Proposal: Creating a Simple Guide to Building Wordnets

Abstract

Building a wordnet can be a daunting task, especially at the beginning. This document is a short proposal on the creation of a simple guide to building wordnets. The primary goal of the guide is to be a clear and friendly resource to help someone new to building wordnets to get started, and also be something more experienced builders can refer to when they build new wordnets.

1 Introduction

Building a wordnet, or any lexical resource, can be daunting, and many thoughts can go through the mind of someone who is starting on a wordnet: what words should be included? How does one decide? Should one go for breadth or depth? Is that limited to only a particular domain? How much time will be needed?

An existing guide such as Vossen (-)'s presentation on building wordnets provides a good overview of the task at hand, but it could be a lot of information to digest for someone new to building wordnets. The proposed simple guide is intended to make the on-boarding process easier.

This proposal will (i) state the goals of the guide, (ii) provide a table of contents for the guide, and (iii) state briefly the style of content/writing that is to be expected.

2 Goals

The primary goal of the guide is to be a simple and user-friendly resource for someone new to building wordnets (or other lexical resources), by providing clear explanations to get started. At the same time, it should also be useful enough for more experienced builders, giving them a useful "cheatsheet" for when they embark on starting another wordnet.

The guide will also include more recent developments, such as the Collaborative Interlingual Index (CILI, Bond et al. (2016)), and how builders of new wordnets can contribute to or make use of them.

Another goal of the guide is to help those with limited resources (manpower, lexical, technology, funding, knowledge of the language, etc) to get as much out there as possible. These could be wordnet builders who work on them as personal or side projects, and/or those whose language do not have adequate or satisfactory lexical resources. They could also be non-native–/non–speakers of the language.

Lastly, the guide will also touch upon open data and open licenses, and provide some information on choosing an appropriate license. The guide will also provide some help for builders to prepare and getting wordnets ready for release to the public.

A stretch goal is to gather the thoughts and experiences of existing wordnet builders (of varying coverage), such as the methods they used to get started, and the common issues they encountered in the process, and make them part of the guide. While the wordnets' respective papers would have covered these, having them in the guide will be convenient for the readers.

3 Proposed Content

In this section, we propose a table of contents for the guide. These are only suggested content, and any additional input will definitely be welcomed. Existing guides can also be used as a guideline to the type of content to provide.

- Introduction
- Gathering Resources
  - The lexical and non-lexical resources a builder can potentially use to build up their wordnet's lexical inventory, such as dictionaries, books, corpora, etc.
- Building
  - What words to choose? What to exclude?
  - Expand or merge?
  - Language-/domain-specific words

    - Proposing new synsets (to CILI, for instance)
    - Writing definitions and giving examples
- Verifying and Checking
- Releasing the Wordnet
    - How to release (licenses and formats)
    - Where to release
    - When to release

4 Style

The writing style will be relaxed and informal, and as much as possible, be in plain English. Some humour will certainly be welcomed. Terminology used for wordnets or linguistics will be explained clearly, keeping in mind that people building a wordnet might not have lexicographical or linguistic experience or knowledge.

Paragraphs will be short, and common issues and problems, as well as tips, will be highlighted clearly. Links to further resources will also be provided. Other media, such as images, will be useful as well.

5 File Format & Miscellaneous

Two file formats will be used: HTML and PDF. The HTML file will be the main format, as it provides greater flexibility, compatibility and up-to-date–ness. The PDF, on the other hand, will be available for the purpose of portability. Other formats will certainly be possible.

6 Sample Guide

The guide is already in the process of being written. An early alpha version of the guide is presently available at [URL to be provided in final submission]. A sample extract of some sections of the guide is provided in the appendix.

7 Closing

We proposed the creation of a simple guide to building wordnets, to be a getting-started guide for new builders, and also a reference for more experienced builders. Among the goals of the proposed guide is also to provide some information and tips for those who might have limited resources to dedicate to the project.

Ultimately, our hope for the guide is to make the creation of a wordnet as pain-free as possible, and perhaps even fun. It will certainly not be the exhaustive and ultimate guide to building wordnet. Nevertheless, we hope that it will be useful in leading new builders in the right direction, and be their warm welcome to the wordnet community.

