Automatic generation of a coarse grained WordNet

Rada Mihalcea and Dan I. Moldovan
Department of Computer Science and Engineering
Southern Methodist University
Dallas, Texas, 75275-0122
{rada, moldovan}@engr.smu.edu

Abstract

Several principles for the automatic transformation of WordNet into a coarser grained dictionary are proposed. A new version of WordNet is derived, leading to a reduction of 26% in the average polysemy of words, while introducing a small error rate of 2.1%, as measured on a sense tagged corpus.[1]

1 Introduction

WordNet is well known to the Natural Language Processing (NLP) community as a valuable resource: more and more applications requiring machine readable dictionaries or world knowledge encoded in semantic networks use WordNet. Certain applications, such as text inference or knowledge processing (Harabagiu and Moldovan, 1998), need a large set of relations among concepts and the large number of concepts and semantic links found in WordNet is well suited. On the other hand, other applications such as semantic or conceptual indexing, and sometime word sense disambiguation or machine translation, do not always need such a fine grain distinction between concepts. This constitutes one of the most often *"critiques"* brought to WordNet: the fact that sometime word senses are so close together that a distinction is hard to be made even for humans.

We propose a methodology for automatic generation of a coarser WordNet by either collapsing synsets (sets of synonyms) very similar in meaning, or dropping synsets very rarely used. The results obtained are encouraging: we show that using the rules presented in this paper one can automatically generate a new version of WordNet, which we call *EZ.WordNet.1*, with an average polysemy reduced with 26%, while the level of ambiguity introduced is only 2.1% as measured on SemCor (corpus tagged with WordNet senses). We also derive an alternative version *EZ.WordNet.2*, using the same rules but with different parameters, with a reduction of 39% in average polysemy and an error rate of 5.6%.

Several applications can benefit from such a method of defining relatedness between synsets:

1. When evaluating Word Sense Disambiguation (WSD) systems it is important to know which senses of a word are similar and which are not (Resnik and Yarowsky, 1999). A WSD system misstagging the musical sense of the word *bass* with its fish sense will be penalized more than a system confusing two different musical senses of the word *bass*.
2. WSD for certain applications such as semantic indexing (Gonzalo et al., 1998) do not need to make such fine distinctions between senses. Also, similar meanings collapsed together bring more candidate words for the task of query expansion (Moldovan and Mihalcea, 2000).
3. By reducing the semantic space, the task of WSD algorithms is made easier.
4. Machine translation applications, given that close senses tend to have the same translation.

The work most similar to ours is the one reported in (Peters et al., 1998). Automatic sense clustering is performed using the relations defined in WordNet: *sisters*, *autohyponymy*, *twins* and *cousins*. The effects of clustering are evaluated using EuroWordNet, namely by measuring the compatibility of language specific wordnets. Besides the measures they are using, which were proved to be useful for sense clustering and applications, we propose additional methods for increased polysemy reduction.

2 WordNet and synsets similarity

WordNet (Fellbaum, 1998) covers the vast majority of nouns, verbs, adjectives and adverbs from the English language. Table 1 shows the number of words, respectively the number of synsets defined in WordNet 1.6 for each part of speech.

One or more senses are defined for each word. Depending on the number of senses it has, a word can be (1) *monosemous* (one sense) or (2) *polysemous* (two or more senses). Table 1 shows also the number of polysemous and monosemous words and the average polysemy measured on WordNet. It results an average polysemy of 2.91 for polysemous words and 1.34 if monosemous words are included.

[1]This research has been supported in part by NSF grant EIA-0078854

Part of speech	Word forms	Synsets	Total senses	Polys. words	Monos. words	Total senses polys.words	Avg. polysemy (- monos)	Avg. polysemy (+ monos)
Noun	94473	66024	116318	12562	81911	34407	2.73	1.23
Verb	10318	12126	22067	4565	5753	16314	3.57	2.13
Adj	20169	17914	29882	5372	14797	15085	2.80	1.48
Adverb	4545	3574	5678	748	3797	1881	2.51	1.25
TOTAL	129505	99638	173945	23247	106258	67687	2.91	1.34

Table 1: Statistics on WordNet: number of word forms, number of synsets, number of monosemous and polysemous words, average polysemy in WordNet 1.6

Humans tend to frequently use words with higher polysemy, and this makes the distinction of word senses a problem. This fact is proven by statistics derived from SemCor (Miller et al., 1993) showing that words used in common texts have an average polysemy of 6.55. Table 2 lists the total number of occurrences of word forms in SemCor, and the number of senses defined in WordNet for these words.

