Dense Component in The Structure of Wordnet

Ahti Lohk[1], Heili Orav[2], Kaarel Allik[1], Leo Võhandu[1]

Tallinn University of Technology[1], University of Tartu[2]
Akadeemia tee 15a, Tallinn, Estonia[1]
Liivi 2, Tartu, Estonia[2]
ahti.lohk@ttu.ee, heili.orav@ut.ee, kaarel.allik@ttu.ee, leo.vohandu@ttu.ee

Abstract

This paper introduces a test-pattern named a dense component for checking inconsistencies in the hierarchical structure of a wordnet. Dense component (viewed as substructure) points out the cases of regular polysemy in the context of multiple inheritance. Definition of the regular polysemy is redefined – instead of lexical units there are used lexical concepts (synsets). All dense components are evaluated by expert lexicographer. Based on this experiment we give an overview of the inconsistencies which the test-pattern helps to detect. Special attention is turned to all different kind of corrections made by lexicographer. Authors of this paper find that the greatest benefit of the use of dense components is helping to detect if the regular polysemy is justified or not. In-depth analysis has been performed for Estonian Wordnet Version 66. Some comparative figures are also given for the Estonian Wordnet (EstWN) Version 67 and Princeton WordNet (PrWN) Version 3.1. Analysing hierarchies only hypernym-relations are used.

Keywords: wordnet, test-pattern, dense component

1. Introduction

Wordnet (Miller and Fellbaum, 1998) as a lexical resource is attractive due to its hierarchical structure of synonym sets (synsets), which is helpful for many natural language processing (NLP) tasks. Wordnet is mostly used for machine translation, automate analysis of text and word sense disambiguation, but also for text categorization, information retrieval, text mining and even for creating new wordnets (Morato et al., 2004). Polysemy as a feature of wordnet hierarchical structure may complicate the NLP (Veale, 2004) and affect the quality of these applications. At the same time, the polysemy may help to find and define new semantic relations between lexical units or synsets which in turn help to improve utility of wordnet in NLP tasks. (Barque et al., 2009) and (Freihat et al., 2013) use regular polysemy patterns to discover these new semantic relations. In our research we redefine the meaning of regular polysemy. To find the cases of regular polysemy in the hierarchical structure of wordnet we use a test-pattern named a *dense component* which is viewed as a substructure of the wordnet hierarchy. Generally defining, the *dense component* is a bipartite graph that has at least two synsets with at least two identical parents, but could contain additional synsets with some common parents (synsets with dotted line) as shown in Figure 1.

With respect to the state of the art, regular polysemy in wordnet is viewed as a status where at least two lexical units (members of synset) from the same or different level in hierarchical structure are related to the combination of one of the following:

1. lexical units (from higher level synsets) (Peters and Peters, 2000; Freihat et al., 2013);
2. "conceptual signposts"(Peters and Peters, 2000)[1];
3. top ontology concepts, unique beginners or domain category names (Buitelaar, 1998; Freihat et al., 2013).

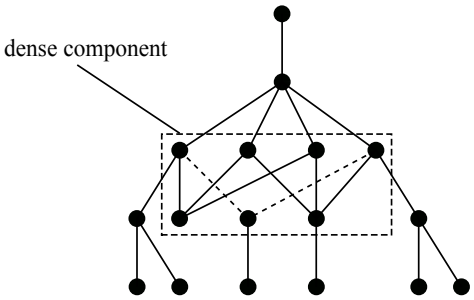

Figure 1: Dense component in a hierarchical structure

However, in our view we apply the same idea of regular polysemy (RP) but instead of abovementioned categorization we use synsets as lexical concepts. So redifining the RP we say that RP is a status where at least two synset have at least two hypernyms with similar relations between those hypernyms.

The paper fills the gap in the state-of-the art by asking the main research question of how to check and evaluate regular polysemy in the hierarchical structure of wordnet? To answer the question, we present a test-pattern named *dense component* view on substructures of the wordnet hierarchy in case of regular polysemy.