8 Appendix: Guide sample

We selected a few sections from an early, "alpha" version of the guide as a sample, to give an idea of the content and style.

Here is an extract from **Gathering Resources > Lexical Resources**:

> Our recommendation is to begin with the smaller dictionaries. Pocket dictionaries — such as the traveller dictionaries for tourists and young learners — are useful starting points, as they generally contain the most common words to be used, and as such give you an idea of which words to start off.
>
> Another advantage of pocket dictionaries is that they typically provide single words (translations), rather than definitions, which will help you get the right words. However, a disadvantage of having such single words is the ambiguity or vagueness that can arise, since you will not get the exact meaning or nuances of the word, which is essential for selecting the right synset.
>
> [...]
>
> Online, searchable dictionaries and databases are very useful for getting to the words you need. Multi-lingual resources like Wiktionary might contain even more information about the words (including examples, derivations and composition) which might help you quickly build up your lexicon.

From **Gathering Resources > Other Resources > Corpora**:

> Corpus-based approaches are increasingly popular these days for empirical language research. We can make use of corpora in your language to gain an idea of what the most common words in that language are, and add them to your wordnet.
>
> [...]
>
> You can build your own corpora from different texts. Where possible, it is good to gather texts from different sources and genres, such as stories (novels, folktales), newspapers, blog posts, social media posts, and others. The latter few sources are good for more colloquial uses of the language, which you might miss out if using only newspapers or more formal texts. Also, texts which are native to the language (folktakes and legends) could give you a better chance of getting language- or culture-specific words, when compared to translations.

Various subsections from **Building**:

> **Expand or Merge?**
>
> When you base your wordnet on the PWN synsets, you are taking the *expand* approach. [...] The use of the Core synsets is an example of this. It is a fast and efficient way to begin your wordnet, as the structure is already there, and it can be integrated with existing wordnets without much hassle. However, the structure might not be entirely relevant or adequate for your language.
>
> If you created your own concepts and semantic links separately, you can take the *merge* approach. This means taking your concepts and links and aligning them with the wordnet structure. This might not be as quick as using the expand approach; however, it could be a more accurate and complete representation of your language, and you do not have to worry about meeting a coverage requirement.
>
> [...]

Language-specific and domain-specific words

If you use the translation approach, you might miss out on many words in the language you are working on (domain-/culture-specific words), that do not have synsets in the Wordnet.

In the initial stages, if speed of coverage is your goal, you might choose to exclude these words. However, for your wordnet to be more complete for your language, you will eventually have to look at these words. There might even be some words which are very basic to the language (such as specific personal pronouns) which do not have an accurate or appropriate synset. As these words do not have an existing synset, you will need to create one yourself, and also provide the semantic relations (as much as possible).

[...]

From **Building > New Synsets > Writing Definitions**:

Your proposed new synset should have an English definition and preferably also one in the wordnet's language itself. Definitions should be *accurate* but *concise*, and should strive to use simple terms where possible. Avoid using the word or derivatives of that word in the definition itself. The reverse — when the word is itself the derived term — however, should not be an issue (Eg: *writer* "a person who is able to write and has written something").

Definitions can be written using a hierarchical approach: begin by defining the main category-type of the word (the *genus* or *class*), and then narrow down to the specifics (the *species*), with additional details to further narrow things down if necessarily.

- Computer: a machine$_{\text{[genus/type]}}$ for performing calculations automatically$_{\text{[species/specific]}}$
- Album: a book$_{\text{[genus/type]}}$ of blank pages with pockets or envelopes$_{\text{[species/specific]}}$; for organizing photographs or stamp collections etc$_{\text{[more specific]}}$

"Simpler" words might be harder to define!

From **Building > New Synsets > Linking synsets**:

Concepts in a wordnet are linked to one another — ideally there should not be any orphan synsets. Your new concept should therefore also be linked to another concept in the wordnet. [...] For example, if your languages differentiates the concept of *cooked noodles* from *uncooked noodles*, your new concepts for these two will likely link to the concept of *noodles*. You might also want to link to attributes such as *cooked* and *uncooked*.

References

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016, 2016.

Piek Vossen. Building wordnets, -. http://www.globalwordnet.org/gwa/BuildingWordnets.ppt.