Completing the Princeton Annotated Gloss Corpus Project

Alexandre Rademaker
IBM Research and FGV/EMAp
alexrad@br.ibm.com

Bruno Cuconato
IBM Research and FGV/EMAp
bcclaro@gmail.com

Henrique Muniz
IBM Research and FGV/EMAp
hnmuniza@gmail.com

Alexandre Tessarollo
FGV/EMAp and Petrobras
alexandretessarollo@gmail.com

Alessandra Cid
FGV/EMAp
alessandracorreacid@gmail.com

Abstract

In the Princeton WordNet Gloss Corpus, the word forms from the definitions ("glosses") in WordNet's synsets are manually linked to the context-appropriate sense in the WordNet. The glosses then become a sense-disambiguated corpus annotated against WordNet version 3.0. The result is also called a semantic concordance, which can be seen as both a lexicon (WordNet extension) and an annotated corpus. In this work we motivate and present the initial steps to complete the annotation of all open-class words in this corpus. Finally, we introduce a freely-available annotation interface built as an Emacs extension, and evaluate a preliminary annotation effort.

1 Introduction

The Princeton WordNet Gloss Corpus is a corpus of the manually annotated synset definitions (glosses) from the Princeton Wordnet (PWN) (Fellbaum, 1998). The corpus is available for download in the PWN website as one of the stand-off packages that supplement the WordNet 3.0 release.[1] Although it has been already recognized as a precious resource, the project of semantically tagging all PWN glosses was not finished. According to the PWN website, the corpus contains 206,711 words (including collocations) yet to be disambiguated. In simple terms, our goal is to complete the disambiguation of all open-class words in this corpus, and here we present our preliminary findings and methodological decisions.

Previous efforts address this same goal in older versions of PWN using automatic or semi-automatic methods (Harabagiu et al., 1999; Moldovan and Novischi, 2004). Here we aim at high-quality human annotation of the glosses, leveraging the lessons learned and directives developed for the project in Princeton but adapting them to our tools and priorities. Data is available at https://github.com/own-pt/glosstag using the same open license used by Princeton for the current version of the data.

The definitional glosses were introduced in PWN primarily to help humans identify the meaning of the synsets, but recently, many word sense disambiguation (WSD) algorithms use the network structure of PWN in combination with the glosses to improve the identification of the most plausible sense for a given word in a corpus (Agirre and Soroa, 2009; Banerjee and Pedersen, 2002; Basile et al., 2007). By semantically disambiguating the words in the glosses, we add pointers from each word to its synset, and this increases the connectivity between the WordNet synsets by approximately an order of magnitude, hopefully improving the performance of these algorithms.

Another reason for such an effort is to ensure the completeness of PWN. By completeness we mean the property of a lexico-semantic resource that all words used in the definitions of the concepts are also themselves explained in this same resource. Hopefully, this completeness could also help us ensure quality in our long-term endeavor, the expansion of PWN to highly technical domains such as those of the geosciences, agriculture, and law. Once more concepts are added or redefined, we will redefine and add glosses that we intend to disambiguate, forcing us to use the newly added senses in a productive cycle of editing, testing, and correcting.

We begin this paper by discussing the original dataset and how we interpreted it converting to a

[1] http://wordnetcode.princeton.edu/glosstag.shtml

more friendly format. Next we describe our annotation interface and some of our implementation decisions. We continue by discussing some of the issues we encountered while sense-tagging the glosses corpus. Finally, we evaluate our ongoing annotation work, and discuss related work and conclude.

2 Sense tagging

Semantically tagging (or sense annotating) a corpus is a task of constructing a semantic concordance – a textual corpus and a lexicon so combined that every content word in the text is linked to its appropriate sense in the lexicon (Miller et al., 1993). Two different strategies for building a semantic concordance are known: the sequential and the targeted approaches.