The structure of the paper is as follows: Section 2 gives additional background for understanding the main body of paper. Next, Section 3 presents formalized algorithm of dense component. Section discusses the inconsistencies taxonomy. Section 5 evaluates the dense component yielding a

[1]"A pair-wise combinations of nodes in the WordNet hierarchy that are preferably more specific than the unique beginners but still general enough to encompass several words and constitute semantically homogenous groups"

numerical overview and finally, Section 6 concludes the paper.

2. Features of Wordnet Dictionaries

Wordnets share properties for the concepts of polysemy that are part of the definitions of the test patterns. On the other hand, regular polysemy is only part of one test-pattern definition, namely the pattern $dense\ component$. In the remainder, Section 2.1. gives general structural features for wordnet and Section 2.2. polysemy versus regular polysemy.

2.1. Wordnet-like dictionaries

The fundamental approach for designing WordNet-like dictionaries came from Princeton WordNet (Miller, 1990). Each WordNet shares particular structural features. First, synonym sets (synsets) group many synonyms that share the same meaning and are also referred to as concepts. Semantic relations connect synsets to each other, e.g., by *hypernymy*, *meronymy* for creating a hierarchical structure, and *caused by*, *near synonym* that do not create a hierarchical structure. In this article, we consider only hypernymy-hyponymy relations as objects of analysis. Furthermore, there is no extension limitation for the approach to different semantic relations that shape the hierarchical structure. For details about Estonian Wordnet, we refer the reader to (Õim et al., 2010). Furthermore, Princeton WordNet has 117 773 synsets and 88 721 hypernym-hyponym relations. In Estonian Wordnet Versions 66, these values are 58 566 and 51 497 respectively, while for Versions 67, the values are 60 434 and 52 678 respectively. Princeton WordNet has hypernym-hyponym relations only in cases of nouns and verbs; in Estonian Wordnet in case of nouns, verbs and adjectives.

2.2. Regular polysemy

According to (Ravin and Leacock, 2000), polysemy is multiplicity of meanings of words. In wordnets, polysemy should appear as one concept with several hypernyms. If the latter are regularly included then the polysemy itself is regular. The best known definition of regular (also systematic or logic) polysemy gives (Apresjan, 1974). In (Langemets, 2010) Apresjan's definition is simplified: regular polysemy is a status where at least two words have at least two meanings with similar relation between those meanings. For example, if the word *school* has meanings *institution* and *building* than the same is true about a *hospital*. The latter is also an *institution* as well as a *building*. According to (Freihat et al., 2013), *institution-building* is an example for a polysemic pattern. Our goal in regular polysemy cases is to check if the polysemic pattern is justified with respect to regular polysemy.

Following section formulates a mathematical concept of dense component.

3. Definitions and algorithm of the dense component

Let $G = (Y, A, E)$ be a bipartite graph whose partition has the parts Y and A; $E \subseteq A \times Y$ is the set of edges. Let $|Y| = m$ and $|A| = n$.

Our goal is to glue together some nodes from A under certain conditions. Therefore it is convenient to represent the result by

$$\hat{G} = \{g_i : g_i =< L_i, N_i >; \quad i = 1, \ldots, k; 1 \leq k \leq n\},$$

where L_i is the set of glued nodes from A and N_i is the set of neighbours of L_i.

For a natural number τ we define a binary relation $R(\hat{G})_\tau \subseteq \hat{G} \times \hat{G}$:

$$R(\hat{G})_\tau = \{(g_i, g_j) : g_i \in \hat{G}, g_j \in \hat{G}, |N_i \cap N_j| \geq \tau\}.$$

We say, that g_i and g_j from $\hat{G}$ are τ-*connected*, if $(g_i, g_j) \in R(\hat{G})_\tau$. Obviously, $R(\hat{G})_\tau$ is symmetrical and reflexive and the emptyness of $R(\hat{G})_\tau$ can be detected in time $\mathcal{O}(k^2 \cdot m)$.