Part of speech	Word occurrences	Senses defined in WN	Average polysemy
Noun	88398	382765	4.33
Verb	90080	944368	10.48
Adj	35770	157751	4.41
Adv	20595	55617	2.70
TOTAL	234843	1540501	6.55

Table 2: Statistics on SemCor: number of word occurrences, number of senses defined in WordNet and average polysemy

The average polysemy of 6.55 is very high, and this is why WordNet is classified as a fine grained dictionary. There are cases when even humans have difficulty making distinction between WordNet senses. In this context, it would be useful to define a methodology indicating which senses are more related to each other, and consequently obtain a reduction in the average polysemy.

Similarity measures in WordNet.

There are already some similarity relations defined in WordNet. These relations are defined only for nouns and verbs, and do not always succeed in properly indicating two synsets as similar in meaning.

For verbs, there is a *VERBGROUP* pointer, linking synsets similar in meaning; these pointers are manually determined by lexicographers. For nouns, three types of synset relations are defined: (1) *sisters*, representing synsets with a word in common and with the same direct hypernym; (2) *cousins*, which are synsets with a word in common and with at least one pair of direct or indirect hypernyms related, based on a predefined list and (3) *twins*, which are synsets with three or more words in common. Additionally, (Peters et al., 1998) mentions a fourth relation, (4) *autohyponymy*, referring to words whose senses are each others direct hypernym or hyponym. No such relations are defined for adjectives and adverbs.

Several problems are associated with these relations: (1) coverage of verb groups is incomplete (as mentioned in the WordNet manuals); (2) the measures for noun synsets similarity are sometime too strong, and sometime too loose; for example, these relations will wrongly group together {*house#3*} with its meaning of *"a building in which something is sheltered or located"* and {*house#5, theater, theatre*} meaning *"a building where theatrical performances or motion-picture shows can be presented"*, but they will fail to group as similar the synsets {*decapitation#1, beaheading*} defined as *"execution by cutting off the victim's head"* and {*decapitation#2, beaheading*} meaning *"killing by cutting off the head"*.

The methods we propose are able to cluster similar synsets together based on a set of rules derived from sound principles. These methods can be also applied to adjectives and adverbs.

3 Semantic principles: tests for ambiguity

When talking about word meanings and ambiguity, a major issue is how to actually determine if a word is ambiguous or not. (Cruse, 1986) describes three principles used by humans in testing the ambiguity of words. In the following, we enumerate these principles and show how they can be translated into methods for measuring the ambiguity level among between WordNet synsets.

PRINCIPLE I. *If there exists a synonym or one occurrence of a word form which is not a synonym of a second, syntactically identical occurrence of the same word in a different context, then that word form is ambiguous, and the two occurrences exemplify different senses.*

The example given as an illustration of this principle is the word *"match"*: in the contexts *"Gay struck the match"*, respectively *"The match was draw"*, it has two different synonyms, namely *"lucifer"* (or *"friction match"*) and *"contest"*. It results that this word is ambiguous in the two given contexts.

This principle can be reformulated as it follows: if a word, in two different contexts, has the same synonyms, then that word has similar meanings in the given contexts.

Translated in WordNet terminology, the context of a word meaning is represented by its synset and its relations to other synsets. We can infer that if a word has two senses with the same synonyms, then it is hard to distinguish among the two senses, and thus we can collapse the appropriate synsets together:

> **Rule SP1.1** If S1 and S2 are two synsets containing at least two words, and if S1 and S2 contain the same words, then S1 and S2 can be collapsed together into one single synset S12.

