Some Structural Tests for Wordnets, with Results

Ahti Lohk
Tallinn University of Technology
Akadeemia tee 15a
Tallinn, Estonia
ahti.lohk@ttu.ee

Heili Orav
University of Tartu
Liivi 2
Tartu, Estonia
heili.orav@ut.ee

Leo Võhandu
Tallinn University of Technology
Akadeemia tee 15a
Tallinn, Estonia
leo.vohandu@ttu.ee

Abstract

This paper proposes some test-patterns (viewed as sub-structures) to evaluate the hierarchical structure of wordnets. By observing hierarchical structure, both top-down and bottom-up experiments are carried out on four wordnets: Princeton WordNet (version 3.1), Cornetto (version 2.0), the Polish Wordnet (version 2.0) and the Estonian Wordnet (version 67). The top-down approach is used to find small hierarchies, which are defined as having up to three levels of subordinates starting from unique beginners (rootsynsets). The bottom-up perspective is looking at the links that appear due to polysemy, and yet these are not. These redundant links form "asymmetric ring topology", and should be eliminated. Finally, an additional particular feature of large closed subsets will be introduced. Addressed views provide an opportunity to evaluate and/or improve the structure of wordnet hierarchies. This paper also provides an overview of the current status of these four wordnets from the according to our proposed test patterns.

1 Introduction

No linguist doubts the importance of wordnets. There are currently about 60 different wordnets worldwide. There are different views on the amount of information that is put into the system of synsets. But Miller and Fellbaum's primary goal, to create a large hypernym/hyponym relational style synset system is the same everywhere. Groups of specialists are involved in every implementation of wordnet for a given language. Every specialist has her/his subjective view about the relational connections between synsets.

It is important that every team has a strong belief in the high quality of the system they have created.

The theory and practice of building and checking computer chips with many millions of elements has proven that one has to build an independent test system to check designer created connections. As wordnets are similarly complex systems, we aim to build such a test system for wordnets.

The task of tests is to create lists of different types of inconsistencies which any Wordnet has at the given moment. Structural inconsistencies do not always translate to a wordnet error. The last word in checking wordnet lists always belongs to a lexicographer. What is truly crucial is that such lists are comprehensive. Tests must check all structurally weak areas of a given wordnet at any given moment.

After a lexicographer has made needed corrections, there follows a repetition of the same test. Such an iterative process has only one goal – to come to a clear understanding of all the weak places a given test can find.

Every created test has a different power. Some tests point with 100% probability to an error made by a lexicographer, although the error rate is usually below 100%. Such tests also have an important lexicographic value, as a long list of inconsistencies usually points to a complicated linguistic problem lacking a unique solution.

In this article we study only hypernym/hyponym relations.

2 Background of the wordnets

2.1 Princeton WordNet (PrWN)

Wordnets (Fellbaum, 1998) have emerged as one of the basic standard lexical resources in the language technology field. Princeton WordNet (PrWN) and most other wordnets are structured into synsets. A synset is usually described as capturing a lexicalised concept. Synsets are linked by conceptual relations with names borrowed from linguistic work on lexical semantics, such as hypernymy, holonymy, meronymy and so on.

More than 60 languages followed suit for building wordnets for their vernacular and very different compilation strategies have been applied. Some teams have decided to translate PrWN and adjust the result of that translation. Some word-

net developers have chosen an opposite route, such as expanding from the most frequent words or from top concepts as it has seen in ontological approaches.

The following is a brief introductory description of three databases from the Fenno-Ugric language family, and the Germanic and Slavic branches of the Indo-European language family.

2.2 Cornetto

The goal of Cornetto[1] was to build a lexical semantic database for Dutch, following the structure and content of Wordnet and FrameNet. Cornetto comprises information from two electronic dictionaries: the *Referentie Bestand Nederlands*, which contains FrameNet-like structures, and the *Dutch wordnet* (DWN) which utilises typical wordnet structures. DWN has a similar structure as the English WordNet although the top-level hierarchy was developed from an ontological framework and more horizontal relations are defined. The database has 70,371 synsets and 119,108 lexical units.

2.3 Polish Wordnet (plWN)

Work on PolNet began in 2005 (Derwojedowa, 2008), and its thesaurus is currently composed of nearly 116,000 synonym sets. The plWN development was organised in an incremental way, starting with general and frequently used vocabulary. The most frequent words from a reference corpus of the Polish language were selected.

2.4 Estonian Wordnet (EstWN)

The Estonian Wordnet began as part of the EuroWordNet project (Vossen, 1998), and was built by translating base concepts from English to allow monolingual extension. Words (literals) to be included were selected on frequency basis from corpora. Extensions have been compiled manually from Estonian monolingual dictionaries and other monolingual resources. After the start several methods have been used, for example domain-specific, i.e there have been dealt with semantic fields like architecture, transportation etc, there are some endeavors to add derivatives automatically and the results have been used of sense disambiguation process. Version 67 of EstWN consists of 60,434 synsets, including 82,515 words.