For given $\hat{G}$ and $(u, v) \in R(\hat{G})_\tau$ we denote

$$glue(u, v, R(\hat{G})_\tau) = (\hat{G} \backslash \{u, v\}) \cup z,$$

where $z =< L_u \cup L_v, N_u \cup N_v >$.

Algorithm 1 τ *-closure*
$l := 0$; $\hat{G}^0_\tau := \{< \{g_i\}, N_i >: g_i \in A, N_i = \{y : (g_i, y) \in E\}\}$;
while $R(\hat{G}^l)_\tau \neq \emptyset$
do choose $u, v : (u, v) \in R(\hat{G}^l)_\tau$;
$\hat{G}^{l+1}_\tau := glue(u, v, \hat{G}^l_\tau); l := l + 1$;
od $\hat{G}^+_\tau := \hat{G}^l_\tau$;

The result of the algorithm, $\hat{G}^+_\tau$ is called τ-*closure* of G.

Every step of the cycle glues two nodes, therefore the Algorithm 1 halts after at most $n - 1$ steps.

Due to commutativity and associativity of the set union ($\cup$), the τ-closure does not depend on the order of choosing elements in the body of the cycle. Therefore $\hat{G}^+_\tau$ is unique for G.

Definition: Dense component is every item g in graph $\hat{G}^+_\tau$ (Algorithm 1) and it is corresponding to Fig. 1.

$$g \in \hat{G}^+_\tau \quad (2)$$

In next section we explain what kind of inconsitencies can be found with help of previously described algorithm of finding dense components.

4. Inconsistencies of Substructure

4.1. Inconsistency taxonomy

Inconstency types lexicographer is focusing on are following:

1. **Non-justified regular polysemy** – in accordance with Section 2.2., linguists have to check if the regular polysemy is justified or not. Furthermore, having expanded view of dense component (i.e. additional synsets connected to dense component), perspective of regular polysemy may help to detect situations where there exist other synsets that are not connected to the same polysemic pattern as shown in Figure 2.

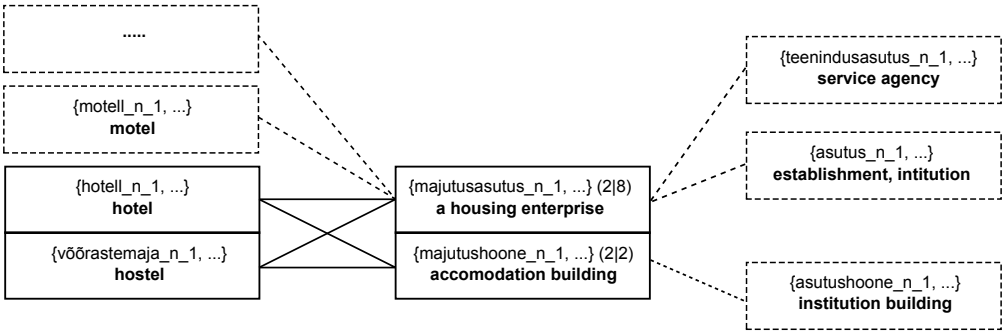

Figure 2: Dense component, non-regular use of polysemy

2. **Ignoring the principle of economy (redundant semantic relation)** – this inconsistency is typical to an asymmetric ring topology in cases where one branch is redundant such as in Figure 2. (Liu et al., 2004; Richens, 2008).

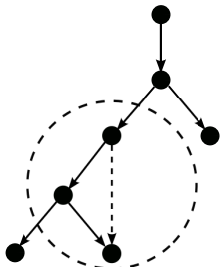

Figure 3: Asymmetric ring topology, dotted line is redundant

3. **Inappropriate semantic relationship** – it implies that a semantic relationship type must change. This inconsistency affects every test-pattern.

