Multilingual Dictionary Linking and Aggregation: Quality from Consistency

Kun Ji, Shanshan Wang, and Lauri Carlson

University of Helsinki,
Department of Modern Languages
{kun.ji,shanshan.wang,lauri.carlson}@helsinki.fi

Abstract. The growth of Web-accessible dictionaries and term data has led to a proliferation of platforms distributing the same lexical resources in different combinations and packagings. Finding the right word or translation is like finding a needle in a haystack. The quantity of the data is undercut by the doubtful quality of the resources.
Our aim is to cut down the quantity and raise the quality by matching and aggregating entries within and across dictionaries. In this exploratory paper, our goal is to see how far we can get by using information extracted from multiple dictionaries themselves. Our hypothesis is that the more limited quantity of data in dictionaries is compensated by their richer structure and more concentrated information content. We hope to take advantage of the structure of dictionaries by basing quality criteria and measures on linguistic and terminological considerations.
The plan of campaign is to derive quality criteria to recognise well-constructed dictionary entries from a model dictionary, and then attempt to convert the criteria into language-independent frequency-based measures. As a model dictionary we use the Princeton WordNet. The measures derived from it are tested against data extracted from BabelNet.

Keywords: Information extraction, Quality checking, Aggregation, Merging, Linked data, Edit distance

1 Introduction

Interactive Web and crowdsourcing have produced easily accessible lexical resources of unprecedented size. Lexical resources such as WordNet or Wiktionary are complemented by encyclopaedic data collections such as Wikipedia and Wikidata. The quantity of the data brings along problems of quality, such as errors, duplication and unclear provenance. Automatic methods including machine learning techniques are called on to manage the wealth, but may also contribute to the disorder.

Typical dictionary data categories, or fields, differ in availability, unambiguity and information potential. These three aspects often vary inversely: word labels are abundant and simple, but polysemous; semantic relations are unambiguous and informative, but scarce; subject field classifications and glosses have great information potential that is hard to make precise. We must combine different properties and vary our methods according to type.

This position paper introduces our line of research (cf . [1]) that tries to develop language-independent linguistically-motivated distributional methods for quality checking and aggregating such linguistic linked data. We first illustrate our approach with a selection of the kind of quality criteria we have in mind. As an example of such a measure, we describe a simple distance measure which is a variant of Levenshtein edit distance [2]. The measure is tested against labels, subject fields, and glosses extracted from multilingual dictionary BabelNet [3].

The experiments indicate that a flat edit distance measure is less suited for longer pieces of text. We are working on a more sophisticated language model that takes into account the linguistic structure of glosses.

The rest of the paper is structured as follows: Section 2 discusses related work. Section 3 investigates candidate indicators/properties for quality checking and aggregation of multiple dictionary data. Section 4 compares these properties and describes the frequency-based distance measure. Section 5 describes our progress implementation and evaluation. Sections 6 discusses the results of our work. Section 7 presents our conclusions future work plan.

2 Related Work

Ide and Veronis [4] argued that dictionaries have too little information for extracting from knowledge bases. That is not our task: basically, we just want dictionaries with fewer errors and duplicates. It remains true that dictionary checking may benefit from external information sources. The hard part is to avoid that these introduce more noise than they help suppress.

Navigli and Ponzetto compile BabelNet [3] using word sense disambiguation and machine translation as external sources. WordNet and Wikipedia are linked by a mapping between WordNet senses and Wikipage titles. Missing translations are collected from Wikipedia inter-language links and by machine translating occurrences of the labels within sense-tagged corpora. They report 82 percent mapping accuracy. We want to locate and fix the remaining errors.

Eckard et al [5] match a French dictionary with a machine translated French WordNet looking for hypernym relations, using manually prepared regex patterns to parse dictionary definitions.

Semantic relatedness (SR), more generally, measures how much two (strings of) words or concepts are related, counting all kinds of relations between them. Zhang et al. [6] present a hybrid SR method that generates a connection graph between labels using WordNet semantic relations and Wikipedia contexts and measures semantic relatedness by the density of the graph between two labels. Semantic similarity can be indirectly measured by semantic relatedness. It may thus bring useful evidence for our task, which is dictionary alignment and aggregation. Token level edit distance is known as WER (word error rate) in speech recognition and machine translation research [7].