> *Example:* Sense #1 and #3 for *paper* are:
> S1 = {*newspaper, paper*} (*a daily or weekly publication on folded sheets*)
> S3 = {*newspaper, paper*} (*a newspaper as a physical object*)
> These two synsets can be collapsed together, as they are very similar in meaning.

There are cases when the synsets contain only one single word, and thus we cannot directly apply this principle. A good approximation of the *synonymy* relation is represented by the *hypernymy* links found in WordNet: if a particular sense of a word has no synonyms, its meaning can be judged by looking at its hypernym. Another rule is inferred:

> **Rule SP1.2**
> If S1 and S2 are two synsets with the same hypernym, and if S1 and S2 contain the same words, then S1 and S2 can be collapsed together into one single synset S12.

> *Example*: Consider senses #1 and #2 for the verb *eat*:
> S1 = {*eat*} (*take in solid food*)
> ⇒ {*consume, ingest, take in, take, have*}
> S2 = {*eat*} (*eat a meal*)
> ⇒ {*consume, ingest, take in, take, have*}
> These two senses are fine distinctions of the possible meanings of *eat*, and the two synsets S1 and S2 can be collapsed together.

By relaxing this first principle in its requirement of having *all* the synonyms of two different word meanings identical in order to fuse the appropriate synsets, to a requirement of having at least K identical synonyms, we can infer Rule SP1.3. By taking K=3, we obtain the *twins* similarity measure.

> **Rule SP1.3** If S1 and S2 are two synsets with at least K words in common, then S1 and S2 can be collapsed together into one single synset S12.

> *Example* Sense #1 and #3 for noun *teaching* are:
> S1 = {*teaching, instruction, pedagogy*} (*the profession of a teacher*)
> S3 = {*education, instruction, teaching, pedagogy, educational activity*} (*activities that impart knowledge*)

PRINCIPLE II. *If there exists a word or expression standing in a relation of oppositeness to one occurrence of a word form, which does not stand in the same relation to a second, syntactically identical occurrence of the same word form in a different context, then that word form is ambiguous, and the two occurrences exemplify different senses.*

The example given for this ambiguity test is the word *"light"*, which is determined to be ambiguous in the sentences *"The room was painted in light colours"*, respectively *"Arthur has a rather light teaching load"*, as it has two different antonyms: *"dark"* and *"heavy"*.

The reformulation of this principle is: if a word, in two different contexts, has the same antonyms, then that word has similar meanings in the given contexts. We can translate this in WordNet terms, and measure the grade of similarity among two synsets with the following rule:

> **Rule SP2** If S1 and S2 are two synsets representing two senses of a given word, and if S1 and S2 have the same antonym, then S1 and S2 can be collapsed together into one single synset S12.

> *Example*: Senses #1 and #2 for the verb *dress* are:
> S1 = {*dress, get dressed*} (*put on clothes*)
> antonym ⇒ {*undress, discase, uncase, unclothe, strip, strip down, disrobe*}
> S2 = {*dress, clothe, enclothe, garb, raiment, tog, garment, habilitate, fit out, apparel*} (*provide with clothes or put clothes on*)
> antonym ⇒ {*undress, discase, uncase, unclothe, strip, strip down, disrobe*}

Finally, the last test of ambiguity refers to paronymic relations [2] related to a particular word.

PRINCIPLE III. *If there exists a word which stands in a paronymic relation to one occurrence of a word form, but does not stand in the same relation to a second, syntactically identical occurrence of the same word form in a different context, then that word form is ambiguous, and the two occurrences exemplify different senses.*

This principle is illustrated using the noun *"race"*. In the sentence *"The race was won by Arkle"*, this noun has the verb *"to race"* related to it, while the same noun in the sentence *"They are a war-like race"* has two other related words, namely *"racial"* and *"racist"*. By having different paronymically related

[2] Relations involving identity of the root, but different syntactic categories (e.g. *act* - *actor*)

Part of speech	Probability of having sense number								
	1	2	3	4	5	6	7	8	>8
Noun	78.52%	12.73%	4.40%	2.07%	0.98%	0.52%	0.34%	0.17%	<0.1%
Verb	61.01%	19.22%	7.89%	4.12%	2.64%	1.47%	0.98%	0.65%	<0.5%
Adj	80.98%	12.35%	3.96%	1.41%	0.51%	0.25%	0.16%	0.16%	<0.05%
Adv	83.84%	11.24%	3.67%	0.61%	0.42%	0.15%	0.03%	0.009%	<0.009%

Table 3: Statistics on SemCor: distribution of senses for nouns, verbs, adjectives and adverbs

words, the noun *"race"* is determined to have different senses in these two examples.

