Clustering WordNet Senses Utilizing Modified and Novel Similarity Metrics

CS229 Final Project Report

Christopher Thad Hughes, Sushant Prakash

(The Neural Activation Profile algorithm, which returned some of the more interesting results, was not completed in time for the poster presentation.)

Introduction

We approach the problem of clustering senses in Princeton's WordNet (Fellbaum 1998), a manually created dictionary/thesaurus which attempts to model the structure underlying human concepts. A synset, the fundamental unit in WordNet, is represented by a group of synonyms and a gloss definition, and is connected through a variety of semantic links, such as hypernyms (type-of) or meronyms (part-of), to other synsets. A particular word is associated with one or more synsets, each representing a particular sense of the word. While this electronic database provides an inventory with which to do Word Sense Disambiguation, the fine-grainedness of the senses – which sometimes even humans have trouble distinguishing between - has posed a problem in achieving reasonable performance.

Our goal is to learn to automatically cluster senses of a given word that are sufficiently close in meaning. We pose the problem as the decision, given two synsets, of whether or not those synsets belong in the same cluster. We train and test a binary logistic regression classifier on a set of hand-labeled clustering data from the GALE OntoNotes project, using a variety of synset similarity metrics as features. We used 85% of the nouns and verbs for training, and put the other 15% of nouns and verbs aside for testing. To obtain data points, we looked at all possible pairs of senses for each given word. For our train set, we had 2382 negative, 258 positive noun pairs, and 2956 negative, 990 positive verb pairs. For test, we had 823 negative, 89 positive noun pairs, and 482 negative, 235 positive verb pairs.

Previous work

Researchers have used a variety of measures on the information contained within WordNet to determine similarity. One simple heuristic is just the shortest path between the two synsets following semantic links. Slightly more complicated measures, such as (Wu and Palmer) take into account the Least Common Subsumer (LCS) of the synsets and their relative depths. First introduced by (Resnik), several measures, including (Lin) and (Jiang and Conrath), estimate probabilities over concepts from a standard corpus, and then compare the information content of the two synsets and their LCS.

While all of these rely on semantic links, a very important source of information, and the one most useful to humans performing the task, is the similarities in the glosses and examples of the synsets. The Lesk measure attempts to deal with this by counting the word overlaps. Banerjee and Pedersen (2002) modified the measure by also counting word overlaps present between synsets semantically connected to the original two, i.e. the hypernyms, the hyponyms, and so forth. Even with this modification, the measure remains somewhat crude and calls for improvement.

Our Features

Lesk Overlap Generalization (LOG)

One obvious shortcoming of the Lesk measure is its inability to deal with synonyms. Since it superficially measures overlap with word equality, glosses that express the same concept, but with different synonymic words, will appear to have low similarity. Having sense-tagged WordNet glosses would help immensely with this problem, but since that is not the case we must consider all combinations of the senses of the two words in question. Given a set S of similarity measures between synsets, we calculate a similarity vector v*ij* for each pair (i,j) where i exists in *Senses*(w1), j exists in *Senses*(w2). We then keep the vector v*max*, which is the vector that maximizes the most entries; this corresponds to an optimistic attitude for the relatedness between the two words. After doing this for each pair of words, we return the top N similarity vectors as the pairwise features for the synset.

This measure provides a great deal of flexibility over the previous Lesk measure, as it not only allows for high similarity between synonyms, but also takes into account soft relations between words that almost or somewhat mean the same thing. Moreover, as S, the set of synset similarity metrics is added to or improved, so will the performance of this measure increase.

Lesk with Information Overlap (LIO)

Two more shortcomings of the Lesk measure are clearly evident. Firstly, any overlap is treated the same, no matter the word(s) contained in the overlap. Even if the standard stop words ("the", "of", etc.) are filtered out, some word overlaps will just be more common than others, and will thus not imply as great a deal of similarity.

The second shortcoming involves the length of the overlap. There is an intuition that a long overlap should be weighted very highly, arguably even more than the sum of several short overlaps. In the Lesk formulation, this intuition is implemented by squaring the overlap length before adding it to the cumulative score. This decision, however, seems quite arbitrary as it does not have any theoretical foundation beyond that of the original, vague intuition.