3 Quality Checking for Dictionary Merging

Dictionaries are a good case of both the availability and need of matching and aggregating Web accessible data. There are a legion of mono- and multilingual dictionaries, glossaries, thesauri and other vocabulary collections in the Web, some but not all in RDF, some public and collectively maintained, many commercial but openly accessible for querying. This multiplicity is also an encumbrance. In language technology, one of the most common feature requests from human translators is ways to simplify the search of equivalents in the host of available sources.

Besides explicit URLs, dictionaries abound in implicit internal and cross-dictionary links, created by shared labels (words, collocations), subject field classifications, glosses, grammar and other properties. By aggregating dictionary entries, we also implicitly address the problems of (i) identifying valid such links (ii) discarding misleading and duplicated links, and (iii) making useful links explicit.

3.1 Terminology

To begin with, we define here some of the key concepts of our dictionary ontology.

By a *label* we mean a language-identified baseform (lemma), represented in RDF as `"base"@lang`. A monolingual or multilingual dictionary minimally generates a cover (set of possibly overlapping subsets) of the labels. The cover represents the neighbourhoods generated by a synonym or equivalent relation. The members of the cover are called *synsets*, or equivalent sets (*eqsets*) if the dictionary is multilingual. An Eqset is a multilingual synset. Separated by language codes, eqsets form disjoint unions of synsets. Synsets/eqsets can be seen to represent concepts or meanings.

A *sense* is a pairing of a label and a synset that the label is a member of. The more synsets a label belongs to, the vaguer it is. A label is n-way polysemous if it belongs to n synsets. Dually, the more members a synset has, the wider its meaning is. Special language labels are less polysemous than general language labels. Ideally, terms should be monosemous (per subject field). Polysemy is said to happen between subject fields, which supplies another consistency test. To estimate if a label is a term, one may check the size of its synset. To check if a term has been translated by a general language label, compare the sizes of their synsets.

Subject field headings show which domain or subject field a specific term belongs to. When the same label appears in different meanings, subject field classifications are used to distinguish meanings. A *gloss* is the definition or explanation associated to a label, which provides direct meaning of the concept in order to check if concepts are same or not. *Hypernyms* and other semantic relations serve the same purpose, as do *part of speech* and other grammar categories.

Synsets and eqsets are target units that we reconstruct in our dictionary alignment. Labels, subject fields, glosses and other indicators are properties from which our distance measure to be inferred. In this paper, we restrict our attention to labels, subject fields, and glosses.

Synsets/eqsets may overlap because labels may belong to many synsets, by way of vagueness (synonymy is not exact, synset boundaries are negotiable) or polysemy (a label may belong to different but semantically related synsets). A synset (eqset) is thus a syntactic representative of a meaning. The relation "synonymy" or "equivalence" means "in the same neighbourhood". It is an equivalence relation within the synset, but not transitive through shared members of overlapping synsets.

Hence overlapping synsets cannot be merged in general. Even if consistent, the conjoined synset may be narrower than the originals, and the inferred equivalences have less application than its premises. This is why WordNet translations cannot be just merged into the synsets that they translate. We need to find criteria for when synsets are safe to merge and what is the risk.

Translation equivalences are more informative than monolingual synsets because of mismatches between languages. Ambiguous words in one language may have unambiguous equivalents in another. Synonyms and hypernyms that are lexical in one language might be phrasal in another. This is particularly true about direct translations of WordNet to another language, as lexical gaps in the other language are often filled by phrasal definitions or paraphrases.

3.2 Merging Dictionaries

To match two dictionaries, we may pool together the equivalent sets from two dictionaries (with some markings to tell where they came from) and test if the combined dictionary satisfies various quality criteria. In theory, we merge two dictionaries by merging best matched entries and apply the quality criteria to the result. In practice, merging and checking may happen interleaved.