(Miller et al., 1993) presented one of the first tools developed for supporting the work on building a semantic concordance with PWN, the ConText. The tool was constructed to support sequential tagging. In this approach, the annotator starts with the corpus and proceeds through it word by word. This procedure has the advantage of immediately revealing deficiencies in the lexicon: missing words, missing senses, and indistinguishable definitions. The sequential process was chosen because of their priorities at that time, as they aimed to make substantial improvements in the PWN. Another tool supporting the sequential approach to building semantic concordances was described by (Bond et al., 2015). The tool was introduced after a brief survey on other tools for sense tagging, none of them actively maintained and freely available at that time.

In the targeted approach, the work starts with the lexicon: we focus on a polysemous word, extract all sentences from the corpus in which that word occurs, categorize the instances and write definitions for each missing sense, and create a pointer between each instance of the word and its appropriate sense in the lexicon; we then repeat the process by choosing another word to focus on. The targeted approach has the advantage of concentrating the annotation effort on a single word, producing better definitions. However, the previously listed flaws in the lexicon would not appear so straightforwardly in this targeted strategy. Consequently, this strategy has the potential of being more successful when the lexicon has already reached a more stable stage. The targeted strategy was the one chosen for the Wordnet Gloss Corpus initial phase; it is described in the original annotators' guidelines that we had access to, and we have decided to follow it as close as possible.

The original Wordnet Gloss Corpus project employed an interface called Mantag, implemented in the Perl programming language.[2] Unfortunately, the tool has many dependencies on legacy code that we were not able to solve.

For our continuation of the Wordnet Gloss Corpus annotation project, we decided to implement a serverless application that can be used offline. This decision reflects the prevailing understanding that semantic annotation is a difficult task that is best done individually and in an environment conducive to concentration. In our tool, each annotator can perform their work independently, making annotations on overlapping parts of the corpus or not. The annotators' data can then be consolidated, possibly including discussions aiming at agreement in the cases where annotations diverge. We have also differed in our technology of choice compared to (Bond et al., 2015). Instead of choosing a web framework we have decided to implement our annotation tool in the extensible and free text editor Emacs,[3] taking advantage of the editor's support for multiple platforms and its rich ecosystem. The annotation interface needs no internet access, depends only on Emacs and its libraries, and can be run either from a graphical interface or from a terminal window.

The annotation interface works as follows: given the directory where the data files are stored, it indexes all tokens to be annotated by their lemmas. This index is persisted to disk so that this indexing does not need to be re-run. The user is then prompted for a lemma and (optionally) a PoS tag; if any matching pairs of lemma and PoS tag are found in the index, a new buffer is opened, containing the glosses where the targeted lemma was found. Colours differentiate token's status: pre-annotated tokens are shown in one color, while tokens yet to be annotated are shown in another; tokens annotated in the current iteration are also shown in a different color. Multiword expressions are marked by subscripts in their constituent tokens (whether they are adjacent or not), while sense and PoS annotations are shown as superscripts. The annotation interface offers the user

[2] https://www.perl.org
[3] https://www.gnu.org/software/emacs/

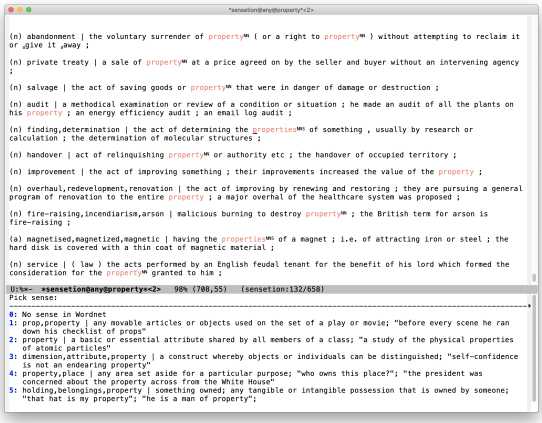

Figure 1: Sense tagging interface

the following capabilities:

1. assign a token zero or more senses;
2. change a token's lemma;
3. mark a token to be ignored (closed-class words or other fragments that are considered meta annotations on the glosses);
4. mark an annotation as having low confidence;
5. create and dismantle multiword expressions.