To deal with both of these problems, we propose that a language model be built to estimate the probabilities of sequences of words. We can then weight an overlap w1,w2,..wn with the information of that sequence, defined as -log(p (w1,w2,..wn)). Notice that this will give less weight to common sequences of words (which would be more likely to overlap), and more weight to rare sequences. This method also captures the intuition behind higher weighting of long sequences, as a sequence of several words will naturally have a much lower probability than the occurrence of a single word. For our experiments we trained a unigram, bigram, and trigram language model on all the WordNet glosses and examples to use for this method.

Relation Weight Learning (RWL)

With Banerjee and Pedersen's modification to the Lesk algorithm, overlaps in the synsets of the hypernyms, meronyms, and other semantically related nodes, are also taken into account, but then these are summed up either uniformly or by hand-chosen weights. We modified their WordNet Similarity package to output each of the measures separately so that we could perform gradient descent to automatically learn the weightings of the different relations. We applied this technique not only to the Similarity::lesk measure, but also the Similarity::vector_pairs measure which also looks at the similarity of extended glosses (Consequently a significant bug in vector_pairs was found and reported; Siddharth Patwardhan put out a new release the next day).

Classifier Results

	Train Max F-Score	Test F-Score
Nouns: Original Lesk Measure	25.29%	22.15%
Nouns: Lesk with RWL	32.98%	23.45%
Nouns: Lesk with Information Overlap	29.64%	21.70%
Nouns: LOG with Original Metrics	39.24%	35.31%
Nouns: RWL with Original Metrics	38.24%	28.57%
Nouns: All Original Similarity Metrics	31.75%	30.19%
Verbs: Original Lesk Measure	40.06%	48.43%
Verbs: Lesk with RWL	42.38%	47.56%
Verbs: Lesk with Information Overlap	41.40%	48.11%
Verbs: LOG with Original Metrics	42.94%	45.58%
Verbs: RWL with Original Metrics	43.42%	48.65%
Verbs: All Original Similarity Metrics	41.67%	49.52%

Looking at the above table, we immediately see an anomaly in the F-scores for verbs: the test set performance is better than the train set performance. We are not sure why this is the case, but we suspect it has to do with the fact that we have relatively few data points, and thus may not be getting a representative sample in either or both sets. In general we can see that the verbs have a higher score than the nouns, most likely caused by the fact that they have a higher rate of clustering.

Modifying the Lesk measure with the Information Overlap (LIO) did not help training performance much and slightly hurt test performance in both cases. One reason for this may be that the language model was not trained well enough from only the WordNet glosses. A supplemental corpus could be used to obtain more data. In addition, a more complex

language model, incorporating higher n-grams or utilizing parses could help. It is a bit of a circular problem in the sense that a proper lexical resource would better allow computers to "understand" language, but the computers need to "understand" the language in the gloss definitions to build the proper lexical resource.

Lesk with learned weights for the different relations (RWL) turned out to increase training performance a notable amount for nouns, but only slightly helped the test performance, and hurt a little bit for the verbs. Disappointingly, when we substituted this modified Lesk measure into the ensemble of metrics, performance actually decreased, a likely cause being overfitting of the training data.

Generalizing the overlaps for Lesk (LOG) seems promising considering the improved performance obtained on nouns. However, it is the most computationally intensive of our measures due to it computing similarities for all pairs of senses for all pairs of words between the glosses.

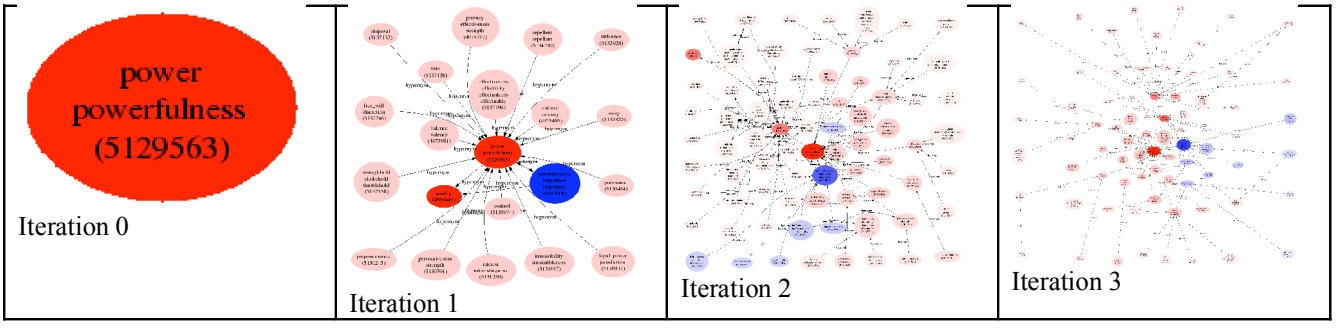
Neural Activation Profile (NAP)