4. **Wrongly inherited domain category** – if one synset inherits simultaneously two different domain categories, one of them is wrong (Liu et al., 2004). The gloss of the synset indicates which of the categories is most appropriate (Miller and Fellbaum, 1998). Unfortunately, this inconsistency is applicable only on PrWN, because its every synset has the information about the domain category in contrast to EstWN.

4.2. Some examples

In this section we present three dense component examples with their specific inconsistencies. In order to facilitate the work of lexicographer, additional synsets connected to dense component are marked using dotted line. Mostly connected synsets (usually located in the middle of the figures 2, 4, 5) are called parents of the dense component. Every parent contains information about its number of subordinates (represented in brackets). First number shows connections to subordinates in the dense component, and second one refers to total number of subordinates (see Figures 2, 4, 5).

In **Figure 2**, we have typical case where the regular polysemy is allowed – *hotel* is simultaneously the building and the institution. While the nature of the *motel* is similar to the *hotel* we expect that the *motel* is connected to same polysemic pattern as the *hotel*.

In **Figure 4**, we see the case where the concept *cinnabar* mistakenly has got three hypernyms. According to the definition of *cinnabar*, only one hypernym was left for *cinnabar – mineral*. Colors as part of material have been changed to *holonym* instead of *hypernym*.

In **Figure 5**, we meet the asymmetric ring topology case. In order to facilitate the work of the lexicographer all these relvant synsets can be higlighted as shown in the case of *artistic production, art*. At the same time this is the case where one co-hypernym ({*artistic production, art*}) becomes to be parents for another ({*applied art*}). I.e., {*artistic production, art*} links with {*glasswork, ...*} and {*leatherwork, ...*} have to be removed.

5. Evaluation

In this section we compare EstWN Version 66 to 67 to see the changes that have taken place in wordnet hierarchy after correcting the dense components by the lexicographer. Hereby, we focus on four different changes as follows: the number of multiple inheritance, sizes of dense components, the number of dense components and distribution of errors.

5.1. The number of multiple inheritances

Every polysemic case in dense component is related to multiple inheritance, i.e. with synsets that have at least two parents/hypernyms in wordnet hierarchy. Therefore correcting a dence component it affects multiple inheritances as well.

Nr of parents	**EstWN, v66** (number of synsets)	**PrWN, v3.1** (number of synsets)	**EstWN, v67** (number of synsets)
5	1	1	–
4	5	3	1
3	68	30	32
2	1 603	1391	1 131
SUM	1 677	1 425	1 164

Table 1: Multiple inheritance counts before and after analysis and correction of the dense components.