The first capability is the main functionality of the tool. When selected, the user is asked to confirm the token's PoS tag, and then a dialogue box is shown with all possible senses to that lemma and PoS tag pair, along with their defined terms and glosses. The user can then select or deselect a sense, or explicitly say that there is no sense for that word in WordNet (see Figure 1).

All other commands are there to allow the correct sense annotation of a word. In case its lemma is wrong, there is no way of presenting the user with the correct sense options unless the lemma is corrected; if a token is part of a multiword expression but is not already marked as so, annotators are able to mark it themselves. The dismantling of a multiword expression is necessary for the cases where the token is wrongly assigned as part of a multiword expression, as *rock* and *bass* in Example 2b, where both are marked as part of the multiword expression *rock_bass* (a kind of fish). All commands are available through customizable mnemonic keyboard shortcuts or by a menu.

3 Data Preparation

The Princeton WordNet Gloss Corpus is distributed in two different XML formats: standoff and merged files.[4] We choose to work with the merged files because they are more concise and are precisely described by a document type description (DTD). We have split this data into files containing 100 glosses each, with one annotated gloss per line encoded as an S-expression, a notation for tree-like data.

Every WordNet gloss contains a sense definition. The gloss can be preceded by a domain classification fragment and/or an auxiliary fragment (usually in parenthesis, but not always), and optionally followed by more auxiliary fragments and zero or more examples. In the original XML, all these components are marked up with nesting elements. The tokens are marked up with parts of speech, potential lemma forms, and (optionally) a small set of semantic classes (indicating whether the token is punctuation, abbreviation, acronym, number, year, currency, or some kind of symbol). Collocations are delimited by special markup which can even indicate discontiguous forms. Words and collocations that have been disambiguated are further annotated with WordNet sense keys. To facilitate the implementation of the interface, we have adopted a flat data format where a gloss is a list of tokens, each one of them represented by a property list (see Listing 1). All nesting elements for boundary-marking tokens in the XML files were converted to key-values pairs in the respective tokens. Further details are publicly provided in a README file along with the data itself.

The data conversion is followed by a validation step to ensure that our understanding of the data was right and that no information was lost. Although the XML validation using the DTD takes care of many validation issues, we did find encoding errors and nonexistent sense-keys in the corpus. For the encoding errors, before the conversion, we searched for and replaced invalid characters by UTF-8 legal codes.

Most of the cases of invalid sense keys turned out to be instances of adjective satellites whose sense keys had been wrongly marked with synset type 3 instead of 5, but some cases were tokens marked with

[4] http://wordnetcode.princeton.edu/glosstag.shtml

```
(:ofs "02744323" :pos "n"
 :keys
  (("arterial_road%1:06:00::" . "arterial_road"))
 :gloss "a_major_or_main_route"
 :tokens
  ((:kind :def :action :open)
   (:kind :wf :form "a" :lemma "a" :pos "DT"
    :tag "ignore")
   (:kind :wf :form "major" :pos "JJ" :tag "man"
    :lemma "major%1|major%2|major%3"
    :senses (("major%3:00:06::" . "major")))
   (:kind :wf :form "or" :lemma "or" :pos "CC"
    :tag "ignore")
   (:kind :wf :form "main" :tag "man"
    :lemma "main%1|main%3" :pos "JJ"
    :senses
     (("main%5:00:00:important:00" . "main")))
   (:kind :wf :form "route" :tag "man" :sep ""
    :lemma "route%1|route%2" :pos "NN"
    :senses (("route%1:06:00::" . "route")))
   (:kind :wf :form ";" :pos ":" :tag "ignore"
    :type "punc") (:kind :def :action :close)))
```

