Ideas and notes:

# 26/07

* The first part of an encyclopedia article could constitute a more informative, less strict definition. Or a context to exploit in Paragraph2Vec manner.  
  We can use DBpedia’s introductory paragraph (is it also *rdfs:comment* ?), and possibly also the *dbo:abstract* (English) property to retrieve the equivalent.
* One of the problems of GNNs is that you can not easily stack many layers, because you end up introducing noise from irrelevant distant nodes.  
  Idea: use a GRU-GNN, and expose all the layers, learning a combination of the output of all the layers, not just the last one. (as done in ELMo)
* Autoencoder: it seems to have no place in the architecture at the moment.  
  (Unless we encode the definition, and then recreate it)
* Generally, good non-Wikipedia dictionaries have an API with a price.

For easy reproducibility, the dictionary data base should be free.

* Could it be good to HTML-parse from thefreedictionary.com, that reunites several sources? No, too much effort for too little comparative gain.  
  They do not have an API due to licensing issues.
* More opportune to use the APIs and the link connections between:
  + BabelNet
  + (and/or) WordNet
  + DBpedia
  + Wiktionary
  + OmegaWiki

¨

## Analysis of BabelNet:

Example currently examined: ‘sunlight’, for simplicity. Including some observations from the more complex ‘plant’, as well.

12 results, 4 concepts.

However, 2 of these concepts are “fake concepts” and reasonably they should be named entities.

* **{sunlight • sunshine • sun}** The rays of the sun
* **{light• visible light • visible radiation • sunlight}** (physics) electromagnetic radiation that can produce a visual sensation
* **{Sunlight (Benson)}** Sunlight is an oil painting by Frank Weston Benson currently in the permanent collection at the Indianapolis Museum of Art.
* **{Sunlight (DJ Sammy song)}**

How to differentiate them?

We could look at the number of semantic connections: 1490, 2244, 147, 40. …

*However*, considering the ‘plant’ example, one of the senses ({plant}: An actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience) has only 9 connections while still being relevant.

Alternatively, although it may be more difficult to implement in practice, some categories could be included/excluded.

For instance, “Categories: 2003 singles, DJ Sammy songs” could be excluded, whereas “Atmospheric radiation, Climate forcing, IARC Group 1 carcinogens, Light sources, Solar energy, Sun” would be included…

Let us dive now into the 1st synset : **{sunlight • sunshine • sun}** The rays of the sun

1) Definitions from: WordNet, Wikipedia, Wikidata, Wiktionary, OmegaWiki

In this particular example, all the definitions are relevant.

Question: Are they identical, or do we need to separately access Wiktionary/OmegaWiki to get better/more definitions?

There is a difference in focus: the BabelNet quote is meant for that particular sense of the synset, whereas the Wiktionary entry includes all (e.g. also “(figuratively) Brightness, hope; a positive outlook.”)

And in the cited example, that figurative sense is not represented at all in BabelNet, and not even in OmegaWiki either.

Therefore, yes, it is worth it to separately access Wiktionary. To avoid eccessive overlaps, it is opportune to extract only WordNet and Wikipedia definitions from BabelNet.

2) Several examples. Examples can be useful as a more focused, restricted context. (in the context co-occurrences source for the GNN, they can have a greater weight).

However, the examples here can be about any term of the synset.

So either: filter out all the synset words (“We were warmed by the bright [sunshine].”, “[Sunlight] on the skin gives you vitamin D”), or ignore this entirely and search for examples elsewhere.

3) Semantic relations. Currently not used in this task.

4) Images. “”

5) Translations. “”

6) Sources

- WordNet senses: (for each element of the synset). In the current version, it can be considered redundant. We have the definition and the synset grouping already.

It could be useful to get a link to WordNet’s semantic relations for the specific target word.

- Wikipedia page: the first paragraph is useful as an encyclopedia-style extended definitions. However, it coincides with DBpedia’s introduction.

- OmegaWiki senses: useful. We can get a direct connection on the specific sense of the target word (the others in the synset can be ignored)

- WikiData: useless

7) Categories

8)Compounds: maybe they could be used to locate phrases? (e.g.: collecting sunlight, intense sunlight, Sunlight Solar energy, sensitivity to sunlight)

However: Compounds for ‘plant’ is too long a list, with 30+ phrases (“… plant engineering, flora of Japan, Nuclear power plants, aquatic plant, marine plants, flowering plants, list of plants.”)