A multilingual dictionary may list bigger or smaller translation equivalences (multilingual synsets) depending on the precision and completeness of the translations. An optimal translation equivalence is to be reconstructed from many such smaller translation equivalences in the different sources.

To find if two translation equivalences (say, from different sources) can be merged, we may try to unify them by merging them and assuming some equivalences (e.g. based on label identity). The merge creates a lot of new equivalences. Some of the new equivalences may be explicitly present or *attested* in data, some not.

We may consider a binary translation pair like `"bank"@en - "pankki"@fi` as a base case of an eqset. In general, translation is not symmetric. By a translation norm, a translation should not add information (change a true text into a false text), but the opposite is not required: a translation may lose information in the source to satisfy other desiderata of the translation brief. The relation 'x entails y' is a partial order (transitive and antisymmetric). Symmetry can be restored by narrowing the context (e.g. with a subject field heading), or by giving up transitivity: the weaker notion 'y may translate x' is a symmetric non-transitive similarity relation. For constructing larger eqsets, the narrowing solution is preferable, so we should prefer binary translation pairs whose symmetry is attested in the data.

If we deconstruct WordNet synsets/eqsets into a binary relation of pairwise synonymies/equivalences between word senses, they form an equivalence relation whose quotient sets are the synsets/eqsets. When such equivalence pairs are all attested, it is easy to reconstruct synsets from them by just forming the partition of the set into strong components (strongly connected graphs, cliques). Each such component is a synset/eqset.

The clique test is at the strict end on a scale of attestedness. It construes synsets out of strongly connected components of the binary equivalent relation. When the evidence for synsets less complete, we may weaken the tests, with increased risk. Assuming transitivity and antisymmetry (the translation norm that translations are no narrower than original), we may check for equivalence by looking for cycles ([8]). Say we have three dictionaries, `en-fi, en-sv, fi-sv`. We may merge `en-fi, fi-sv` and check the result against `sv-en`:

```
"bank"@en < "pankki"@fi, "pankki"@fi < "bank"@sv, "bank"@sv < "bank"@en
```

Given multiple sources, there are two dimensions of degree of attestedness: number of distinct attested equivalences and number of duplicate attestations from different sources. Using such counts, we may construct quantitative variants of the consistency tests.

3.3 WordNet as Gold Standard

We test our criteria and measures on WordNet by deconstructing WordNet synsets down to senses or labels and then seeing to what extent we are able to reconstruct the synsets from them. The deconstruction of WordNet synsets can be done at word sense level or down to label level. Assume given the following synsets.

```
[ synset 1; label ’man’@en,’human’@en; property1 ’Noun’;  property2 ’human being’ ] .
[ synset 2; label ’man’@en; property1 ’Noun’; property2 ’adult male’ ] .
```

1. **Word senses** Sense deconstruction for the synsets produces three senses:

   ```
   [ sense 1; label ’man’@en; property1 ’Noun’; property2 ’human being’ ] .
   [ sense 2; label ’human’@en; property1 ’Noun’, property2 ’human being’ ] .
   [ sense 3; label ’man’@en; property1 ’Noun’, property2 ’adult male’ ] .
   ```

   Deconstructing synsets to word senses that inherit properties from the synsets, can we reconstruct the synsets by merging the senses? The answer is trivially yes if we retain the synset id or other key properties of synsets (like gloss). This exercise is more relevant when word senses come from different dictionaries.
2. **Word labels** Label deconstruction produces two labels:

   ```
   [ label ’man@en’; property1 ’Noun’; property2 ’human being’,’adult male’ ] .
   [ label ’human@en’ ; property1 ’Noun’, property2 ’human being’ ]
   ```

Distributing properties inherited from synset all the way to labels, can we reconstruct senses and synsets by splitting the labels, without knowing how the properties were clustered in the senses?

In (1) the senses keep properties together, whereas in (2) we lose the information regarding which properties go with which sense. In (2) there will be many more items to merge. In the example above, the three senses cannot be reconstructed from the label deconstruction since the ambiguity of 'man' has been lost. The three senses can be merged back to two synsets from sense deconstruction because the properties of the two senses agree. For other similar or more complicated cases, we try reconstruction at varied granularity – word sense, label sense or combined to find an optimal merging solution.

