: GeoNames entry: Panorama Racławicka (links resolved and annotated with labels)

Class	Sub	Description
A	24	country, state, region, …
H	137	stream, lake, …
L	48	parks,area, …
P	18	city, village, …
R	22	road, railroad , …
S	253	spot, building, farm, …
T	98	mountain, hill, rock, …
U	62	undersea …
V	18	forest,heath, …

Table 1: Top Level Feature Classes

PreferredName	an official/preferred name
ShortName	a short name ***California*** for ***State of California***
Colloquial	a colloquial or slang term ***Big Apple*** for ***New York***
Historic	the was used in the past ***Bombay*** for ***Mumbai***

GeoNames also includes non-language data in these fields: external links, mainly to wikipedias and dbpedias; postcodes, airport codes and more. We currently do not use them, but they are a potential source for more translations.

The GeoNames database is built from official public sources, the quality of which may vary. Through a wiki interface, users are invited to manually edit and improve the database by adding or correcting names, move existing features, add new features, etc. Ahlers (2013) showed that there are many inaccuracies, especially in the granularity of coordinates (e.g., due to truncation and low-resolution geocoding in some cases), as well as

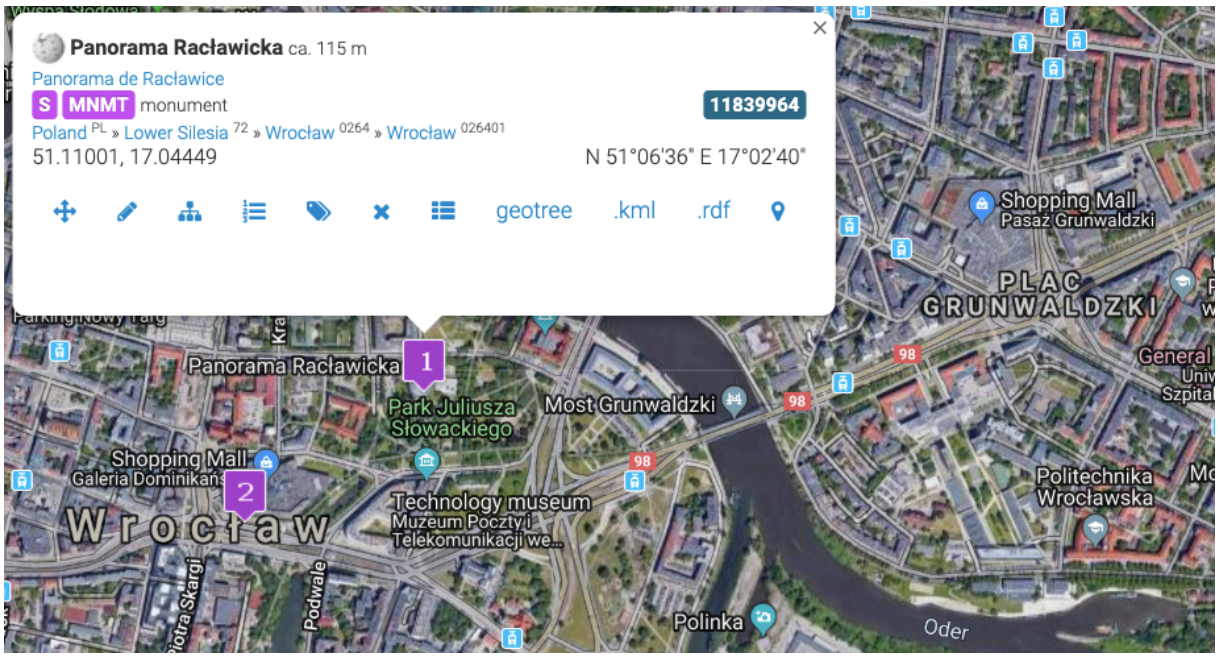

Figure 2: GeoNames entry for the Panorama Racławicka

wrong feature codes, near-identical places, and the placement of places outside their designated countries. However, he also pointed out that there was no other freely available resource with more accuracy.

2.2 Geo-WordNet (GWN-link)

This resource links locations in PWN 3.0 to GeoNames, for example PWN 08997487-n[5] ***Republic of Singapore*** is linked to GeoNames GN: 1880251. There are 1,964 entries so linked.