It may have another use, instead, not necessarily directly connected to the word embeddings: we may collect compound expressions to build an encyclopedia of phrases and idiomatic expressions

9) External Links

- DBpedia: Sunlight

Extremely useful, direct connection to target word in DBpedia. Once again, it appears that the most useful properties are *dbo:abstract* and *rdfs:comment*, that provide an extended definition.

Note: Are these 2 properties identical. Checking on ‘plant’…

It appears that the *rdfs:comment* is a core subset of the *dbo:abstract*

## Onwards

Observation: Generally, the definitions “from WordNet”*in the synsets that contain the word* constitute a superset of the definitions *for the target word* (e.g. “plant”) in WordNet.

Possibilities:

- Restrict: Examine only the BabelNet synsets which correspond (word-by-word) to a WordNet definition

- Widen: Examine all synsets, and include everything, thus pulling the WordNet definitions of words different from the target word as well

- Choose by rank: we include the synsets where the target word is 1st or 2nd. If it is 3rd or beyond, ignore.

Note: maybe a better WordNet – BabelNet correspondence is determined by whether the elements in the synset are 100% coinciding? They do not \*always\* coincide, so synset + definition is probably a better idea to get the correspondence…

It would appear that the “Choose by rank” policy is the most promising, as the most likely to include relevant additional senses (see for ‘plant’ without adding too much unimportant information / noise from other words (e.g. including ‘pan’ into ‘pot’, or ‘light’ into ‘sunlight’)

* We could also use different weights for the “rank/core” contributions…
* Or we could use a “restrict-WordNet” policy, while including the relevant (eg. Close, or key concepts) synsets among the synonims

# 27/07

Describing again the task:

A Graph NN can be applied over:

1. dictionary definition co-occurrence graph
2. context-window occurrences
3. (synonims & antonyms)

The aim is to combine the input of context co-occurrence methods and dictionary information

1)

BabelNet: Select the relevant synsets, and extract the WordNet and Wikipedia definitions

DBpedia: From BabelNet>External links, go to the DBpedia page and get *rdfs:comment*

Wiktionary: Go to the Wiktionary page of the target word, and either use the API or parse to get English > Noun / Verb > List of definitions

OmegaWiki: get the English definitions from the page/content of the target word, all of them.

## Selecting the relevant synsets from BabelNet

We review the examples ‘sea’, ‘plant’, ‘high’:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Synset for: Sea | Definition | Key | In WordNet | Connections | Relevant? |
| sea | (a division of an ocean or a large body of salt water partially enclosed by land) | Y | Y | 4.4K + | Y |
| ocean • sea | Anything apparently limitless in quantity or volume | N | Y | 5 | Y |
| sea | Turbulent water with swells of considerable size | N | Y | 7 | Y |
| ocean • body of water • bounding main • sea | A large body of water constituting a principal part of the hydrosphere | Y | C | 3.1K + | Y |
| Lunar mare • mare • lunar maria • sea | The lunar maria are large, dark, basaltic plains on Earth's Moon, formed by ancient volcanic eruptions. | N | C | 842 | C |
| Sea (advertisement) | Sea is an advertising campaign launched by Diageo in 2007 to promote Smirnoff brand vodka. | N | N | 76 | N |
| Seamester • mester Global Programs • Sea | Sea|mester Global Programs is an organization which offers academic, study abroad programs on board two sailing vessels, Ocean Star and Argo. | N | N | 43 | N |
| seah (unit) • Sea (unit) | The se'ah or seah is a unit of dry measure of ancient origin used in Halakha, which equals one third of an ephah, or bath | N | N | 42 | N |
| Sea | Genus of insects | N | N | 29 | N |
| sea | Heraldic figure | N | N | 13 | N |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Synset for: Plant(name) | Definition | Key | In WordNet | Connections | Relevant? |
| industrial plant • plant • works | Buildings for carrying on industrial labor | Y | Y | 66 | Y |
| flora • plant • plant life | (botany) a living organism lacking the power of locomotion | Y | Y | 4.2K + | Y |
| plant | An actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience | N | Y | 9 | C |
| plant | Something planted secretly for discovery by another | N | Y | 4 | C |
| factory • manufactory • manufacturing plant • plant | A plant consisting of one or more buildings with facilities for manufacturing | Y | C | 1.6K + | C |
| assembly line • production line • line • plant | Mechanical system in a factory whereby an article is conveyed through sites at which successive operations are performed on it | N | C | 428 | N |
| factory (trading post) • factorij • manufactory • plant | "Factory" was the common name during the medieval and early modern eras for an entrepôt – which was essentially an early form of free-trade zone or transshipment point. | N | N | 392 | N |
| Glossary of cue sports terms • Plant (snooker) | The following is a glossary of traditional English-language terms used in the three overarching cue sports disciplines: carom billiards referring to the various carom games played on a billiard table without pockets; pool, which denotes a host of games played on a table with six pockets; and snooker, played on a large pocket table, and which has a … | N | C | 386 | N |
| shill • Plant (person) | A decoy who acts as an enthusiastic customer in order to stimulate the participation of others | N | C | 367 | C |
| Glossary of professional wrestling terms • Plant (professional wrestling) | Professional wrestling has accrued a considerable nomenclature throughout its existence. | N | N | 332 | N |
| Creature type (Dungeons & Dragons) • Plant (Dungeons & Dragons) | … | N | N | 320 | N |
| PLANT | Corporation | N | N | 133 | N |
| Control System • Plant (control theory) | A plant in control theory is the combination of process and actuator. | N | C |  | N |

