# Aligning Wiktionary With Natural Language Processing Resources

**Ben Casses**
Western Carolina University
Cullowhee, NC 28723
bncasses1@catamount.wcu.edu

## Abstract

Recently, significant progress has been made towards mapping various natural language processing resources together in order to form more robust tools. While most efforts have gone towards connecting existing tools to each other, recently several projects have involved aligning the popular NLP resources to open collaborative projects such as Wikipedia. Such alignments are promising because they link the specific but frequently narrow NLP data to high coverage open resources. This project explores the effectiveness of some variations of the Lesk Algorithm in connecting specific Wikipedia senses to corresponding senses in other NLP resources. The purpose of this project is to present a potential method of semiautonomous alignment for Wiktionary that will serve to augment other NLP resources.

## 1 Introduction

Existing natural language processing (NLP) resources can be used in each step of the language comprehension task. Some resources also contain mappings to others. Since some extensive tasks like language comprehension involve the use of multiple resources, such mappings can be very useful. This project will study some methods that may help to reinforce or partially automate mappings between some of these resources through the use of domain based disambiguation and gloss comparisons as in the Lesk Algorithm (Lesk, 1986). The following introduces some of the popular NLP resources and discusses their relative advantages.

### 1.1 FrameNet

FrameNet is a collection of semantic frames. It contains detailed breakdowns of the function and contextual meaning of a given verb along with all possible participating semantic participants and examples. FrameNet is organized into a hypernym-hyponym hirearchy with more specific terms inheriting structure from their hypernyms. FrameNet's exhaustive detail into semantic participants makes it valuable for studying distance relationships between frames. *Robbery*, for example, inherits directly from *Committing_crime*, indirectly from *Misdeed* and uses *Theft*. [1]

### 1.2 OntoNotes

OntoNotes is a collaborative project that aims to produce "richer model of text meaning" (Hovy et al., 2006). OntoNotes contains 2,445 verb entries organized into different senses with brief definitions and examples. OntoNotes is the most externally connected of the sense based NLP resources, it contains mappings to FrameNet, PropBank, VerbNet and WordNet. [2]

[1] http://framenet.icsi.berkeley.edu/

[2] http://verbs.colorado.edu/html_groupings/

### 1.3 PropBank

PropBank contains information on 5,384 verbs. Each verb entry is divided into different syntactic structures based on common use. Structures are subdivided into components *referred to as arguments*. Propbank has value as a parsing and a disambiguating resource. A verb in a given sentence could be matched to one syntactic structure by its surrounding arguments. Once matched, the roles of the arguments are disambiguated according to the roleset. PropBank contains mappings to VerbNet. [3]

### 1.4 VerbNet

VerbNet is an online database of verbs. Verbs are collected into 274 Levin (1993) style verb classes containing relevant semantic and syntactic information. The value of VerbNet lies in its depth of study into its verb members and their relationships. A given verb may be considered synonymous with its fellow members and hyponymous to its class. This can be useful for the purposes of translation and comprehension if a given verb in VerbNet is not understood but one of its fellow members is. VerbNet contains mappings to FrameNet, OntoNotes, PropBank and WordNet.[4]

### 1.5 Wiktionary

The effectiveness of Wiktionary in NLP tasks has already been established by Zesch and Müller (2008). At the time of this writing, English Wiktionary had "1,813,199 entries with English definitions from over 350 languages" (wik, 2010). The usefulness of Wiktionary is in its open, collaborative nature. Because anyone can contribute to Wiktionary, it is far more encompassing than any project developed by an individual or more structured group could be. It is also current, while other dictionaries must be revised and updated occasionally, Wiktionary entries are constantly being appended as the nature of the language changes. *The author observed the number of entries increasing by 20,000 over a period of three weeks.* Wiktionary's advantage can also be a disadvantage. Its open nature leads to format inconsistencies between definitions that make automated processing difficult. The potential also exists for incorrect or deliberately misleading entries such as those known to have occurred in Wikipedia (Snyder, 2007) (Chesney, 2006).

### 1.6 WordNet

WordNet is a long term project hosted by Princeton University. Entries in WordNet are organized into parts of speech and distingushed by sense. Wordnet is highly interconnected with each term referring to related terms including hypernyms, hyponyms, troponyms, frames and peers. Because of its interconnectedness, WordNet is useful for sense disambiguation. An unknown term could be generalized to its hypernym, for example: if the term "shuffle" is not understood, its hypernym, "walk" may be. [5]

## 2 Related Work

Several different mappings between existing NLP resources already exist. Combinations of resources have been created in different ways with most projects involving multiple alignment techniques to improve accuracy.