Listing 1: Property list encoding of WordNet 3.0 synset `02744323-n`

an undocumented and non-existent sense key `purposefully_ignored%0:00:00::`. The name suggests that this was a virtual sense, created as a way of manually marking tokens as to be ignored in the sense annotation. A case like the annotation of 'ng' in Example 1a[5] seems to support this view. However there is also evidence to believe that the non-existent sense key was created to mark cases where the appropriate sense for a word did not exist in WordNet, as in 'designating' in Example 1b. We also found four cases where a 'purposefully ignored' sense was assigned together with some other sense; these we have revised and corrected manually. These cases include the aforementioned Example 1a (where it had also been tagged as the unit *nanogram*), Example 1c (where *waves* also had been tagged as "(physics) a movement up and down or back and forth"), and Example 1d (where *Mediterranean* was also tagged as "of or relating to or characteristic of or located near the Mediterranean Sea").

(1) a. produced with the back of the tongue touching or near the soft palate (as 'k' in 'cat' and 'g' in 'gun' and 'ng' in 'sing') (01156750-a)

b. designating the player judged to be the most important to the sport; "the most-valuable player award" (01279431-a)

c. atomic events are explained as interactions between particle waves (06107850-n)

d. small dried seedless raisin grown in the Mediterranean region and California (07752966-n)

[5]The identifiers in the end of the examples stand for the synset IDs of PWN 3.0.

4 Challenges

The challenges of this project encompass many aspects: the amount of work, the particularities of the glosses compared to sentences in an ordinary text, and the mismatch between the 'continuous' sense boundaries of words in utterances and the 'discrete' boundaries defined by a lexicon.

The Princeton Wordnet Gloss Corpus contains 117,659 glosses composed by definitions and examples, comprising more than 1,621,129 tokens. So far, 449,355 tokens have been annotated, 118,856 of them automatically. Considering only the taggable tokens, i.e., the open-class words, 206,711 tokens are estimated to remain untagged. From these untagged tokens, we have so far annotated approximately 500 tokens during the development of our tool and training of the annotators. To deal with the amount of work in the next phases of this project, we plan to prioritize the annotation by focusing on domain-specific words most relevant to other projects of our team.

The sense of a word in a text is determined by its context – the more context information we use, the easier is the determination of the right sense of its polysemous words. Compared to other corpora, synset glosses provide relatively little context. Figure 2 summarizes the sizes of glosses (number of characters) by part-of-speech. As we can see, the majority of the glosses has less than 100 characters (76% of them). Moreover, most of the glosses are not complete sentences, e.g. 'secured with bastions or fortifications.' The annotator has to carefully consider the words that are being defined by the gloss and its relations to other synsets, in order to compensate for these obstacles. For some cases, such as 'allomorphs' in 'pertaining to allomorphs' and 'park' in 'The young man

was caught soliciting in the park' the unique viable solution is to allow multiple sense annotation, as described in Section 2.