Note: Verbs and names can be recognized by the end of their ID:

Example: bn:00046568**n** versus bn:00091692**v**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Synset for: Plant (verb) | Definition | Key | In WordNet | Connections | Relevant? |
| set • plant | Put or set (seeds, seedlings, or plants) into the ground | N | Y | 18 | Y |
| implant • embed • engraft • plant | Fix or set securely or deeply | N | Y | 9 | Y |
| found • constitute • establish • plant | Set up or lay the groundwork for | N | Y | 8 | Y |
| plant | Place into a river | N | Y | 5 | C |
| plant | Place something or someone in a certain position in order to secretly observe or deceive | N | Y | 8 | Y |
| implant • plant | Put firmly in the mind | N | Y | 4 | Y |
| set up • establish • found • plant | Set up or found | N | N | 6 | C |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Synset for: high (noun) | Definition | Key | In WordNet | Connections | Relevant? |
| high | A lofty level or position or degree | N | Y | 5 | C |
| high • high-pressure area | An air mass of higher than normal pressure | N | Y | 174 | Y |
| high | A state of sustained elation | N | Y | 6 | Y |
| high | A state of altered consciousness induced by alcohol or narcotics | N | Y | 8 | Y |
| heights • high | A high place | N | Y | 3 | Y |
| high school • senior high school • high | A public secondary school usually including grades 9 through 12 | Y | Y | 23.1K + | Y |
| high • high gear | A forward gear with a gear ratio that gives the greatest vehicle velocity for a given engine speed­­ | N | Y | 10 | C |
| Top (technical analysis) • High (technical analysis) | In technical analysis, a top is an event in which a security's market price reaches a high, then a higher high, and then a lower high | N | N | 83 | C |
| Ledisi discography • High (Ledisi song) | This article contains the discography of American soul and R&B singer-songwriter Ledisi. | N | N | 54 | N |
| High (Young Rising Sons song) | "High" is a song recorded by New Jersey band Young Rising Sons, released as the band's debut single on July 22, 2014. | N | N | 21 | N |
| High (Royal Headache album) | High is the second studio album by Australian punk rock band Royal Headache, released on 21 August 2015 by What's Your Rupture?. | N | N | 17 | N |
| High (tectonics) | A high in structural geology and tectonics an area where tectonic uplift has taken place relative to its surroundings. | N | N | 9 | Y |
| High (computability) | In computability theory, a Turing degree is high if it is computable in 0′, and the Turing jump is 0′′, which is the greatest possible degree in terms of Turing reducibility for the jump of a set which is computable in 0′. | N | N | 7 | C |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Synset for: high (verb)  *(Not in WordNet)* | Definition | Key | In WordNet | Connections | Relevant? |
| high | To hie; to hasten. | N | N | 0 | N |
| high | To rise. | N | N | 0 | N |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Synset for: high (adverb) | Definition | Key | In WordNet | Connections | Relevant? |
| high • high up | At a great altitude | N | Y | 1 | Y |
| high | In or to a high position, amount, or degree | N | Y | 3 | Y |
| luxuriously • high • richly | In a rich manner | N | Y | 3 | Y |
| high | Far up toward the source | N | Y | 0 | Y |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Synset for: high (adjective) | Definition | Key | In WordNet | Connections | Relevant? |
| high | Greater than normal in degree or intensity or amount | N | Y | 15 | Y |
| high | (literal meaning) being at or having a relatively great or specific elevation or upward extension (sometimes used in combinations like `knee-high') | N | Y | 20 | Y |
| eminent • high | Standing above others in quality or position | N | Y | 4 | Y |
| high-pitched • high | Used of sounds and voices; high in pitch or frequency | N | Y | 18 | Y |
| high • in high spirits | Happy and excited and energetic | N | Y | 4 | Y |
| gamey • gamy • high | (used of the smell of meat) smelling spoiled or tainted | N | Y | 3 | Y |
| high • mellow | Slightly and pleasantly intoxicated from alcohol or a drug (especially marijuana) | N | Y | 10 | Y |