- Through common fields shared by both tools. (Loper et al., 2007) (Pazienza et al., 2006) (Giuglea and Moschitti, 2004)
- Through mutual restrictions, where a given entry in one database could match with a given entry in another database, but restrictions prevent this combination, thus narrowing the possiblilities. (Shi and Mihalcea, 2005) (Loper et al., 2007) (Giuglea and Moschitti, 2006)
- Through frequency analysis such as greatest numbers of common synonyms. (Shi and Mihalcea, 2005) (Giuglea and Moschitti, 2006) (Giuglea and Moschitti, 2004)

[3] http://verbs.colorado.edu/propbank/framesets-english/

[4] http://verbs.colorado.edu/verb-index/index.php

[5] http://wordnetweb.princeton.edu/perl/webwn

- Through supervised learning techniques where connections are trained. (Loper et al., 2007) (Giuglea and Moschitti, 2004)

- And through manual mapping of connections. (Pazienza et al., 2006) (Shi and Mihalcea, 2005).

In general, these methods could be divided into three categories: manual methods that require human control, statistical methods that involve matching around comparisons, and learning methods that involve machine learning techniques.

Zesch and Gurevych (2010) compared the effectiveness of several different semantic relatedness analysis methods. They considered four distinct semantic relatedness measures:

- Path based, where the distance between two terms in a graph is considered *edge counting*.

- Information Content based, where the number of documents containing both terms is considered.

- Gloss based, where the quantity of common words in each term's gloss is considered.

- Vector based, where vectors are constructed from multiple documents and the frequency of occurrence of the given term in each document is considered.

Recently work has begun on mapping and utilizing the collaborative resources such as Wiktionary and Wikipedia. The Ubiquitous Knowledge Processing Lab has developed api's for both Wiktionary[6] and Wikipedia[7] and made use of them as NLP resources (Zesch et al., 2008).

## 3 Problem Definition

The task of aligning NLP resources involves several issues. Each of the NLP resources considered in the introduction have different strengths, but some might be more difficult to align to Wiktionary or less effective when combined. There are structural concerns where two resources do not follow the same layout or present the same information. There are also potential problems where two resources do not expose the same granularity. Only two resources containing the same information could have a one-to-one cardinality.

### 3.1 Structure and Organization

Not all of the NLP resources discussed in the introduction follow the same structure. WordNet, for example is categorized by frames where each frame contains multiple syntactic structures with interchangable member verbs. Entries in PropBank, however, are centered around a specific predicate or term and divided into usages.

Wiktionary is organized by term with each term divided into different senses. It will be more meaningful to map the Wiktionary senses to the senses of another NLP resource that is organized similarly. Of the resources organized in this way that were discussed earlier, OntoNotes offers the most advantages due to its interconnectedness. A given Wiktionary sense mapped to OntoNotes could be followed to each other resource that OntoNotes is already connected to.

### 3.2 Granularity

Term-to-term matching is generally trivial, involving only a mutual lookup for a given term. Sense-to-sense mapping becomes more difficult for several reasons. A sense-to-sense connection between two different resources indicates that both senses could be considered to be the "same", but the two senses will generally not contain the same information. For example, one Wiktionary sense of the verb *make*, "To indicate or suggest to be", was aligned to "cause to become, or to have a certain quality" in OntoNotes even though both senses contain different information. Additionally, because of the different ways these resources were constructed, they feature a different degree of granularity around their terms. *Arrive*, for example, has three senses in Wiktionary and one in Ontonotes. Granularity differeneces indicate that the cardinality of this mapping will be many-to-many.

[6] http://www.ukp.tu-darmstadt.de/research/software/jwktl/

[7] http://www.ukp.tu-darmstadt.de/research/software/jwpl/

### 3.3 Size

Wiktionary is a rapidly growing resource. With the observation that on Wiktionary there may be up to 1,000 new definitions added and many existing definitions modified daily, manual mapping is not a feasable method of alignment. Results would quickly become incomplete or inaccurate. Since it is universally editable, it is not guaranteed that all Wiktionary entries follow the same layout. In some cases formats are inconsistent, information may also be missing, incorrect or presented out of order. Inconstistency makes automated retrieval and matching difficult.