In the general situation of combining different dictionaries, what get merged are just such "senses", "terms" or "entries", which combine one or more labels plus some other properties. The risk is in merging labels or/and properties belonging to incorrect or repeated senses. The merger can lose information. Errors and duplicates may arise. The task is to aggregate the entries to obtain the most likely and meaningful synsets. Starting from a set of partial descriptions of shared meaning, we try to merge the descriptions into manageable clusters. We next look at some statistics on the different indicators in WordNet.

4 Comparing Properties

In the previous section, we presented criteria that may be used in aggregating synsets and eqsets. Our criteria depend on notions of identity or sufficient similarity among labels and other dictionary fields/properties, which is the topic of this section.

We have not yet touched the problem of matching similar but not identical properties. For other properties besides labels, such as glosses, matching is not straightforward. In the general case, we want to deal with graded measures.

The problem of matching two properties is not independent of matching the whole entries. Matching property contents is an argument for matching the entries, and vice versa. We set aside this complication for now.

4.1 Sharing of Labels between Synsets

The English WordNet RDF has about 100K synsets and 200K senses. It includes translations in 21 languages. The most complete one is Finnish (300K), with Malay, Japanese, Indonesian, and French next (over 100K each).

To appreciate our chances in the reconstruction, we studied how far labels alone go in measuring the similarity of synsets. Less than 0.1 percent of hypernymous synset pairs in the English WordNet share one or more labels. About 4 percent of hypernymous eqset pairs share labels in at least one language. This is another indication that translating the English WordNet creates redundant distinctions for the target languages. For random synset pairs, the corresponding percentages are one or two magnitudes smaller. So label sharing is a good, though rare indicator of synset similarity.

Listing 1 shows some of the closest WordNet synsets measured in shared labels and translation respectively. The first column is the number of shared labels, the next two columns are the synset ids, followed by a sample shared label and glosses for the two synsets.

Listing 1 Closest WordNet Synsets Measured in Shared Labels and Translation

```
 10 wn31:107741018-n wn31:112599160-n "mung bean"@eng ’mung seed’’mung plant’
  6 wn31:107137720-n wn31:107407761-n "scream"@eng  ’cry’’noise resembling cry’
  6 wn31:104647089-n wn31:104717403-n "severity"@eng ’excessive sternness’’hard to endure’

129 wn31:200825727-v wn31:200826456-v "admonish"@eng ’take to task’’censure severely’
 94 wn31:400046739-r wn31:400473918-r "extremely"@eng ’extreme degree’’extraordinary degree’
 81 wn31:200346415-v wn31:201654152-v "start"@eng  ’take first step’’get off the ground’
```

4.2 String Relationships in Hypernyms

A fraction of hyponymy relations are recognisable from their syntactic makeup as phrasal species terms, each composed of a hypernym denoting the genus and modifiers specifying differentia, for example *skilled workman* < *workman* . In the English WordNet, about one quarter of hypernym relations have this form, mostly phrasal verbs and special field terms. Another fraction are suffixal (in English, typically compounds), like *workman* < *man* . When all of the above types are included, 22 percent of hypernym relations contain at least one English substring relationship. Apparently, substring relationships are a useful indicator, but not strong enough alone.

4.3 Distance Measure

To have a quantitative measure for the distance between similar labels and other dictionary fields/properties, we implemented a language-independent character frequency-based edit distance measure. The same measure is designed to be applicable to subject field labels and glosses, possibly with different additional information sources and parameter settings.

Our distance measure is a two-level frequency weighted Levenshtein (edit) distance measure [2]. It is designed to be language-independent as far as feasible, using only information available in the dictionary itself. With this desideratum in mind, the measure derives edit costs (weights) from character and token frequencies extracted from the input data or imported from external sources.

4.3.1 Character-based Distance Measure for Comparing Labels

We first calculate Levenshtein edit distances between tokens, with edit costs weighted by frequencies of characters per string position. Specifically, character cost grows with the variety (number of different characters) per position and the information value (inverse frequency) of the character at the position.