# 29/07

Observations:

* Not all synsets have a DBpedia entry. For instance, there are 4 synsets corresponding to WordNet senses for ‘plant’, and only the 2 main senses have External Links > DBpedia.
* Choice: due to the presence of other independent sources (Wiktionary, OmegaWiki), and due to the later insertion of synonims and antonyms, we decide to use a **Restrict** policy: we include the synsets of nouns (and also verbs, adjectives and adverbs) that have a direct correspondence with a sense of the target word in WordNet.
* About the subsequent Synonyms&Antonyms step: considering the WordNet synsets *and* the key concepts, all the other words in the sets will be counted as synonyms.

Operating with the nltk interface for WordNet:

We start from a target word, e.g. ‘plant’

We retrieve all the synsets for it.

[Synset('plant.n.01'), Synset('plant.n.02'), Synset('plant.n.03'), Synset('plant.n.04'), Synset('plant.v.01'), Synset('implant.v.01'), Synset('establish.v.02'), Synset('plant.v.04'), Synset('plant.v.05'), Synset('plant.v.06')]

Remember that POS-tagging and the different roles and meanings of a word are not addressed in this task. The purpose is to obtain graph-based, dictionary-enhanced word embeddings, not multi-sense.

Therefore, we simply collect all the definitions, without having POS-based differences.

(Just use synset.definition() on all)

# 31/07

Working with the HTTP API for BabelNet.

Given a target word, we can:

- select its synsets

- exclude those where the synsetType is NAMED\_ENTITIES instead of CONCEPTS

[problem: How do we deal with words like: “New York”, “Copenhagen”, “London”?

We have to rely on the other sources: WordNet, DBpedia and Wiktionary]

- *Restrict to WordNet* policy: in the list of senses:

- if there isn’t any WordNetSense, drop

- go into properties > simpleLemma. If the original target word is not contained in any of the lemmas, then drop. If it is found, keep.

- collect the definitions, through: glosses > [one of the dicts in the list] gloss , where the source must be either WN or WIKI

Observations: As it is, the definitions pulled from the BabelNet synsets are a superset of the Wordnet ones, because they also include related synsets.

While it is true that all accepted entries must have a correspondence in WordNet (although not necessarily directly with the target word), maybe it would be appropriate to move some of the related entries to the synonyms?

DBpedia access: we pull the encyclopedia definitions of the target words (written in CamelCase), proceeding as follows:

**PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  
PREFIX dbres: <http://dbpedia.org/resource/>  
PREFIX dbo: <http://dbpedia.org/ontology/>  
  
SELECT DISTINCT ?encyclopedia\_def  
WHERE { dbres:"""** + target\_word + \  
**""" rdfs:comment ?encyclopedia\_def   
 FILTER (LANG(?encyclopedia\_def)='en')}**

*rdfs:comment* provides 3-4/7-8 rows. *dbo:abstract* has a more extensive definition, but I consider it too extensive, especially if compared with the other sources

# 01/08

To extract information from Wiktionary, we can use the wiktionaryparser Python package. Here the target word should be in lowercase, not CamelCase.