[1] http://www2.let.vu.nl/oz/cltl/cornetto/index.html

3 Related works

The most similar research to our paper has been done by Tom Richens, who has studied the anomalies in the WordNet verb hierarchies (Richens, 2008). Under the notion of topological anomalies, he notes three types of sub-structures in the hierarachical structure of WordNet that should be checked: “cycles”, “rings” (these in turn are classified into “asymmetric ring topology” and “symmetric ring topology”) and “dual inheritance”. He emphasizes that if “dual ineritance” (which also includes “asymmetric ring topology” and “symmetric ring topology”) appears, it merits investigation.

In his paper, Richens refers to the work of Pavel Smrž (Smrž, 2004) and Yang Liu (Liu, 2004). Smrž proposes twenty-seven tests for quality control in wordnet development. In most cases these tests are dealing with editing errors like “empty ID, POS, SYNONYM, SENSE (XML validation)” or “duplicate literals in one synset”, but some of them are errors of hierarchical structure, like “cycles”, “dangling uplinks”, “structural difference from PWN and other wordnets”, “multi-parent relations”.

Lin proves and refers to two kind of hypernymy faults in WordNet (about version 2.0): rings and isolators, and asserts that “In the future, some amendments should be made to solve these issues during the evolution of WordNet” (Liu, 2004).

Research about quality and evaluation of WordNet are made also by Aron N. Kaplan et al. (Kaplan, 2001), Philippe Martin (Martin, 2003), Raghuvar Nadig (Nadig, 2008) and Tomáš Čapek (Čapek, 2012).

4 Top-down view, small hierarchies

A top-down view of the structure will begin walking through the unique beginner separating all hierarchical structures (see Fig. 2), which end after the root of the concept on three next levels. This view can be useful for detecting small hierarchies that have somehow remained unconnected to a higher hierarchy. A large number of small hierarchies points to a lack of feedback (see Table 1).

PrWN was originally constructed with 25 unique beginners (rootsynset). These rootsynsets were later connected to a single unique beginner labeled ”entity” (Miller, 2007). From Table 1, it can be seen that in the PrWN there are only 11

Princeton WordNet	
rootsynset	**352** (n-12, v-340, a-0)
1 add. level	155 (n-11, v-144, a-0)
2 add. levels	81 (n-0, v-81, a-0)
3 add. levels	48 (n-0, v-48, a-0)
Cornetto	
rootsynset	**497** (a-454, n-2, v-2, r-12, c-27)
1 add. level	285 (a-263, r-11, c-1)
2 add. levels	148 (a-137, r-1, c-10)
3 add. levels	40 (a-37, n-1, c-2)
Polish WordNet	
rootsynset	**861** (n-531, v-35, j-295)
1 add. level	586 (n-335, v-25, j-226)
2 add. levels	159 (n-100, v-9, j-50)
3 add. levels	49 (n-34, v-0, j-15)
Estonian WordNet	
rootsynset	**169** (n-129, v-4, a-36)
1 add. level	128 (n-94, v-0, a-34)
2 add. levels	18 (n-16, v-0, a-2)
3 add. levels	6 (n-6, v-0, a-0)

Table 1: Number of rootsynsets and number of hierarchies that have only up to three additional levels of subordinates. (Numbers in brackets are about parts of speech as it is shown in every WordNet database.)

noun root synsets with one additional level of hierarchy, which is probably either due to human error, or unfinished work.

According to Table 1, Cornetto has only two noun and two verb hierarchies. That shows that every added synset is located directly into a large hierarchy. (Rootsynsets for the nouns are iets:2 and niets:1, translated as ”something” and ”nothing”.)

The much smaller number of Estonian Wordnet’s rootsynsets (169) is due to the fact that the team has gradually started to take into account the specific nature of the information obtained by structural tests. For example, in version 65, the number of rootsynsets was 303. Most of the decrease in rootsynsets is due to the fall of noun rootsysets has been reduced from 248 to 129.

It may be wise to take advantage of the low number of verb root concepts of EstWN to improve other wordnets’ verb hierarchies. This is certainly the case when the number of root concepts is too big.

The number of small hierarchies can be reduced considerably trying to locate them in the bigger hierarchy. This approach is a particular issue in the noun and verb trees.