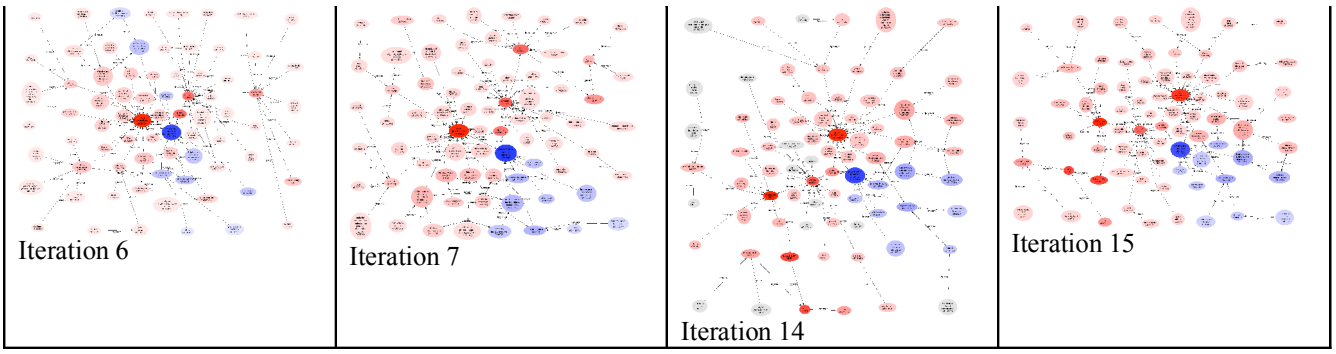
Late in the project, we had the insight that a concept is intricately connected to many other concepts, and that WordNet models at least some of this structure using a graph of synsets and the semantic edges between them. Interestingly, most of the work we have seen on measuring word sense similarity, including many of the measures described above, assumes (or even imposes) a tree-like structure on the synset graph, including a root node and a strong inheritance hierarchy. Our insight was that moving away from the tree mentality and towards a true graph or network might be useful.

We were partially inspired by our simple understanding of the neurons in the brain, which seem to function as at network that passes energy between the nodes in a complex feedback loop. We imagined interpreting WordNet as a similar structure, with the synset nodes representing neurons and the edges between them representing synapses.

With this analogy in mind, we developed the concept of a Neural Activation Profile (NAP). An NAP is a vector of real numbers representing the activation energy for every single node in the WordNet graph. For any synset, we can compute the synset's NAP by initializing the energy of each node in the graph to zero, and then assigning a positive (unity) activation energy to that synset. We then iteratively spread the energy from the initial node throughout the entire WordNet graph. During each iteration, each node that contains energy keeps some fraction of its energy for itself, and passes the rest of its energy to its neighbors. (This process of iteratively passing information to neighbors is realized efficiently using an algorithm similar to the Bellman-Ford shortest path algorithm). We say the algorithm has converged with the angle between two successive NAP vectors (energies as each node) dips below a threshold. Even with a WordNet graph of over 100,000 nodes, convergence typically took between 20-50 iterations, in total lasting less than a second in our Java implementation. The edges over which the energy is passed can be arbitrarily weighted; we used unity weights for all edge types except antonym edges, which were weighted with negative one. This means that antonyms of the starting synset can receive negative energy and pass it onward.

The figures below (generated by producing GraphViz .dot files from our Java program) show the nodes with the highest activation energy over the first few iterations of NAP computation for the word sense "power, powerfulness." Red color is used for positive energy, blue for negative, and intensity of the color shows the amount of energy present at the node. Only the 75 nodes with the highest energy are displayed, although many thousands of others have non-zero energy after the first iterations.