Reformulated, this principle becomes: if a word, in two different contexts, is paronymically related to the same words, then the word has similar meanings in the given contexts.

The translation into WordNet terminology cannot be done for all parts of speech, as there are no WordNet relations between verbs and nouns, therefore paronymic relations of the form *act* - *actor* cannot be extracted. This principle can be used for adjectives and adverbs, where a *pertainymy* relation is defined. The following rule is derived:

Rule SP3 If S1 and S2 are two synsets representing two senses of a given word, and if S1 and S2 have the same pertainym, then S1 and S2 can be collapsed together into one single synset S12.

Example: Senses #1 and #5 for the adverb *lightly* are:
S1 = {*lightly*} (*without good reason*)
pertainym ⇒ {*light#5*} (*psychologically light*)
S5 = {*lightly*} (*with indifference or without dejection*)
pertainym ⇒ {*light#5*} (*psychologically light*)

4 Probabilistic principles

Besides the principles presented in the previous section, the polysemy of WordNet can be reduced based on the frequency of senses and the probability of having particular synsets used in a text. By dropping synsets with very low probability of occurrence, we can reduce the number of senses a word might have.

We need (1) a distribution of sense frequencies for the different parts of speech and (2) a method of deriving the probability of a synset occurring in a text, starting with the probabilities of its component words.

To determine the distribution of word sense frequencies, we used again SemCor, as the only corpus available in which all words are sense tagged using WordNet. Table 3 show the sense distributions for nouns, verbs, adjectives and adverbs.

Let us denote a synset with S = { W_{i_1}, W_{i_1} ... , W_{i_n}}, meaning that the synset is composed by words having senses $i_1, i_2 \ldots , i_n$. If we denote with P_{i_k} the probability of occurrence of a word having sense i_k, then the probability of occurrence P_S for the synset S is equal with the summation of the probabilities of occurrence for the component words, i.e. $P_S = \sum_{k=1}^{k=n} P_{i_k}$.

In order to reduce the granularity of WordNet without introducing too much ambiguity, we use this formula together with probabilities derived from SemCor, and drop those synsets with a probability of occurrence P_S smaller than a given threshold. The following rule is derived:

Rule PP1 If S is a synset S = { W_{i_1}, W_{i_1} ... , W_{i_n}} with the probability of occurrence $P_S = \sum_{k=1}^{k=n} P_{i_k} < Max_P$ then S can be considered as a very rarely occurring synset and it can be dropped.

Example: The noun synset S = { *draft#11, draught#5, drawing#6*} (*the act of moving a load by drawing or pulling*) has the probability of occurrence $P_S = P_{11} + P_5 + P_6 = 0.1\% + 0.98\% + 0.52\% = 1.6\%$. For Max_P set to 2.0, this synset can be dropped.

Note that this way of computing the probability of a synset does not make reference to the component words themselves, but to their senses, and thus we do not have to deal with the problem of data sparseness that would result from the limited size of the corpus.

5 Applying the principles on WordNet

We applied these semantic and probabilistic principles on WordNet and generated two new versions, called *EZ.WordNet.1* and *EZ.WordNet.2*.

The semantic principles resulted in collapsed synsets, while the probabilistic principles determined which synsets can be dropped. By applying these rules, we obtain a reduction in the number of synsets, and implicitly a reduction in the number of word senses.

There are two variables used by the reduction rules: the K minimum number of common synonyms among two synsets, as required by *Rule SP1.3*, and Max_P, which is the maximum probability threshold for *Rule PP1*. Depending on the values selected for these parameters, one can obtain sense inventories closer to the original WordNet, but with a smaller reduction in polysemy, or versions of WordNet with a higher reduction in polysemy but with more synsets modified respect to WordNet.