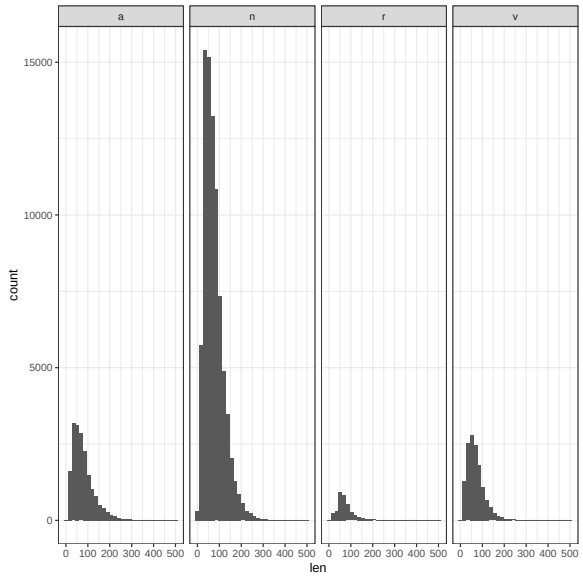

Figure 2: Glosses' sizes (number of characters) by part-of-speech

Regardless of the nature of the sentences, disambiguation of senses is a notoriously hard task that may not be disconnected from the constant revision of the lexical resource being used as a sense inventory. This line of thought is supported by the way previous work on building semantic concordances was conducted. (Miller et al., 1993), carried out sense annotation while expanding and refining PWN, with the annotations continually signalling omissions and inaccuracies in the resource. We have already noted some cases of inconsistencies in PWN such as the case of 'deposit' presented in Figure 3. The dashed red lines point to the two possible senses for the word 'deposit' (bold) in the synset 01576001-v. Note that although the synset 01528069-v seems to be the best option, as it is the direct hyperonym of 01576001-v. The synset 01575675-v is the best matching considering the example in 01576001-v and its hyponym 01988755-v. This situation suggests that 01576001-v should have a different position in the network.

More recently, (Kilgarriff, 1997) has already pointed out 'word senses are only ever relative to a set of interests' and (Rudnicka et al., 2019) emphasizes this point, remarking that dictionaries (or wordnets) and corpora are in two different levels: "Dictionaries and wordnets are metalinguistic

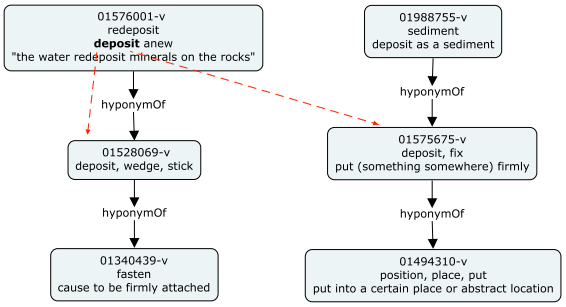

Figure 3: possible senses for an occurrence of the word 'deposit'

generalizations, while corpora are real texts; dictionaries and wordnets include decontextualized isolated items, corpora consist of contextualized continuous text."

Some cases of multiword expressions (MWE) seem to support our belief that sense annotation and PWN maintenance should be joint work. First, we need to define and enforce heuristics to determine when a given word sequence is a multiword expression (being sense annotated as a single entity), and when its component tokens should be annotated individually. The compositionality and conventionality criteria from (Farahmand et al., 2015) may help, however these criteria are not as clear-cut as we would like them to be. Take the case of 'first degree' and the example 'all of the terms in a linear equation are of the *first degree*' in its definition (synset 05861716-n); we can annotate it as 'first degree' (this same sense being defined in the synset where the example is given); but there is no sense for 'second degree', or 'third degree', which are equally valid. This leads us to consider that it should be annotated individually, and that the 'first degree' sense should be removed from PWN.

One can conclude that sense tagging the PWN glosses is a never-ending task, but we believe it is possible (and useful) to achieve definitional completeness in restrictive domains. The question that we face is how to make it feasible and synchronized with the changes in the lexicon (senses, words, and relations). Admittedly, we will need to implement tools for tracking the changes in the dictionary and signal for re-annotation all potentially affected glosses.

Finally, the challenges related to the corpus'

size and to MWEs also interact. To make the annotation process easier we would like to have a certain degree of automation. We have inherited from the original data expressions that have been incorrectly tagged as MWEs, as in the bold words in the sentences 2a and 2b. While it is easy to recognize and fix this kind of error, the other way around is more challenging: identifying an expression that should be added or that is already defined in the lexicon.

(2) a. bearing or producing or containing **calcium** or calcium **carbonate** or calcite (02674398-a)

b. English **rock** star and **bass** guitarist and songwriter who... (11167952-n)

5 Evaluation

We have carried out a preliminary annotation effort to test our interface, train our annotators, and refine our guidelines. In this section we report the issues we found and the results we obtained.