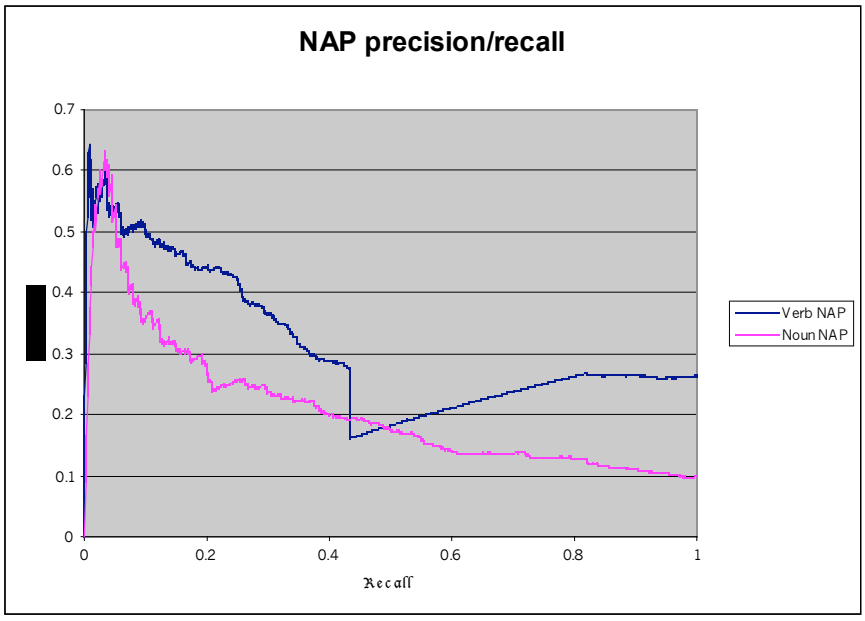




Once we have computed a NAP vector for two different word senses, we can compare the senses by computing the cosine similarity between the NAP vectors. The advantage of using this method, as opposed to previous methods that examine only a single path between senses, is that many different paths of influence between nodes are taken into account.

Neural Activation Profile Results

The classification results based on NAP similarity are interesting in that they reveal ambiguities in the word sense clustering task. Many of the false positives generated by the NAP similarity algorithm show examples in which the hand-labeled test data has been mislabeled. In fact, the top 10 word sense pairs predicted to be most similar by the NAP similarity are hand-labeled as not being similar senses, although we believe that some of them are (see Appendix). The maximum f-score achieved for nouns using only the NAP similarity was 0.28 for nouns and 0.418 for verbs. Precision and recall curves for the NAP similarity are shown below.



The tables in Appendix A show the best examples of false positives, true positives, false negatives and true negatives computed with the NAP similarity measure. Interestingly, the false positives comprised the word sense pairs with the overall highest NAP scores. All of the examples below had an angle of zero-degrees between their NAP vectors, to within the precision of the machine. However, one has to wonder if a "basketball center" and a "hockey center" represent two different senses of the word "center." Similar reasoning would seem to imply that "wheat bread" and "white bread" are two different uses of the word "bread." The same can be said for many of the rows in the table below. However, the "rent" row in the table below does actually represent two different (and opposing) senses of the word "rent." Note that in the "focus" row, the implied row was also present.

The table below lists the word sense pairs with the highest NAP similarities that were marked in the hand-labeled data as being the same word sense. We argue that the similarity in sense between these sense pairs and most of the sense pairs from the "false positives" table above are not that different, meaning that many of the "false positives" are actually mislabeled data. Again, all of the sense pairs in the table below had a difference of zero degrees between their NAP vectors.

The table below shows the sense pairs that were labeled as being the same sense, but that had the lowest NAP similarity scores. We realize that all of theses sense groupings are appropriate that that the computed NAP similarity does not reflect the similarity of the senses. These truly are false negatives.

The senses pairs below are those with the highest difference in NAP similarity that were also marked in the hand-labeled data as not being the same sense of the word in question. We are quite happy that the NAP similarity measures for all these sense pairs are different, because they do indeed represent different word senses. One of the most interesting examples below is the sense pair for "wear," which contains two senses with opposite meaning.

Interestingly, the sense pairs with the highest NAP differences are all verb sense pairs. We hypothesize that this is because WordNet imposes a tree-like hypernym hierarchy on all nouns, having them all descend from a single root node. This means that activation energy trickles up to the root nodes in the noun hierarchy and then back down again, where some of it moves to completely unrelated or even antonym nodes. The verb hypernym hierarchy is much weaker, and because the verbs don't all join at a single root, the NAP vectors for unrelated nodes don't loose energy to the upper nodes in the hierarchy.