## 4 Proposed Solution

Due to the differences in content and the lack of existing connections, a gloss based method was selected for aligning Wiktionary to OntoNotes. One advantage of using a gloss method for comparison is that it can act in a naïve fashion. No sense content need be understood or categorized, a sense is just a "bag of words". Even in occasions where sense parsing does not work as anticipated due to format inconsistency, for example, gloss comparisons will not suffer significantly. A missed tag from one resource, such as *</title>*, is highly unlikely to have a correspondance in another resource.

The gloss comparison in this project will consider common *n*-grams between the compared documents or senses as an indication of similarity. Consider the existence of the uncommon unigram *lathe* in table 1.

| Wiktionary |
|---|
| turn: To shape (something) symmetrically by rotating it against a stationary cutting tool, as on a **lathe**. |
| OntoNotes |
| Shape by rotating, Examples: After purchasing the wood, I ripped all the pieces to length, then turned the legs on a motorized **lathe**. |

Table 1: aligned Wiktionary [10] and OntoNotes [11] senses of turn

*Lathe* does not appear in any other senses from either resource therefore it indicates a relationship between these two senses.

## 5 Experiments

Two different experiments were performed, the first involved comparing Wiktionary senses directly to OntoNotes senses. Both experiments attempted to determine the most appropriate sense-to-sense mapping for fifteen verbs, show in table 2, that were selected from three sets of driving directions from Google Maps[12], Yahoo Maps[13] and MapQuest[14].

For each comparison between two documents, or one document and a corpus, a similarity value was computed as the sum of the values of each instance of each $n$-gram in common between the two sources. The comparison with the greatest similarity value was considered to be the correct choice. Ties involving a correct answer were considered incorrect as they did not effectively disambiguate.

### 5.1 Direct Method

Let $N$ denote the set of senses $\{N_1, N_2, N_3, ..., N_j\}$ for a given term in OntoNotes and let $W$ denote the set of senses $\{W_1, W_2, W_3, ..., W_k\}$ for the same term in Wiktionary. The matching senses are those with the greatest similarity. To compute similarity, let each sense $S_a$ from $W$ and $N$ contain a set of $n$-grams $\{S_{a1}, S_{a2}, S_{a3}, ..., S_{ap}\}$. The similarity between two glosses is equal to the sum of the values of the intersection of their terms. $Sim_{N_x,W_y} = \sum_{r \in I} V(r)$ where $I = N_x \cap W_y$ and $V$ is a value filter discussed in 5.4. This process derived a closest match $x \in N$ for each sense $y \in W$.

### 5.2 Transitive Method

The second method, the transitive comparison, involved first comparing glosses for a given term from Wiktionary, $W$ and OntoNotes, $N$ to a set

[12] http://maps.google.com/

[13] http://maps.yahoo.com/

[14] http://www.mapquest.com/

of $n$-grams in a domain specific corpus $C$. As with the previous method, similarity scores were calculated based on the values of the intersection of terms, but in this case, a best similarity score was derived separately for OntoNotes and Wiktionary, $Sim_{N_x} = \sum_{r \in I} V(r)$ where $I = N_x \cap C$ and $Sim_{W_y} = \sum_{r \in I} V(r)$ where $I = W_y \cap C$. The senses $N_x$ and $W_y$ with the greatest similarity scores to $C$ were considered the correct sense for the domain and matched to each other. This process derived a single closest disambiguated sense pair $(y \in W, x \in N)$ for each term.

| | | |
|---|---|---|
| arrive | avoid | bear |
| become | continue | enter |
| go | head | keep |
| make | merge | start |
| take | turn | welcome |

Table 2: driving verbs

The Wiktionary source text was downloaded from a publicly available data dump [15]. The OntoNotes source text was gathered from the OntoNotes [16] site.

### 5.3 Testing

Volunteers were initially asked to select only the most appropriate sense to the driving domain from Wiktionary and OntoNotes for each of the fifteen verbs in table 2. Of the thirty selections, the volunteers were found to be in agreement on a single sense only 58% of the time. Because this amount of disagreement either indicates a many-to-one correspondance or some incorrect selections by the volunteers, it was decided that further disambiguation was necessary for testing. The volunteers were combined into a single group and asked to decide on a most appropriate sense *or more than one in the case of a deadlock*. The volunteers were then asked to select the most appropriate OntoNotes sense for 63 additional senses in Wikipedia from the fifteen verbs.