In the original paper (Buscaldi and Rosso, 2008), locations in PWN 2.0 were linked to Wikipedia to get coordinates. In Geo-WordNet 3.0,[6] the source of geographical data was GeoNames, and the mapping was to PWN 3.0.

This mapping is very useful, but does not extend the vocabulary of PWN, it merely adds more data (links to GeoNames, latitude and longitude).

2.3 GeoWordNet (GWN-super)

GeoWordNet takes a different approach and links the top level categories of GeoNames to wordnet synsets (Giunchiglia et al., 2010). The GeoNames entries are then treated as synsets. This gives an integration of WordNet, GeoNames and the Italian part of MultiWordNet (Pianta et al., 2002).

The GeoWordNet Public Dataset[7] is an impressive collection and contains 3,698,238 entities, 3,698,237 part-of relations between entities, 334 concepts, 182 relations between concepts, 3,698,238 relations between instances and concepts, and 13,562 (English and Italian) alternative entity names.

However, in the data made publicly available, there is no link to the original GeoNames data (so, for example, you cannot look up latitude and longitude). Further, locations that already exist in Multiwordnet are not linked, so for ***Republic of Poland*** a new node is created, and it will appear to be ambiguous: there will be two concepts, one from the wordnets and from GeoNames (although this ambiguity is spurious).

2.4 GeoDomainWordNet

GeoDomainWordNet aims to link the resources more loosely (Frontini et al., 2016). By treating GeoNames as linked open data they make sure that the full up-to-date version of GeoNames will be linked to. They took the GeoNames upper categories and linked them to synsets, either directly, or with a hyponym or meronym relation. For example ***section of lake*** is not a category in wordnet, but can be thought of as a meronym of ***lake***

[5] Wordnet synsets are linked to the Open Multilingual Wordnet.

[6] `http://timm.ujaen.es/recursos/geo-wordnet-3-0/`: note we could not find the data here, but got it from Bogdan Ivanov's NLTK wordnet extensions `https://github.com/bogdan-ivanov/wnext`

[7] Retrieved from here: `http://diversicon-kb.eu/dataset/geowordnet`

PWN 09452395-n. In this way, all categories are connected.

This approach has the same drawback as with GeoWordNet — if a location appears in both GeoNames and PWN (or Italwordnet: Toral et al., 2010), then it it will appear to be ambiguous.

3 GeoNames Wordnet (gnwn)

Our goal is the same as Geo(-)(Domain)Wordnet: to link the data in GeoNames to wordnets. We take advantage of the work they have done already to make what we believe is a better integration in the following ways.

- We make synsets for the feature codes and link them to the Collaborative Interlingual Index (CILI) using the mappings from GeoDomainWordNet, with additional mappings for newly added codes (§ 3.1)
  - the synset names encode the feature code names, so it is easy to retrieve them
- We have a script to build a wordnet for any language in GeoNames in the GWA LMF format (Vossen et al., 2016)
  - synsets are linked as instances of the feature codes
  - synset names encode the GeoName ids, so it is easy to retrieve them
  - synsets that are already in Princeton Wordnet, and thus in the CILI are linked (using the mapping from Geo-Wordnet)
  - GeoName admin codes are linked as location-meronyms — this is completely novel.

The code to and revised mappings used to create the lexicons is available at `https://github.com/fcbond/geonames-wordnet`. Entries that are not already in CILI will not be added — as GeoNames already curates and manages the GeoNames IDs, it is better to not duplicate effort. Instead we will encourage wordnet users to add places to GeoNames.

3.1 The Feature Codes

Figure 3 shows an example of a feature code mapping. There are 645 of these, with 18 newly added.

Figure 4 shows a new entry. In this case there is no corresponding entry in the ILI, so instead the synset is linked to another synset ***power station*** (gnwn-S.PS) which is linked to the ILI (`i57632`).

Synset	gnwn-S.MNMT
Definition	"a commemorative structure or statue"
ILI	`i82178`

Figure 3: Synset for Monument (S.MNMT)

Synset	gnwn-S.PSN
Definition	"nuclear power station"
`hypernym`	gnwn-S.PS

Figure 4: Synset for Nuclear Power Station (S.PSN)

In this way, all supertypes are linked to some existing entry in the wordnets.

3.2 Locations

In this section we give two examples of locations. Each location has a synset (Figures 5 and 6). The synsets each have an `instance_hypernym` and a note giving the GeoNames name (to make it easier to debug the wordnet). The Panorama Racławice synset also has a `mero_location` to the City of Wroclaw.