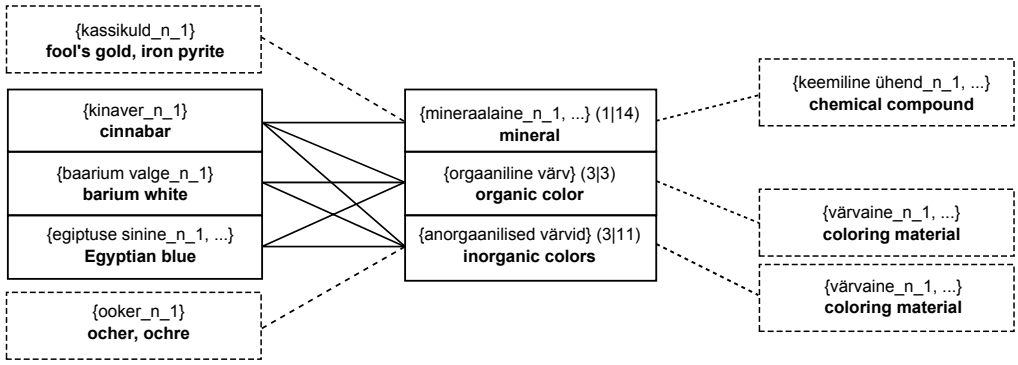

Figure 4: Dense component, wrong semantic relation

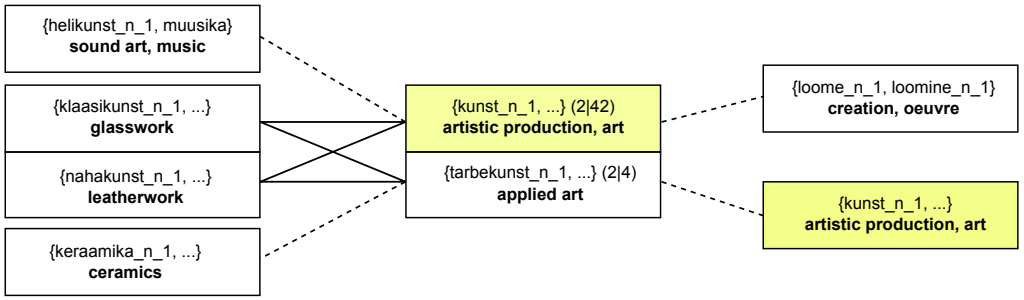

Figure 5: Dense component, asymmetric ring topology

Looking at the Table 1, we see that after correction of dense components there are no synsets with 5 parents in Version 67. Synsets with 3 parents are reduced about 50% and dual inheritance is reduced about by 500 cases.

5.2. Size and number of the dense components

According to the number of parents in dense components we present in Table 2 ten components with the highest number of parents with their occurances for two EstWN versions and for one PrWN version.

A considerable change after correction of dense components can be observed in their number of occurences. In the last row of Table 2 we see that the number of dense components is fallen from 121 to 24. The number of the biggest dense components (according to the number of parents) and the number of small dense components have also significantly decreased. E.g., both wordnets EstWN Version 66 and PrWN Version 3.1 include the same number of the smallest dense component (2 x 2) – 59. After correction this number dropped to 11.

Nr	**EstWN, v66** (synsets x parents) x nr	**PrWN, v3.1** (synsets x parents) x nr	**EstWN, v67** (synsets x parents) x nr
1	(5 x 9) x 1	(3 x 5) x 1	(3 x 3) x 1
2	(6 x 6) x 1	(2 x 5) x 1	(2 x 3) x 1
3	(116 x 4) x 1	(4 x 4) x 1	(8 x 2) x 1
4	(5 x 4) x 1	(3 x 4) x 1	(7 x 2) x 1
5	(3 x 4) x 1	(2 x 4) x 2	(6 x 2) x 1
6	(2 x 4) x 3	(9 x 3) x 1	(5 x 2) x 1
7	(19 x 3) x 1	(4 x 3) x 2	(4 x 2) x 2
8	(10 x 3) x 1	(3 x 3) x 3	(3 x 2) x 5
9	(8 x 3) x 2	(2 x 3) x 7	(2 x 2) x 11
10	(4 x 3) x 1	(9 x 2) x 2	–
SUM	121	107	24

Table 2: Dense components (bipartite graphs) sizes in EstWN (v66), PrWN (v3.1) and EstWN (v67). First ten components.

5.3. Distribution of corrections

In Table 3 we give a detailed overview about corrections that were made by the lexicographer. This table is based on comparing dense components from EstWN Version 66 to Version 67 manually. The sum of the first column numbers (106+14+65+39+14) in Table 3 is not equal to 121, because in many types of corrections have been included by the same dense components.

The number 106 presented in the first row points to the situation where dense component as a pattern is useful particularly in the checking of justness of regular polysemy cases. If regular polysemy is not justified, it means that some semantic relations have just been removed. due to background synsets that were added to every dense component (represented to dotted lines) we could Detecting that *principle of economy was not followed* in the second row.

While asymmertic ring topology is possible in cases where direct link exceeds/overpasses more then one level of hierarchicy, we can not expect that dense component refers to all these kinds of inconstistencies.