After the committee decisions, there were only two situations that involved one Wiktionary sense mapping to multiple OntoNotes senses. For analysis in these cases, each of the selected OntoNotes senses was considered to be equally correct, that is, if the sense programmatically determined to be the most approprite matched any of the *correct* senses, it was considered a success.

There was one trivial situation where a term had only one sense. The verb *arrive* had only one OntoNotes sense, making determination trivial and potentially skewing results, therefore there were 72 non-trivial determinations to be made in the direct gloss comparisons and 30 determinations to be made in the transitive comparisons.

### 5.4 Filters

To avoid false matches from insignificant, common words such as *the, of, is*, a word frequency list was formed from the Wiktionary data dump. First, all tags were removed leaving only text. Next, a list of the frequency of appearance of each individual word was created. This list was used to construct two filters. The first filter, *gentle* assigned a value $v = -Log_2 \frac{f}{F+1}$ to a given term found $f$ times in Wiktionary where the greatest frequency for any word was $F$. The second filter, *severe*, was created to determine if more restrictive scoring achieved any better results. Values for the second filter were set to $v = -Log_2 \frac{f \cdot 20}{F+1}$ and assigned to $10^{-10}$ if zero or less.

Two different metrics were explored for evaluating $n$-grams. The first was the sum of the values of the terms $V = v_1 + v_2 + ... + v_n$, the second was a geometric mean, $V = \sqrt[n]{v_1 \cdot v_2 \cdot ... \cdot v_n}$. In both cases, the filters created greater emphasis on $n$-grams containing rare words. This allowed the trigram *the cat is*, for example to be slightly more significant than *cat* alone but not as significant as *sneaky orange cat*. Only unigrams, bigrams and trigrams were counted.

[15] http://dumps.wikimedia.org/enwiktionary/latest/

[16] http://verbs.colorado.edu/html_groupings/

### 5.5 Direct Comparison

The direct comparisons involved evaluating a given sense in Wiktionary against each sense for the same term in OntoNotes. Within Wiktionary, for example, there are 13 senses for the verb *make*. Each of these senses was compared to the 17 senses for *make* in OntoNotes. For each Wiktionary sense, the sense pair with the greatest similarity value was considered to be the correct mapping, so for *make* there were 13 possible correct mappings. These results were compared to the direct comparisons made by the volunteers.

### 5.6 Transitive Comparison

The transitive comparisons involved evaluating each sense from a given resource against a domain specific corpus for disambiguation assistance. Only the domain relevant sense decided by the volunteers was considered to be the correct answer, so there was only one correct sense for each verb in each resource.

Two corpora were created for testing transitive comparisons. The first was formed from the sources for the 15 verbs, driving directions from Google, MapQuest, and Yahoo Maps. The second corpus was created from Google searches for "driving +turn +go +keep +bear".

## 6 Results

For each method, the results were considered compared to the volunteer selections. If the sense pair with the greatest similarity score matched the sense pair selected by the volunteers, the matching was considered correct.

### 6.1 Direct Comparison Results

For the direct comparisons, there were 72 non-trivial matches. Four sets of experiments were performed: Sum $n$-gram and Geometric Mean $n$-gram evaluations were performed with the *gentle* and *severe* filters. Although they resulted in slightly different sets of correct answers, neither evaluation or filter method did significantly better. Results are shown in table 3 as accuracy percentages out of 72 possible matches.

For a baseline comparison, a naïve selection was developed that involved matching all Wiktionary senses to the first OntoNotes sense for a given term. The naïve achieved 35% accuracy.

| | Sum Eval | Geom Mean |
|---|---|---|
| gentle filter | 46% | 46% |
| severe filter | 47% | 46% |

Table 3: Direct Comparisons

### 6.2 Transitive Comparison Results

For the transitive comparison, there were 29 non-trivial matches. Eight comparisons were performed. Sum $n$-gram and Geometric Mean $n$-gram evaluations were performed with the *gentle* and *severe* filters in comparisons to both corpora. In the transitive case, the *severe* filter performed slightly better than the *gentle* filter. Results are shown in table 4 as accuracy percentages out of 29 possible matches.

A naïve method, selecting the first sense was used for a baseline comparison to the transitive method. The naïve method achieved 48%.

| original corpus | Sum Eval | Geom Mean |
|---|---|---|
| gentle filter | 48% | 48% |
| severe filter | 48% | 52% |
| Google search | Sum Eval | Geom Mean |
| gentle filter | 48% | 48% |
| severe filter | 62% | 59% |

Table 4: Transitive Comparisons

Discarding all but the most effective methods, these results, 47% and 62% can be compared to (Lesk, 1986) 50%-70% and (Banerjee and Pedersen, 2002) 25% for verbs and 74%-78% for (Kilgarriff and Rosenzweig, 2000).