Part of speech	Rule applied														Total reduced word senses
	SP0		SP1.1		SP1.2		SP1.3		SP2		SP3		PP1		
	(s)	(w)	(s)	(w)	(s)	(w)	(s)	(w)	(s)	(w)	(s)	(w)	(s)	(w)	
EZ.WordNet.1 (K=3 Max_P=2)															
Noun	–	–	349	743	216	216	100	328	2	3	–	–	1074	1142	2432
Verb	244	252	117	242	131	131	71	226	7	7	–	–	889	969	1827
Adj	–	–	115	244	–	–	29	90	8	8	12	12	931	1018	1372
Adv	–	–	23	52	–	–	1	3	6	6	96	100	84	85	246
EZ.WordNet.2 (K=2 Max_P=5)															
Noun	–	–	349	743	216	216	973	2015	2	3	–	–	3159	3344	6321
Verb	244	252	117	242	131	131	677	1257	7	7	–	–	1615	1767	3656
Adj	–	–	115	244	–	–	356	714	8	8	12	12	1628	1795	2773
Adv	–	–	23	52	–	–	47	92	6	6	96	100	158	165	418

Table 4: Reductions in the number of synsets ((s) columns) and in the number of word senses ((w) columns) obtained for each rule.

Part of speech	Polys. words	Monos. words	*EZ.WordNet.1*			*EZ.WordNet.2*		
			Total senses polys.words	Avg.polys. (- monos)	Avg.polys. (+ monos)	Total senses polys.words	Avg.polys. (- monos)	Avg.polys. (+ monos)
Noun	12562	81911	31975	2.54	1.20	28086	2.23	1.16
Verb	4565	5753	14487	3.17	1.96	12658	2.77	1.78
Adj	5372	14797	13713	2.55	1.41	12312	2.29	1.34
Adv	748	3797	1635	2.18	1.19	1463	1.95	1.15
TOTAL	23247	106258	61810	2.65	1.29	54429	2.34	1.24

Table 5: Statistics on EZ.WordNet.1 and EZ.WordNet.2: number of senses for polysemous words, average polysemy for polysemous words only and for all words

Two sets of values were chosen, and consequently we obtained two versions of WordNet:

- *EZ.WordNet.1*, for $K = 3$ and $Max_P = 2$.
- *EZ.WordNet.2*, for $K = 2$ and $Max_P = 5$.

Table 4 shows, for each of these versions, the reduction obtained in number of synsets, respectively in number of senses, for each of the semantic and probabilistic rules. *Rule SP0* is applicable only for verbs and it corresponds to the *VERBGROUP* pointers already defined in WordNet. Combining the information derived from this table with the statistics shown in Table 1, we can calculate the average polysemy for the two new versions of WordNet. Table 5 shows the total senses for the polysemous words, as computed in *EZ.WordNet.1* and *EZ.WordNet.2*, as well as the average polysemy computed for these sense inventories.

A much more important and interesting result would be to measure the reduction in the number of senses for the words commonly used by people, i.e. to determine the reduction in the polysemy of the words in SemCor. We can also make use of this corpus to determine the ambiguity introduced by the new sense inventories with respect to the original WordNet sense tagging.

We compute two measures on SemCor: (1) *the average polysemy* determined as the total number of senses for the words in SemCor, with respect to *EZ.WordNet* 1 or 2, divided by the total number of words, as it results from Table 2; and (2) *error rate*, defined as the total number of words from SemCor that are not defined anymore in the new WordNet versions, divided by the total number of words in SemCor. [3] Table 6 shows these figures computed for *EZ.WordNet.1* and *EZ.WordNet.2*.

Interpretation of results. Looking at both Table 1 and 5, it results an average reduction in the number of senses for polysemous words of 9% for *EZ.WordNet.1*, respectively 20% for *EZ.WordNet.2*, with respect to the original WordNet.