As expected for English, early positions have more variation, while mid vowels and dental consonants predominate at endings.

```
0: v=1 w=2 o=12 c=6 h=3 r=1 b=1 f=4 l=4 k=2 ?=1 s=6 g=1 i=5 a=20 e=10 d=2 p=4 n=1 t=12
1: s=5 g=1 i=8 p=1 t=1 n=16 e=8 a=10 y=1 c=1 o=6 w=1 b=1 f=2 x=4 h=13 u=1 r=9
2: l=2 f=1 r=4 c=1 o=4 y=1 m=2 v=6 t=9 n=10 p=1 d=3 e=3 a=8 i=6 g=1 s=7
3: r=1 a=2 e=9 l=3 t=10 n=2 d=2 f=1 g=2 w=1 i=11 s=4 c=5 o=1 m=4
4: g=3 i=6 s=2 a=1 e=4 n=5 t=9 p=2 o=3 -=1 m=1 u=1 r=4 h=2 l=2 b=1
5: g=4 i=2 s=1 t=1 n=3 d=2 p=1 e=5 a=2 y=6 o=2 c=3 w=1 v=1 l=2 b=2 f=1 r=2
6: e=5 a=1 r=2 p=1 n=3 t=3 l=3 v=1 i=2 c=1 o=2 y=1
7: t=1 n=3 l=1 d=3 e=3 a=1 c=3 y=1 i=1 s=2
8: e=3 g=2 d=1 t=1 n=1
9: e=1 a=1 g=1 n=1
10: t=1 i=1
11: n=1 l=1
12: e=1 y=1
13: d=1
```

The tokens are normalised to types using token distances and token frequencies as guides. The assumption in this reduction is that the dictionary or lemma form is close in character distance to its inflections and derivations and no less frequent than them. If a synonym dictionary is supplied, it is used in the tokenisation, preferring types that occur in the dictionary. Also abbreviations and multiword phrases get tokenised with the dictionary if supplied.

The token distances obtained on the first level are used as token costs in another Levenshtein round that compares multi-token strings (terms, glosses, definitions etc.). This round uses a similar logic to the previous one, using position-sensitive type frequencies to weight edit costs. (The built-in assumption is that key terms occur early in glosses.) Besides the usual edit operations (addition, deletion, substitution) gloss distance adds permutation (by lowering the cost of substitutions of low-frequency terms if they are offset by an opposite substitution elsewhere).

To give more weight to low-frequency terms, the character-based token edit distances are scaled by token frequencies so that long edit distances to low-frequency, high-information tokens (terms) are stretched exponentially at the high frequency end and short distances correspondingly shrunk at the opposite end. Under this metric, the short end edit distances manage to single out inflectional and derivational relations between significant keywords.

```
0.025269 221.000000 reciprocating reciprocal
0.024468 214.000000 functional function
0.023897 209.000000 experiencing experience
0.022639 198.000000 features feature
0.016808 147.000000 characteristic characterized
0.015550 136.000000 organisms organism
0.013835 121.000000 interacting interaction
0.012920 113.000000 accomplishment accomplishing
0.008804 77.000000 substances substance
0.003544 31.000000 independently independent
```

4.3.2 Synonym-enhanced Distance Measure for Comparing Glosses

The character-based distance measure fails to capture similarities between glosses that use unrelated but synonymous words. To remove this limitation, we import semantic relatedness information from the dictionary itself. It was done here by lifting WordNet synset and hypernym relations to a semantic relatedness relation between labels. This construction is lossy in three ways: (1) the further apart two synsets are in hypernym hierarchy, the more loosely they are considered related; (2) the larger the synset, the vaguer its meaning (in general) – special language concepts tend to have fewer synonyms than vaguer or context dependent general language meanings; (3) precision falls with sense count: a polysemous label is less sure an indication of meaning than a monosemous one. We generate a fuzzy set of 1.3M semantically related pairs of labels from English WordNet, weighed by the above counts so that more precise synonymies have more bearing than fuzzier ones.