## 7 Problems

Although each method performed as good as or better than naïve selection, 72 potential matches among 15 verbs may not be enough to make any determination about the validity of the hypothesis. Further tests are needed to reinforce the effectiveness of these methods.

Some OntoNotes senses were troublesome for gloss comparisons. The fifteenth sense of *go*, for example, "miscellaneous idioms..." is a catch-all sense containing more text, therefore more inadvertent matches, than other senses. The seventeenth OntoNotes sense of *make*, "other verb particle constructions", is informative to a human reader, but contains little information for gloss comparisons.

It is likely that a single document may use the same word in two different senses. While this was not the case with the transitive comparison corpora used, it could cause confusion in future experiments.

## 8 Conclusion

These results are far better than random, 28% for transitive and 46% for direct comparison. The most accurate mapping method found was 62% for the *severe* filter sum evaluation on the Google search corpus. While both the direct and transitive methods achieved results slightly better than their corresponding naïve methods, 14% and 17% better respectively, the results are not strong enough to regard these experiments as effective stand alone semiautonomous alignment methods.

It appears that the neither method of evaluating $n$-grams performed significantly better at disambiguating. It is possible that false matches, $n$-grams that correspond to the incorrect sense, were more significant to the outcome than evaluation methods. This could explain why changing the filter caused more of a difference than changing the evaluation method.

## 9 Future Work

A larger selection of verbs to disambiguate should help to better establish the accuracy of this method. Additionally, a different domain for transitive comparisons would further prove the possibilities of this method.

The value of the most likely sense, whether correct or not, often stood out strongly from the other senses, sometimes differing in value by an order of magnitude. This could indicate that the analysis method may be "fooled" by false matches. If these false matches could be isolated and disregarded somehow, results might improve.

The corpus, in the case of the transitive mapping, also caused some false matches, for example, the phrase "crossing into NEBRASKA" in the corpus and the example *...water turned into ice...* in OntoNotes caused *turn* to match to the *become* sense in one experiment. Issues like these might be resolved by scaling the values of $n$-grams based on their distance in the document from the term being considered. The text corpus for comparison could be refined to a domain keyword list which would eliminate some extraneous terms.

The *severe* filter, which performed better than the *gentle* filter, was created after the *gentle* filter in response to the determination that insignificant words still had too much influence. Perhaps a stronger filter could produce better results.

There are several possiblities for augmenting this gloss comparison method. Second order comparisons as used by (Banerjee and Pedersen, 2002) might be beneficial. Additionally, it has been suggested that syntactic limitations from a given sense could be considered in disamgibuation.

Verbs containing troublesome senses as mentioned in the Problems section cannot be aligned through gloss comparisons. In future experiments, such verbs should be discarded from the test set.

## References

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. pages 117–171.

Thomas Chesney. 2006. An empirical examination of wikipedias credibility.

Ana-maria Giuglea and Ro Moschitti. 2004. Knowledge discovering using framenet, verbnet and propbank.

Ana-maria Giuglea and Ro Moschitti. 2006. Semantic role labeling via framenet, verbnet and propbank. In *In Proceedings of COLING-ACL*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 57–60, Morristown, NJ, USA. Association for Computational Linguistics.

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for english senseval.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA. ACM.

Beth Levin. 1993. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago and London.

Edward Loper, Szu ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In *In Proceedings of the 7th International Workshop on Computational Linguistics*.

Maria Teresa Pazienza, Marco Pennacchiotti, Fabio Massimo Zanzotto, and Via B. Arcimboldi. 2006. Mixing wordnet, verbnet and propbank for studying verb relations.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In Alexander F. Gelbukh, editor, *CICLing*, volume 3406 of *Lecture Notes in Computer Science*, pages 100–111. Springer.

Johnny Snyder. 2007. Its a wiki-world utilizing wikipedia as an academic reference.

2010. Wiktionary english version main page.

Torsten Zesch and Iryna Gurevych. 2010. Wisdom of crowds versus wisdom of linguists - measuring the semantic relatedness of words. *Natural Language Engineering*, 16(01):25–59.

Torsten Zesch, Christof Mller, and Iryna Gurevych. 2008. Using wiktionary for computing semantic relatedness. In *In Proceedings of AAAI*.