We have trained four annotators and instructed them to annotate all glosses in which one of these three words occurred: 'derivation' (9 occurrences in 8 glosses), 'formation' (153 occurrences in 146 glosses), and 'incompatible' (8 occurrences in 7 glosses). The word 'derivation' has eight senses available from the PWN, while 'formation' has seven senses and the adjective 'incompatible' has nine senses. These example words were chosen to balance frequency and polysemy degree.

After the annotation, two sessions of discussion were conducted to consolidate annotation decisions. We must note that because training the annotators was the goal of this experiment, the results presented here are still very preliminary.

Considering all three words, only half of the occurrences presented full agreement among annotators. But partial agreements are reasonably common as we can see in the tables 1 and 2. The 'ctx' column (for context) numbers the occurrences of a given word; when the number is a decimal it means that there is more than one occurrence of that word in the same context. The other columns number the possible sense an annotator could choose, including a label for the absence of a suitable sense in PWN ('N'). Table 1 presents the annotations of the 'derivation' occurrences. In five out of eight contexts, most of the annotators agreed on one sense, e.g. in the last context, all annotators agreed on sense 7 although annotator **T** also assigned sense 2. We can also note that annotators A and B agreed six times even though one of them also annotated an additional sense.

ctx	1	2	3	4	5	6	7	8	N
1		T				A	ABR		
2						ABRT	A		
3				T		BT	AR		
4			BR						AT
5.1				ABRT					
5.2				ABRT					
6		AT				R	BTR		
7	B	BRT	AB	B	B	B	B	B	
8		T					ABRT		

Table 1: Annotations of all 9 occurrences of 'derivation' in 8 glosses. A, B, R and T stands for the annotators' initials.

Particularly interesting is that all of the eight occurrences of 'incompatible' have almost always been annotated with the most generic sense "not compatible" (00508192-a). Nevertheless, annotators reported this to be the hardest among the three words to annotate. Even to reach a consensus on the proper sense afterwards was a hard task. Table 2 also shows that for annotators **A** and **T**, in all of the contexts that they examined, sense 3 and sense 5 are indistinguishable.

ctx	1	2	3	4	5	6	7	8	9	N
1	T	AT	ART		ABRT			B		
2		T	AT		ABRT			B		
3.1			AT		ABRT					
3.2			AT		ABRT					
4			AT		ABR					
5			T		ABRT					
6			AT		ABRT					
7			T		ABR			B		

Table 2: Annotations of all 8 occurrences of 'incompatible' in 7 glosses. A, B, R and T stand for the annotators' initials.

When we consider the word 'formation,' there are 82 occurrences with full agreement (Table 3 lines 1, 4, 7, 9, and 37). Line 1, for example, shows that there was full agreement regarding sense 3,"natural process that causes something to form", in 71 occurrences. This same sense was selected other 29 times with partial agreement. In 21 of these 29 cases, at least one annotator also chose the sense "the act of forming or establishing something". One such case was the gloss for 'electronegativity', which states "(chemistry) the tendency of an atom or radical to attract electrons in the *formation* of an ionic bond" (04944513-

n). However, although only one annotator had assigned 'formation' to the mentioned sense, after discussions among the annotators, others agreed that it could also be assigned to it in the given context.

qt	1	2	3	4	5	6	7	N
71			ABRT					
13			ABR		T			
13		B					ABRT	
6						ABRT		
5		T		ABR				
4						BRT		A
2					ABRT			
2					T	BRT		A
2				ABRT				
2			AB	R	T			
2			B		ABRT			
2		B		A			BRT	
2		BR					ABT	
2		BRT					A	
1						R		ABT
1					BRT		A	
1					T	ABRT		
1					T	AR		B
1					T	BR		A
1			AB		RT			
1			AB	RT				
1			B		ABRT	A		
1			B		T	ABRT		
1			BR		T		A	
1			BT	AR				
1		A	BRT					
1		ABR					BT	
1		ABR			T		B	
1		B					ART	
1		B	R	ABRT				
1		BRT	A					
1		BT					ABR	
1		BT					ABRT	
1		BT		ABR				
1		R	ABR				T	
1		T	B	ABR				
1	ABRT							
1	B					RT		A
1	T		ABR					

Table 3: Annotations of all occurrences of 'formation'. A, B, R and T stand for the annotators' initials. The first column is the number of contexts where the same pattern of annotation appears.