The Panorama Racławice synset is linked to translations in three languages. We show two of them here: English in Figure 7 and French in Figure 8. A single location can have multiple names (in Wordnets for different languages) or even multiple names in the same language. GeoNames marks preferred names: if a name is so marked, we take it as a vote of higher usage and add one to the frequency count of '1', so that it will be sorted first. Other information about names (short, colloquial, historic) could be encoded as meta information on the variant, this is left for future work.

4 Results

We show the sizes of the wordnets created for all the curated languages in the Open Multilingual Wordnet 2.0 (Bond and Foster, 2013), along with the lemma for the continent of Asia (GN: `6255147`) in Table 2.

Synset	gnwn-11822362
`instance_hypernym`	gnwn-S.MNMT
`mero_location`	gnwn-7531292
note	Panorama Racławice

Figure 5: Synset for Panorama Racławice

Synset	gnwn-798544
`instance_hypernym`	gnwn-A.PCLI
ili	`i83894`
note	Republic of Poland

Figure 6: Synset for Republic of Poland

Lemma	w1
lang	en
writtenForm	Racławice Panorama
partOfSpeech	n
	Sense gnwn-11839964-w1

Figure 7: English Lemma for Racławice Panorama

As can be seen, not all languages are equally well represented. Some languages include transliterations and this thus inflates the number of lemmas. For many languages, the average ambiguity is high.

There are 3,649,522 synsets, 3,129,147 lemmas and 4,587,108 senses, a substantial addition of knowledge.

The last column of Table 2 shows which place names are most common for each of the 40 languages (if there are fewer than 4 or more than one). Some of these are very common names: ***Stormyra*** is well known as the most common place name in Norway, **本町** *hon-machi* "this town" is a common placename in Japanese and *Kampung Baharu* and 新村 *xincun* "new village" are common names in Malay and Chinese. However some results are surprising: Some equivalent of "Washington County" is the most common placename for Estonian, Basque, Italian, Polish and Romanian! This is because many states in the US have a Washington County, and they have all been diligently translated. A more interesting query may have been: what is the most popular placename in a given country, rather than language.

Lemma	w1919
lang	fr
writtenForm	Panorama de Racławice
partOfSpeech	n
	Sense gnwn-11839964-w1919

Figure 8: French Lemma for Racławice Panorama

5 Conclusion and Future Work

We have created a large collection of lexicons of placenames: the Geoname Wordnet. Looking at 40 languages we had over 3.6 million locations with over 4.6 million senses. We can create lexicons for many more languages: all of those in GeoNames. We hope that this is one more step toward a completely open, linguistic knowledge base.

Each location is treated as a new synset, which is linked by `instance_hypernym` to a small set of supertypes based on GeoNames categories. These are linked to the collaborative interlingual index, based on an extended set of mappings from GeoDomainWordnet. If a location is already in the interlingual index, then it is also linked to the entry, using mappings from the Geo-Wordnet. Finally, we added some additional structure, if GeoNames places the location in a larger location, this is linked using the `mero_location` link. The data and code to produce GeoNames Wordnet are released under the MIT licence.

We have some ideas for future work:

- There are translations and definitions for the feature nodes for the languages (bg, nb, nn, no, ru, sv) in GeoNames, and for Italian in GeoDomainWordnet: we should add them.
- Almost half the names have language unknown, we could try to deduce the language perhaps by seeing which country it is in.
- Many of the names are transliterations: e.g. GN: `10630004` has both **庄内町** and it's latin equivalent ***Shōnai-machi*** while GN: `11209749` **小萩** *ohagi* has both hiragana and katakana (**おはぎ** and **オハギ**). The GWA LMF allows us to treat these as variants, but this requires language specific knowledge.
- We need to make sure all locations are merged across languages, we will propose an extension to CILI based on GeoNames IDs.
- We found some errors in the GeoNames database (spaces in fieldnames and so forth). We will fix these online.

Acknowledgements

We would like to thank Nathanael Kusanda, who helped with the coding.