There is a difference between the polysemy of the words in a dictionary and the polysemy of the words actually used by humans (the words in the *active* vocabulary). This difference is clearly shown in Table 2: the average polysemy of the words in a common text like SemCor is much higher than the average polysemy of the words in a dictionary. Hence, a more representative result is obtained by comparing the average polysemies obtained with different sense inventories on a corpus, such as SemCor.

From Table 6, it results a reduction of 26% in

[3] An observation to be made here relates to the modality of calculating the error rate; the fact that one synset is missing from EZ.WordNet (the effect of dropping synsets) is naturally considered to introduce an error ϵ; the fact that one synset is in EZ.WordNet, but is fused with another synset (the effect of collapsing synsets) should not be considered as introducing the same error ϵ, but $c\epsilon$, with c in the range [0..1] (the synset is not missing, it is only fused with another synset). We consider the worst case, i.e. c=1, and thus the error rates reported here are a maximum over the possible error rates.

Part of speech	Total word occ.	*WordNet* Total senses	*WordNet* Avg. polys.	*EZ.WordNet.1* Total senses	*EZ.WordNet.1* Avg. polys.	*EZ.WordNet.1* Missing senses	*EZ.WordNet.1* Error rate	*EZ.WordNet.2* Total senses	*EZ.WordNet.2* Avg. polys.	*EZ.WordNet.2* Missing senses	*EZ.WordNet.2* Error rate
Noun	88398	382765	4.33	316796	3.58	1461	1.65%	256178	2.89	4668	5.28%
Verb	90080	944368	10.48	668712	7.42	2879	3.1%	542629	6.02	6310	7.0%
Adj	35770	157751	4.41	119044	3.32	545	1.52%	101907	2.84	1366	3.81%
Adv	20595	55617	2.70	45928	2.23	200	0.97%	39732	1.92	818	3.97%
TOTAL	234843	1540501	6.55	1150480	4.89	5085	2.16%	940446	4.0	13162	5.6%

Table 6: Average polysemy and error rate obtained on SemCor for *EZ.WordNet.1* and *EZ.WordNet.2*. We also replicate, for comparison purposes, the total number of senses and average polysemy in WordNet, as shown in Table 2

polysemy, with an error rate of 2.16%. The second version has a larger error rate, namely 5.6%, but it also brings a larger reduction in polysemy of 39%. The error rates of 2.1% and 5.6% are acceptable as it is considered that the accuracy obtained by humans in sense tagging is not larger than 92-94%. Depending on the application, one of these versions or newly compiled versions of WordNet can be used.

6 Conclusions

One of the main problems associated with WordNet is its fine granularity. We presented a methodology for reducing the average polysemy of the words defined in WordNet. We derived a set of semantic and probabilistic rules, used to either collapse synsets very similar in meaning or drop synsets that are very rarely used. A new version of WordNet is obtained, leading to a reduction of 26% in the average polysemy of words, while introducing a small error rate of 2.1%, as measured on SemCor. An alternative version is also created, with a polysemy reduction of 39% and an error rate of 5.6%. These results are encouraging, as a coarse grained WordNet is known to be beneficial for a large range of applications.

Note: A preliminary version of this paper appears in (Mihalcea and Moldovan, 2001).

References

D.A. Cruse. 1986. *Lexical Semantics.* Cambridge Univ. Press.

C. Fellbaum. 1998. *WordNet, An Electronic Lexical Database.* The MIT Press.

J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of COLING-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada, August.

S. Harabagiu and D.I. Moldovan, 1998. *Knowledge Processing on an Extended WordNet*, pages 289–405. The MIT Press.

R. Mihalcea and D. Moldovan. 2001. EZ.WordNet: principles for automatic generation of a coarse grained WordNet. In *Proceedings of FLAIRS-2001 (to appear)*, Key West, May.

G. Miller, C. Leacock, T. Randee, and R. Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, New Jersey.

D Moldovan and R. Mihalcea. 2000. Using WordNet and lexical operators to improve Internet searches. *IEEE Internet Computing*, 4(1):34–43.

W. Peters, I. Peters, and P. Vossen. 1998. Automatic sense clustering in EuroWordNet. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 409–416, Granada, May.

P. Resnik and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–134.