In sum, the synonym-enhanced gloss distance may correctly predict semantic distances between differently worded definitions of the same thing on the one hand, and definitions pertaining to different concepts on the other hand. Table 1 shows the gloss distance for term 'REMOVAL' and 'Removal':

Table 1. Gloss Distances for REMOVAL-Removal

distance	gloss 1	gloss 2
0.075714	wn31:100021914-n	REMOVAL
0.064758	wn31:100021914-n	Removal
0.027266	REMOVAL	Removal
0.000000	wn31:100021914-n	wn31:100021914-n

The glosses of the terms in Table 1 are:

- wn31:100021914-n "any substance such as a chemical element or inorganic compound that can be taken in by a green plant and used in organic synthesis"
- REMOVAL "The formal expulsion or deportation of a non-citizen from the United States when the non-citizen has been found removable for violating the immigration laws A person can be removed for overstaying a visa or for breaking laws including immigration laws"
- Removal "The expulsion of an alien from the United States based on grounds of either inadmissibility or deportability"

Table 2 is a truncated Levenshtein distance matrix for the glosses REMOVAL-Removal. Star marks a substitution, plus addition and minus deletion. The minimum edit path can be traced following pluses down, minuses to the right, and stars diagonally down right.

Table 2. Distance Matrix for REMOVAL-Removal

	The	expulsion	of	an	alien	from	the	United States	based	on
The	0.00*	-1.00	-1.58	-2.16	-2.90	-3.57	-4.20	-5.56	-6.29	-6.87
formal	+0.79*	+1.42	+2.00	+2.58	+3.32	+3.99	+4.62	+5.98	+6.71	+7.29
expulsion	+1.79	0.79*	-1.37	-1.96	-2.69	-3.37	-4.00	-5.35	-6.09	-6.67
or	+2.37	+1.37*	+1.95	+2.53	+3.26	+3.94	+4.57	+5.93	+6.66	+7.24
deportation	+3.53	+2.53*	3.11	2.88	3.40	4.07	4.70	6.06	6.79	7.37
of	+4.11	+3.11	2.53*	-3.12*	-3.85	-4.53	-5.16	-5.82	-6.55	-7.14
a	+4.64	+3.64	+3.06	+3.65*	+4.38	+5.06	+5.69	+6.35	+7.08	+7.67
non-citizen	+5.80	+4.80	+4.22	4.81	5.54*	6.22	6.85	7.02	7.75	8.33
from	+6.48	+5.48	+4.90	5.48	6.22	5.54*	-6.17*	-7.53*	-8.26*	-8.84*
the	+7.11	+6.11	+5.53	6.11	6.85	+6.17	+6.80	+8.16	+8.89	+9.47*

5 Evaluation

To evaluate our distance measure, we extracted the first 4130 synsets from BabelNet using the Java API for the BabelNet 3.6.1 Lucene index download.

5.1 Term Labels

We tested the character-frequency based distance measure on the first 1000 English language BabelNet labels in our extract, listing for each label its nearest neighbour in the set according to the measure. The synonym dictionary was not used here. 177 pairs were identical. 58 percent of the near neighbours came from the same synset. This may be compared to the probability of a random pair of labels coming from the same synset in our data (0.007).

```
"protein folding"  ~  "folding"
"Misfolded protein" ~ "Misfolded"
"Misfoldings" ~ "Misfolding"
"Incorrect protein folding" ~ "Incorrect folding"
"Singleblinding" ~ "Singleblind"
"Dryopithecini" ~ "Dryopithecidae"
"Double-entries" ~ "Double-entry"
```

The list deteriorates towards the end. This can be helped by adding a confidence index and threshold to cut off weakest cases. Another class of false positives are near ties like ”fathering”/”feathering”. They may be captured using a dictionary.

5.2 Subject Field Labels

To test the distance measure on subject field labels, we extracted 9890 BabelNet categories in English from our data and listed the nearest matching category label pairs in that set. In this run, we used the WordNet based synonym dictionary. An excerpt from both ends of the listing:

```
"Philosophy" "Epistemology"
"Polytheism" "Religion"
"Behaviorism" "Psychology"
...
"Knights Grand Cross of the Order of Merit of the Italian Republic"
~ "Grand Cross of the Order of Civil Merit"
"Central Committee of the Communist Party of the Soviet Union members"
~ "Heads of the Communist Party of the Soviet Union"
"People executed by the Bourbon dynasty of the Kingdom of France"
~ "Peers of France"
```

The result may be evaluated again by comparing the proportion of matches in our listing which classify the same synset (0.14) to the probability of a pair of category labels chosen at random to classify the same synset in our data (0.0004).