Language	Code	Synsets	Lemmas	Senses	Asia	Most Common
Arabic	ar	232,575	197,679	252,316	آسيا	الظهرة
Bulgarian	bg	30,518	29,419	38,059	Азия	Чуката
Catalan	ca	13,857	13,270	14,292	Àsia	Irlanda
Danish	da	3,455	3,444	3,557	Asien	—
German	de	56,548	51,334	58,332	Asien	Neuhof
English	en	599,552	481,369	628,376	Asia	Union Township
Spanish	es	407,846	215,948	439,396	Asia	San Antonio
Estonian	et	4,220	3,914	4,334	Aasia	Washingtoni maakond
Basque	eu	8,860	7,444	9,169	Asia	Washington konderria
Persian	fa	272,151	377,549	492,500	—	Ḩoseynābād
Finnish	fi	48,628	30,641	49,420	Aasia	Isosaari
Irish	ga	3,111	2,893	3,169	an Áise	An Baile Nua
Galician	gl	1,954	2,075	2,125	—	A Rioxa, Guadalaxara
Hebrew	he	14,199	20,985	21,875	אסיה	לבנים יד בית
Hindi	hi	2,166	2,245	2,326	एशिया महाद्वीप	चर्च ऑफ गॉड वर्ल्ड, …
Croatian	hr	2,060	2,098	2,157	—	Sveti Martin, Nova Gora, …
Indonesian	id	322,293	217,548	325,413	Asia	Krajan
Icelandic	is	5,293	4,583	5,590	Asía	Tunga
Italian	it	31,631	32,607	40,492	Asia	Contea di Washington
Japanese	ja	103,881	145,047	184,080	アジア	本町
Lithuanian	lt	33,811	28,831	34,383	Azija	Girelė
Marathi	mr	2,210	2,167	2,271	—	डेटन
Malay	ms	36,993	30,797	37,259	—	Kampung Baharu
Burmese	my	712	728	746	—	—
Dutch	nl	17,316	17,108	17,795	Azië	Bergen
Nynorsk	nn	3,006	2,972	3,056	—	Balearane, London lufthamn, …
Norwegian	no	620,012	423,491	685,065	Asia	Stormyra
Polish	pl	21,456	19,578	21,659	Azja	Hrabstwo Washington
Portuguese	pt	64,161	50,208	65,752	Ásia	Sítio São José
Romanian	ro	7,692	6,703	7,988	—	Comitatul Washington
Sanskrit	sa	658	662	666	—	कुरुक्षेत्रम्, सेंट लूसिया, …
Slovenian	sl	1,640	1,644	1,705	—	Otok
Albanian	sq	3,749	5,613	5,990	—	Novosellë
Thai	th	240,365	168,682	256,304	เอเชีย	หนองบัว
Tswana	tn	7	9	9	—	—
Turkish	tr	40,001	31,978	43,143	Asya	Yeniköy
Venda	ve	11	10	11	—	Kuritiba
Xhosa	xh	32	34	34	—	—
Chinese	zh	740,984	495,549	826,003	亚洲	新村
Zulu	zu	273	291	291	—	—
Total	40	3,649,522	3,129,147	4,587,108	—	—

Table 2: GeoNames Wordnet stastics for various languages

We also show the lemma for Asia, and the most common name in GeoNames for each language

References

Dirk Ahlers. 2013. Assessment of the accuracy of GeoNames gazetteer data. In *Proceedings of the GIR Workshop*, page 74–81.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL `http://aclweb.org/anthology/P13-1133`.

Davide Buscaldi and Paolo Rosso. 2008. GeoWordNet: Automatic georeferencing of WordNet. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 1255–1258. Marrakech, Morocco.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Francesca Frontini, Riccardo Del Gratta, and Monica Monachini. 2016. Geodomainwordnet: Linking the geonames ontology to wordnet. In Zygmunt Vetulani, Hans Uszkoreit, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 229–242. Springer International Publishing, Cham.

Fausto Giunchiglia, Vincenzo Maltese, Feroz Farazi, and Biswanath Dutta. 2010. Geowordnet: A resource for geo-spatial applications. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *The Semantic Web: Research and Applications*, pages 121–136. Springer Berlin Heidelberg, Berlin, Heidelberg.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302. Mysore, India.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago - a core of semantic knowledge. In *16th international World Wide Web conference (WWW 2007)*.

Antonio Toral, Stefania Bracal, Monica Monachini, and Claudia Soria. 2010. Rejuvenating the Italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*.

Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual global wordnet grid. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. 419–426.