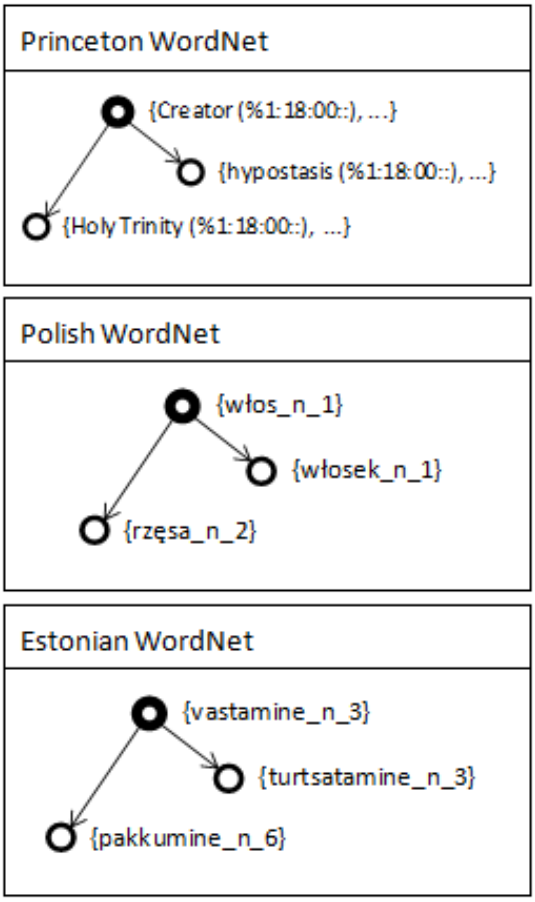

Figure 1: Small hierarchies. Rootsynsets with one additional level.

5 Bottom-up view, asymmetric ring topology

In this view, we are moving from lower level synsets to higher ones starting from synsets with many parents and separating substructures where such synsets are related to other synset directly and indirectly (see Fig. 2). The resulting subset is also referred to as a asymmetric ring topology (Richens, 2008) (see Table 2). This sub-structure may occur if lexicographers have created a new, more precise link to another synset, forgot to remove the previous relation. In this case one synset is connected to hypernym-synsets twice - directly and indirectly through other hypernym-synset (see Fig. 2)

6 The Largest Closed Subset (LGS)

LGS in hierarchical structures has been regarded as a coherent bipartite graph (Lohk, 2013).

	Synsets with many parents	Asymmetric ring topology
PrWN	1,425	30
Cornetto	2,438	306
plWN	10,942	476
EstWN	1,167	69

Table 2: Synsets with many parents and asymmetric ring topology numerically

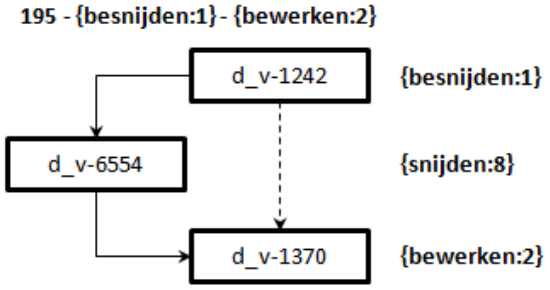

Figure 2: Asymmetric ring topology seen in Cornetto

In many cases LGS seems to be like particular feature of the hierarchical structure that links different hierarchical structures started from unique beginners. It is remarkable that in many cases the upper base of the bipartite graph consists of rootsynsets (see Table 3). Authors think that this conflict arises because the concepts of the root level are put to the same level with non-roots.

In Figure 3 an artificially constructed hierachical structure with one unique beginner (root node) has been shown. Closed subsets are highlighted by rectangles. Our interest is to find only the biggest ones, this is possible when a closed subsynset has at least two parents (represented with thick lines).

According to Figure 3 and Table 3 lower nodes in a closed subset are related to the first number in the second column of the table and upper nodes in a closed subset are related to the second number also in the second column of the table.

In the case of PrWN, every upper base synset in the bipartite graph belongs to the synset ”entity;” in the case of Cornetto, to ”iets:2” (in eng: ”something”); and in the case of EstWN into ”olev” (in eng: essive). Cornetto has one more large closed subset, related to verbs. As can be see in Table 1, the overall number of verb hierarchy is two and second big closed subset of Cornetto (in Table 3) connects these two (root synsets {afspelen:1, gebeuren:1, ..} and {zijn:7, uitmaken:2, vormen:5}).

