Quality Checking and Matching Linked Dictionary Data

Kun Ji, Shanshan Wang, and Lauri Carlson

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

1 Introduction

The growth of web accessible dictionary 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 redundancy and doubtful quality of the resources.

In this paper, we develop ways to assess the quality of multilingual lexical web and linked data resources by internal consistency. Concretely, we deconstruct Princeton WordNet [1] to its component word senses or word labels, with the properties they have or inherit from their synsets, and see to what extent these properties allow reconstructing the synsets they came from. The methods developed should then be applicable to aggregation of term data coming from different term sources - to find which entries coming from different sources could be similarly pooled together, to cut redundancy and improve coverage and reliability. The multilingual dictionary BabelNet [2] can be used for evaluation.

We restrain our current research to dictionary data and improving language models rather than introducing external sources.

2 Methodology

In ([3]) we canvassed our sample of large dictionary and term databases for the descriptive fields/properties of entries to see what sources for matching entries they provide. The following types of properties susceptible to matching could be found in different combinations:

```
1) languages and labels
2) translations / synonyms
3) thematic (subject field) classifications
4) hypernym (genus/superclass) lattice
5) other semantic relations (antonym, meronym, paronym)
6) textual definitions / examples / glosses
7) source indications
8) grammatical properties (part of speech, head word, etc)
9) distributional properties (frequency, register etc)
10) instance data (text or data containing or labeled by terms)
```

Normally, these data sources in a dictionary vary in terms of coverage, unambiguity and information value. Labels are exact, but polysemous; semantic properties are informative, but scarce; distributional properties have a large information potential, but hard to make precise. Subject field classifications are potentially powerful, but alignments between them are often open as well.

Based on the above information sources, we have developed string-based distributional distance measures. Many of them are variations of Levenshtein edit distance [4]. Distributional measures are adaptable by machine learning, language independent and cheap, but fall short with unconstrained natural language (witness MT). A reason to hope that purely distributional methods work better on dictionary data is that dictionaries are a constrained language and self-documenting. For example, the following glosses for "green card" (culled from web-based glossaries) seem at first sight lexically unrelated, but WordNet synsets and the hypernym lattice allow relating many label pairs in them: "permit" is a "document", "immigrant" is a "person", "US" is "United States".

```
GREEN CARD OR PERMANENT RESIDENT CARD: A green card is a document which demonstrates that a
person is a lawful permanent resident allowing a non-citizen to live and work in the United
States indefinitely. A green card/lawful permanent residence can be revoked if a person does
not maintain their permanent residence in the United States travels outside the country for
too long or breaks certain laws.

GREEN CARD:  A permit allowing an immigrant to live and work indefinitely in the US.
```

We have tested and trained measures for some of the properties listed above on WordNet data. Although they find definite correlations, they are too weak taken singly to predict the synsets. The task is to develop a synset distance measure that combines individual measures on the different criteria above to one similarity vector and train it on WordNet data for optimal aggregation of WordNet senses back to WordNet synsets.

3 Conclusion

In this paper, we only use information available in the dictionary entries themselves in the criteria 1-9 above, to see how far dictionary internal information goes to reconstruct synset structure. Use of external information (instance data, language specific parsers, MT) will follow in subsequent papers.

References

1. Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>.
2. 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.
3. Wang. S and L. Carlson 2016. Linguistic Linked Open Data as a Source for Terminology - Quantity versus Quality. Proceedings of NordTerm 2015 (to appear).
4. Levenshtein, Vladimir I. (February 1966). "Binary codes capable of correcting deletions, insertions, and reversals". Soviet Physics Doklady. 10 (8): 707–710.