The case of 'formation' is in agreement with results from (Leacock et al., 1993) that say "the degree of difficulty involved in resolving individual senses is a greater performance factor than the degree of polysemy.". It also suggests a two-step sequential approach to annotation: first the annotators agree on each synset's scope and only then do they proceed to the actual annotation process. This two-step approach will be the object of a future investigation.

As for multiword expressions, expressions such as 'military formation', 'geological formation', 'reticular formation', and 'reaction formation' are removed from the above quantitative analysis but we have discussed them. The expression 'military formation' stands out in many glosses. The expression exists as a MWE but a similar expression, 'naval formation' does not, with both appearing in the gloss "the side of military or naval formation". Annotators discussed whether 'naval formation' should be considered a MWE or whether 'military formation' should not be considered one.

An annotator's familiarity with a particular domain also plays a role in the annotation process, affecting both the senses assigned and the decisions regarding which collocations should be considered MWEs. For instance, the expression 'rock formation' is not part of PWN, but it appears many times in the corpus (see Example 3a).

(3) a. a national park in Utah having colorful *rock formations* and desert plants and wildlife (08603525-n)

b. the gradual movement and formation of continents (11434448-n)

Although some of us believe the expression should be added to PWN, it is not in the lexicon yet, and so, three annotators chose the sense '(geology) the geological features of the earth' for the word 'formation' in all occurrences of the expression. This decision was understandable if we consider that the word 'rock', in one of its senses, naturally evokes the domain 'geology'. The same can also be said for the word 'continent' in Example 3b. But one annotator, a geology expert, consistently took the sense 'a particular spatial arrangement' for the word 'formation' in this expression. His decision was based on the strict interpretation of 'geological formation' as a domain-specific concept also described in https://en.wikipedia.org/wiki/Geological_formation and reinforced by the fact that 'geological formation' in PWN has 'physical object' as its hyperonym, not 'formation' (as a process).

Another issue identified in this small experiment was that of an annotator consistency. In the definition of 'male bonding' and 'female bonding,' the word 'formation' appears in a very similar way ('the formation of a close personal relationship between men/women'), but one of the annotators was not consistent in the annotation of its sense in the two glosses. Finally, Tables 1 and 3 show that some annotators have already identified missing senses in PWN (column N).

6 Previous Work

The recognition that PWN contains a substantial amount of knowledge within its glosses was made clear in (Clark et al., 2008a,b). These articles describe the work on some of the standoff files distributed in the PWN website,[6] including the 'logical forms' of the glosses. The authors also mention the use of the 'logical forms' produced years before by another team at the University of Texas.[7] In the Extended Wordnet (Harabagiu et al., 1999), the disambiguation of the glosses was done automatically over the PWN 2.0 glosses. Although their initial plan was to "develop a tool that takes as input the current or future versions of WordNet and automatically generates an extended WordNet that provides several important enhancements intended to remedy the present limitations of WordNet", the project does not seem to be maintained anymore. In (Clark et al., 2008b) the authors reported that the logical forms generated at that stage are not of high quality in general. Further use of the Extended Wordnet was reported in (Castillo et al., 2004).

The most relevant work to our present effort is the original Princeton WordNet Gloss Corpus project, our starting point.[8] Unfortunately, we are not aware of any publications resulting from the project except the README file distributed with the data and the annotators' guidelines. [9]

Here we emphasize the manual process, focus on the creation of a semantic concordance of PWN glosses and PWN itself in the same lines of (Miller et al., 1993). We are also following as close as possible the directives devised by the Princeton team when they started the original Princeton Wordnet Gloss Corpus. Similar to (Moldovan and Novischi, 2004), our primary goal is the development of better word-sense disambiguation methods and algorithms that can take advantage of the annotated glosses for better results on a domain-specific corpus, such as the one described in (Rademaker, 2018).