5.3 Glosses

The distance measure was tested on 1000 BabelNet glosses extracted from our data, with the nearest matching category label pairs listed in that set. In this run, we used the WordNet-based synonym dictionary. A few examples of the pairs judged nearest in the listing:

```
"Jane Seymour Fonda is an Academy Award-winning American actress,
model, writer, producer and political activist."
~ "American actress and activist"
"The down of birds is a layer of fine feathers found under the tougher exterior feathers."
~ "Soft, immature feathers."
"A dog sled is a sled pulled by one or more sled dogs used to travel over ice and through snow."
~ "A sled, pulled by dogs over ice and snow."
"Dreams are successions of images, ideas, emotions, and sensations that
occur involuntarily in the mind during certain stages of sleep."
~ "A series of mental images and emotions occurring during sleep"
"Duty is a term loosely applied to any action which is regarded as morally incumbent,
apart from personal likes and dislikes or any external compulsion."
~ "Nose."
```

The result may be evaluated again by comparing the proportion of matches in our listing which classify the same synset (0.18) to the probability of a pair of glosses belonging to the same synset in our data (0.002). There were just 14 identical pairs this time.

6 Discussion

As the last example above shows, there is room for improvement here. The measure is sensitive to length, while lengths of glosses may vary considerably. The effect of unequal length may be damped by truncating glosses (say the length of the shorter one), or by dropping high frequency tokens (articles, prepositions, auxiliaries). Quality aside, the Levenshtein measure is resource intensive on long glosses.

Obviously, edit distance is too unstructured for long glosses. We must improve the language model. We are currently working on a frequency-driven parser to compare glosses not as flat strings but as binary tree (dependency) structures, so as to cut down on pairwise comparisons of low-information tokens. Only the n-best edges from the parser are compared using the Levenshtein distance measure. Our parser is a frequency weighted chart parser using binary (dependency) grammar rules extracted from dictionary data.

7 Conclusion

To summarise, this paper presents quality criteria based on WordNet to merge linked lexical resources and to detect duplicates and errors in them. A distance measure to compare linguistic strings was described and tested on WordNet and BabelNet.

The first-round tests suggested how to improve the measure for longer glosses. That done, we may proceed with the WordNet deconstruction/reconstruction exercise to test our approach.

References

1. Wang. S and L. Carlson 2016. Linguistic Linked Open Data as a Source for Terminology - Quantity versus Quality. Proceedings of NordTerm 2015 (to appear).
2. Levenshtein, Vladimir I. (February 1966). "Binary codes capable of correcting deletions, insertions, and reversals". Soviet Physics Doklady. 10 (8): 707–710.
3. R. Navigli and S. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250.
4. N. Ide and J. Veronis. Extracting knowledge-bases from machine-readable dictionaries: Have we wasted our time? In Proc KB&KB93 Workshop, 1993.
5. Emmanuel Eckard, Lucie Barque, Alexis Nasr and Benoit Sagot, 2012. Dictionary-Ontology Cross-Enrichment. Using TLFi and WOLF to enrich one another. COLING Workshop on Cognitive Aspects of the Lexicon, 2012.
6. Z Zhang, AL Gentile, F Ciravegna 2011. Harnessing different knowledge sources to measure semantic relatedness under a uniform model. in Proceedings of EMNLP 2011. http://www.aclweb.org/anthology/D11-1092.pdf.
7. A Marzal, E Vidal. Computation of normalized edit distance and applications. IEEE transactions on pattern analysis and machine intelligence 15/9, September 1993.
8. Navigli, Roberto 2009. Using cycles and quasi-cycles to disambiguate dictionary glosses. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 594–602, Association for Computational Linguistics, 2009.
9. Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>.