In th third row, in about 50% of cases of dense components were engaged in the process of changing the semantic relations. Within this, 162 semantic relations were changed. Hypernym relation was exchanged to near synonym 88 times, to fuzzynym 52 times etc.
Hierarchy was changed 39 times. Main reason was in circumstances where one co-hypernym or co-hypenym became parent to the another.
Only 14 dense components did not need any corrections. However, Version 67 consists of 24 dense components. These 24 had their content as follows:

- 14 of them were without any correction
- 2 of them were changed a little bit
- 8 of them were new

Futhermore, all dense components in Version 66 were revised, 1 868 synsets and 1 181 semantic relations were added into Version 67 as well. For that reason new 8 dense components arised in Version 67.

106	regular polysemy was not justified		
14	the principle of economy was not followed		
65	dense components was connected to changes of semantic relation		
	162	semantic relation was changed to	
		88	near synonym
		52	fuzzynym
		20	holonym
		2	meronym
39	hierarchy was changed in cases		
	14	co-hypernyms/co-hyponyms, one became parents to other one	
	7	connection to a synset is replaced to other one	
	5	new synset was added	
	4	added or removed lexical units from synsets	
	3	synsets were merged	
	2	removed synsets	
	4	hierarchical strcuture was reorganized	
14	no correction needed		

Table 3: Distribution of corrections

6. Conclusion

In this paper, we propose to use a dense component as a test-pattern to detect inconsistencies in substructures of wordnet hierarchy. Dense component is viewed on the one hand as bipartite graph and on the other hand as substructure of wordnet hierarchy and as a visual picture. It consists of at least one regular polysemy and simultaneously at least two synsets with at least two identical parents. Its finding process takes place iteratively trying to find fore current dense component other synsets that have at least two parents among the current dense component (see Section 3). The greatest benefit the dense component may give to lexicographer is helping to check the correctness of regular polysemy, i.e., it helps to see if the regular polysemy is justified or not but it is not limited to that case. Exhaustive analysis made by second author surprised positively because only 12% of dense components did not need any correction. The number of dense components in EstWN Version 66 diminished after corrections from 121 to 24 in Version 67.

7. Acknowledgements

In this paper Ahti Lohk is supported by Estonian National Doctoral School in Information and Communication Technology.

8. References

J. D. Apresjan. 1974. Regular polysemy. *Linguistics*, 12(142):5–32.

L. Barque, F.-R. Chaumartin, et al. 2009. Regular polysemy in wordnet. *JLCL-Journal for Language Technology and Computational Linguistics*, 24(2):5–18.

P. Buitelaar. 1998. *CoreLex: systematic polysemy and underspecification*. Ph.D. thesis.

A. A. Freihat, F. Giunchiglia, and B. Dutta. 2013. Approaching regular polysemy in wordnet. In *eKNOW 2013, The Fifth International Conference on Information, Process, and Knowledge Management*, pages 63–69.

M. Langemets. 2010. *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras*. Eesti Keele Sihtasutus.

Y. Liu, J. Yu, Z. Wen, and S. Yu. 2004. Two kinds of hypernymy faults in wordnet: the cases of ring and isolator. In *Proceedings of the Second Global WordNet Conference*, pages 347–351.

G. Miller and C. Fellbaum. 1998. *Wordnet: An electronic lexical database*. MIT Press Cambridge.

J. Morato, M. A. Marzal, J. Lloréns, and J. Moreiro. 2004. Wordnet applications. In *GLOBAL WORDNET CONFERENCE*, volume 2, pages 270–278.

H. Õim, H. Orav, K. Kerner, and N. Kahusk. 2010. Main trends in semantic-research of estonian language technology. In *Baltic HLT*, pages 201–207.

W. Peters and I. Peters. 2000. Lexicalised systematic polysemy in wordnet. In *LREC*.

Y. Ravin and C. Leacock. 2000. *Polysemy: Theoretical and computational approaches*. MIT Press.

T. Richens. 2008. Anomalies in the wordnet verb hierarchy. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 729–736. Association for Computational Linguistics.

T. Veale. 2004. Polysemy and category structure in wordnet: An evidential approach. In *LREC*.