The structure returned is:

[ #list containing 1 dictionary

‘etimology’: {str} “From Middle English plante, from Old English plante (“young tree or shrub, herb newly planted”), …”

**‘definitions’**: {list} [

{dict} {‘partOfSpeech’:”noun”,

**‘text’**: [*'plant (plural plants)'*, '

(botany) An organism that is not an animal, especially an organism capable of photosynthesis. Typically a small or herbaceous organism of this kind, rather than a tree.',

'(botany) An organism of the kingdom Plantae; …’,…] }

'relatedWords': …

**'examples'**: [ …

{dict} {

'partOfSpeech' = {str} 'verb'

**'text'** = {list} [

*'plant (third-person singular simple present plants, present participle planting, simple past and past participle planted)'*,

…

'relatedWords' = {list} [{'relationshipType': 'related terms', 'words': ['plantation']},…]

**'examples'** = {list} ["…

}

OmegaWiki already provides an API. We navigate the structure returned by the request:

'http://www.omegawiki.org/api.php?action=ow\_express&search='

+ target\_word +'&format=json'

# 02/08

I decided to use the HDF5 data format to store the definitions, since it is a binary format known for its scalability and speed, and it can be used to store tables and organize them in groups/subfolders.

The pandas library provides an interface.

Note: given the current retrieval mechanisms, some of the definitions retrieved from BabelNet-WN are going to coincide with the ones from WordNet.

Handling this aspect also depends on the organization of the definition data.

Organization, prelude:

“The ordinary HDF5 library apparently does not even support concurrent reading of different files by multiple threads…”

On the matter,

“Starting with version 2.5.0, h5py includes support for the HDF5 SWMR features.

The SWMR features allow simple concurrent reading of a HDF5 file while it is being written from another **process**.”

Taking into account that every read/write must be sequential anyway, we have 3 alternatives:

1. Put all the definitions, from all the sources, into one file.  
   n: We can still use different threads to send the web requests, gather the data from them, dump.
2. Use several files, one per source (Wordnet (Wikipedia+WN), BabelNet, Wiktionary, DBpedia, ΩWiki)
3. Use several files, one per each chunk of the vocabulary (e.g. 5000 words)

Moreover: how to deal with the partial overlap BabelNet-WN & WordNet?

The frequency of a word in a set of definitions is an important piece of information: it will add a weight to the edge in the graph.

It is opportune to include only one copy of a definition from WordNet. Policy: *exclude* all WordNet definitions from the BabelNet input, keep only those from Wikipedia. This has the added advantage of not focusing on the synonyms’ definitions.

The workflow becomes the following:

* Use different threads to send the web requests
* Gather the data from them,
* Write all the definitions for that word into a file. We may use either one HDF5 file, or several (one for each vocabulary chunk) depending on the size.

# 03/08, 04/08

Immediately before setting up the definition retrieval, it is necessary to examine:

Which elements are we going to extract from each source?

* WordNet:
  + Definitions
  + Examples
  + Synonyms: from the lemmas in the same synset
  + Antonyms: in adjectives, that have the Antonym semantic relation.
* BabelNet:
  + Definitions: only those from Wikipedia  
    among the other sources: WordNet, Wiktionary and OmegaWiki are already collected separately. Wikidata, Wikiquote, and FrameNet are not as relevant and/or reliable
  + Examples
  + Synonyms: from the lemmas in the selected synsets
* Wiktionary:
  + Definitions
  + Examples
* OmegaWiki:
  + Definitions
  + Synonyms (provided for each meaning)
* DBpedia
  + Encyclopedia definition

Emerging questions:

1. Multi-word expressions.

For instance, the list of synonyms for the term ‘Sunlight’ from BabelNet is:

['sunlight', 'sunshine', 'sun', 'solar\_radiation', 'Natural\_lighting', 'Solar\_Irradiation', 'Solar\_irradiation', 'Solar\_output', 'Solar\_Radiation', 'Solar\_spectrum', …]

And the list of synonyms for ‘sunlight’ in OmegaWiki is:

['solar radiation', 'sunlight', 'sunshine']

In the current version, we take the simplifying choice and eliminate all multi-word synonyms (i.e. with an underscore or with a space). We can do this after we have retrieved them, when it is the moment to store them.

1. Antonyms

What could be a good source for antonyms? \*\*To be continued…

# 05/08

I deem useful to employ multiple threads, since we send the HTTP requests to different sources.  
(For the computation part, it is equivalent to operating sequentially, as Python threads keep a global lock. Multiple processes do not share the same memory, and dealing with IPM is not worth it).

NO: some remote services may be occasionally unavailable (e.g.: BabelNet. It is opportune to write to different hdf5 files for each source, and postprocess afterwards).

Parallelism goes in the dustbin as a consequence.

Note: we need each thread to return values: the definitions/examples/synonyms/antonyms that have been extracted.

We can use concurrent.futures.ThreadPoolExecutor

Note: when storing examples, it is opportune to eliminate duplicates (since part of those in BabelNet are also present in WordNet and Wiktionary)