[6] https://wordnet.princeton.edu/download/standoff-files

[7] http://www.hlt.utdallas.edu/~xwn/

[8] http://wordnetcode.princeton.edu/glosstag.shtml

[9] The authors would like to thank Christiane Fellbaum for sharing the guidelines with us.

7 Conclusion

In this paper we describe our resuming of the Princeton WordNet Glosstag Corpus project.[10] We have assembled a team and created an annotation interface, and have begun our work. As put by (Miller et al., 1993), the semantic annotation of corpora helps improve both the coverage and the precision of the semantic resource being used in the annotation. This work is thus part of our effort in expanding and improving WordNet-like resources in an application-driven and domain-specific way, initially focusing on oil & gas domain applications.

Besides a continuous annotation effort, future work mostly involves improvements in the annotation interface[11] and in annotation methodologies. With respect to the annotation tool, we intend to start supporting the sequential annotation style discussed in Section 2, and to improve its performance. The methodological work involves developing processes for the revision of syntactic annotation (part-of-speech tags, lemmas, and MWE tagging) and for updating the corpus when the underlying WordNet changes. Additionally, we also intend to develop querying and visualization tools to support the annotation and the WordNet's expansion work.

[10] https://github.com/own-pt/glosstag

[11] https://github.com/own-pt/sensetion.el

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational Linguistics and Intelligent Text Processing*, pages 136–145, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pierpaolo Basile, Marco de Gemmis, Anna Lisa Gentile, Pasquale Lops, and Giovanni Semeraro. 2007. Uniba: Jigsaw algorithm for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 398–401, Prague, Czech Republic. Association for Computational Linguistics.

Francis Bond, Luís Morgado da Costa, and Tuan Anh Lê. 2015. Imi — a multilingual semantic annotation

environment. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 7–12, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Mauro Castillo, Francis Real, Jordi Asterias, and German Rigau. 2004. The talp systems for disambiguating wordnet glosses. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 93–96, Barcelona, Spain. Association for Computational Linguistics.

Peter Clark, Christiane Fellbaum, and Jerry Hobbs. 2008a. Using and extending wordnet to support question-answering. In *Proceedings of the 4th Global WordNet Conference (GWC'08)*.

Peter Clark, Christiane Fellbaum, Jerry R Hobbs, Phil Harrison, William R Murray, and John Thompson. 2008b. Augmenting wordnet for deep understanding of text. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 45–57. Association for Computational Linguistics.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Sanda M. Harabagiu, George A. Miller, and Dan I. Moldovan. 1999. Wordnet 2: a morphologically and semantically enhanced resource. In *Proceedings of SIGLEX99: Standardizing Lexical Resources*, pages 1–8.

Adam Kilgarriff. 1997. "i don't believe in word senses". *Computers and the Humanities*, 31(2):91–113.

Claudia Leacock, Geoffrey Towell, and Ellen Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the workshop on Human Language Technology*, pages 260–265. Association for Computational Linguistics.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

Dan Moldovan and Adrian Novischi. 2004. Word sense disambiguation of wordnet glosses. *Computer Speech & Language*, 18(3):301–317.

Alexandre Rademaker. 2018. Challenges for information extraction in the oil and gas domain. In *Proceedings of the XI Seminar on Ontology Research in Brazil (ONTOBRAS)*, São Paulo, Brazil.

Ewa Rudnicka, Francis Bond, Łukasz Grabowski, Tadeusz Piotrowski, and Maciej Piasecki. 2019. Sense Equivalence in plWordNet to Princeton WordNet Mapping. *International Journal of Lexicography*.