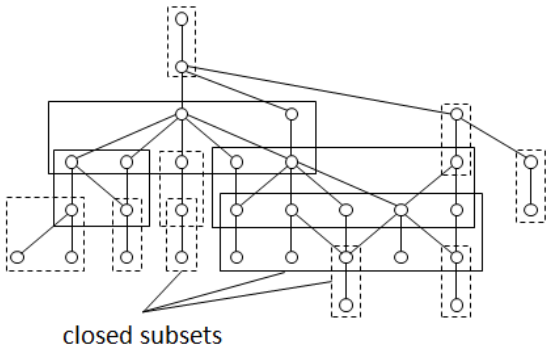

Figure 3: Artifically constructed tree of the WordNet with closed subsets

	The biggest closed subset	Root synsets in closed subset	Synsets of closed subsets that are connected to root synsets
PrWN	1,064 x 126	0	1
Cornetto[1]	11,032 x 589	0	1
Cornetto[2]	4,423 x 545	1	2
plWN	30,794 x 4,683	142	76
EstWN	1,526 x 66	8	1

Table 3: The largest closed subsets

While PrWN is obviously the most studied (see WordNet bibliography[2]) and Cornetto has a commercial version[3], it can be assumed that their hierarchical structure has received more attention (see Table 3, the number of rootsynsets in closed subset is in the case of PrWN and Cornetto 0).

Earlier tests with the Slovenian Wordnet (version 3.0) showed that a very large closed set may not be typical for all wordnets. It turned out that the largest closed subset size in this case was only 248 x 3.

LGS and closed subsets with many hypeonyms may be generally useful if the hypernyms in the upper base of closed sets are separated and their levels of concept are evaluated. Additionally, LGS seems to indicate the correctness (or uncorrectness) of the hierarchical structure, although this

[2] http://lit.csci.unt.edu/~wordnet/

[3] http://tst-centrale.org/nl/producten/lexica/cornetto/7-56

claim has not been definitively verified.

7 Discussion and Conclusion

The most difficult issue for wordnet compilers with regard to noun hierarchical relationships is to find the top hypernyms. The same also occurs in regard to finding the top concepts for the most frequent verbs, both transitive and intransitive. As for adjectives, the situation is even more unclear, as wordnets for various languages deal with adjectives differently. In some wordnets, adjectives are hierarchical (as seen in Table 1: Cornetto, EstWN), but in PWN, adjectives have different types of" semantic connections.

One analyses only the short hierarchies in all wordnet variants, (root level plus up to 3 lower levels) one comes to the realisation that new addons for wordnets have created a situation in which missing feedback has lost the information required to correctly connect synsets.

All wordnets studied here show that the expansion process requires strong and effective feedback.

As is made clear by Table 1, in the top-down perspective, three of the four wordnets studied here require either verb or noun hierarchy correction. However, as Cornetto has only two hierarchies for nouns and verbs, it has somehow excluded small hierarchies. This shows that Cornettos team is using different tools or/and ways for add-ons.

References

Magdalena Derwojedowa, Maciej Piasecki, Stanislaw Szpakowicz, Magdalena Zawislawska and Bartosz Broda. 2008. *Words, Concepts and Relations in the Construction of Polish WordNet* In Proceedings of the Fourth Global WordNet Conference - GWC 2008. pp: 162–177.

Tomáš Čapek. 2012. *SENEQA – System for Quality Testing of Wordnet Data.* Proceedings of 6th International Global Wordnet Conference. Matsue, Japan, 9-13 January 2012. pp: 400-404.

Christiane D. Fellbaum. 1998. *WordNet An Electronic Lexical Database* Cambridge, Massachusetts, London, England: The MIT Press

Aaron N. Kaplan and Lenhart K. Schubert. 2001. *Measuring and improving the quality of world knowledge extracted from WordNet.* University of Rochester, Rochester, NY.

Yang Liu, Jiangsheng Yu, Zhengshan Wen and Shiwen Yu. 2004. *Two Kinds of Hypernymy Faults in WordNet: the Cases of Ring and Isolator.* Proceedings of the Second Global WordNet Conference. Brno, Czech Republic, 20-23 January 2004. pp: 347-351.

Ahti Lohk, Ottokar Tilk and Leo Võhandu . 2013. *How to Create Order in Large Closed Subsets of WordNet-type Dictionaries* Estonian Papers in Applied Linguistics 9 pp: 149–160.

Philippe Martin. 2003. *Correction and extension of WordNet 1.7* Conceptual Structures for Knowledge Creation and Communication: Springer. pp: 160–173.

George A. Miller. and Christiane D. Fellbaum. 2007. *WordNet then and now* Lang Resources & Evaluation, Volume 41, Issue 2. pp: 209–214.

Nadig Raghuvar, Ramanand J and Bhattacharyya Pushpak. 2008. *Automatic Evaluation of Wordnet Synonyms and Hypernyms* Proceedings of ICON-2008: 6th International Conference of Natural Language Processing.

Tom Richens. 2008. *Anomalies in the wordnet verb hierarchy* Proceedings of the 22nd Inter-national Conference on Computational Linguistics: COLING-ACL 2008. pp: 729–736.

Piek Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* Dordrecht: Kluwer Academic Publishers.

Pavel Smrž. 2004. *Quality Control for Wordnet Development.* Proceedings of the Second Global WordNet Conference. Brno, Czech Republic, 20-23 January 2004. pp: 206-212.