NAP Analysis

One of the most obvious facts about the performance of the NAP sense clustering is that the recall seems to be poor. Most of the false positive assertions of similarity are somewhat excusable; however, the false negatives are clearly examples of word sense similarity that was missed by the NAP similarity measure.

Future work

There are severals ways to try to improve on this work, including the better language model for LIO and efficiency speed-ups for LOG. For NAP, one things we would like to try is to automatically learn edge weights for propagating energy. While the function may not be differentiable with respect to those parameters, eliminating the possibility of Gradient Descent or Newton's Method, approximating those algorithms with the Secant Method could provide better results.

Acknowledgements

Thanks to Rion Snow for an initial discussion about the project and Professor Dan Jurafsky for providing the hand-labeled data.

References

C. Fellbaum. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

T. Chklovski and R. Mihalcea. 2003. Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. *Proc. Of RANLP-2003.*

Mihalcea, R. & D. Moldovan. 2001. Automatic generation of a coarse grained WordNet. In [N-WordNet] (pp. 35-41).

Pedersen, T., Patwardhan S., and Michelizzi J. 2004. WordNet::Similarity – Measure the Relatedness of Concepts. *Dem. Of the Human Language Technology Conf.* 2004.

Resnik, P. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. JAIR 11, pp. 95-130.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proc. Of SIGDOC-1986.*

Appendix A: NAP performance examples

NAP false positives

Sense 1	Sense 1 gloss	Sense 2	Sense 2 gloss
center3#16	the position on a hockey team of the player who participates in the face off at the beginning of the game	center1#18	a position on a basketball team of the player who participates in the jump that starts the game
field1#3	somewhere (away from a studio or office or library or laboratory) where practical work is done or data is collected; "anthropologists do much of their work in the field"	field#6, field of operation#1, line of business#2	a particular kind of commercial enterprise; "they are outstanding in their field"
focus#4, focal point#2, nidus#1	a central point or locus of an infection in an organism; "the focus of infection"	focus#7	a fixed reference point on the concave side of a conic section
acknowledge2#5	accept as legally binding and valid; "acknowledge the deed"	acknowledge9#6, recognize2#1, recognise2#8, know6#6	accept (someone) to be what is claimed or accept his power and authority; "The Crown Prince was acknowledged as the true heir to the throne"; "We do not recognize your gods"
rent1#1, lease#1	let for money; "We rented our apartment to friends while we were abroad"	rent#4, hire1#2, charter1#1, lease1#2	hold under a lease or rental agreement; of goods and services
remember7#5	mention favourably, as in prayer; "remember me in your prayers"	commend1#5, remember#6	mention as by way of greeting or to indicate friendship; "Remember me to your wife"
report#1, describe1#2, account2#3	to give an account or representation of in words; "Discreet Italian police described it in a manner typically continental"	report1#4	make known to the authorities; "One student reported the other to the principal"
report1#4	make known to the authorities; "One student reported the other to the principal"	report13#5, cover2#8	be responsible for reporting the details of, as in journalism; "Snow reported on China in the 1950's"; "The cub reporter covered New York City"

NAP true positives

Sense 1	Sense 1 gloss	Sense 2	Sense 2 gloss
car3#4, gondola3#3	the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant	cable car#1, car4#3	a conveyance for passengers or freight on a cable railway; "they took a cable car to the top of the mountain"
car2#5, elevator car#1	where passengers ride up and down; "the car was on the top floor"	cable car#1, car4#3	a conveyance for passengers or freight on a cable railway; "they took a cable car to the top of the mountain"
close up1#3, close15#10	unite or bring into contact or bring together the edges of; "close the circuit"; "close a wound"; "close a book"; "close up an umbrella"	close14#11	bring together all the elements or parts of; "Management closed ranks"
point11#8	mark with diacritics; "point the letter"	point13#7	mark (Hebrew words) with diacritics
point11#8	mark with diacritics; "point the letter"	point12#9	mark (a psalm text) to indicate the points at which the music changes
report#1, describe1#2, account2#3	to give an account or representation of in words; "Discreet Italian police described it in a manner typically continental"	report13#5, cover2#8	be responsible for reporting the details of, as in journalism; "Snow reported on China in the 1950's"; "The cub reporter covered New York City"
hang glide#1, soar3#2	fly by means of a hang glider	sailplane#1, soar2#5	fly a plane without an engine

switch over#1, switch3#1, exchange1#3	change over, change around, as to a new order or sequence	interchange#3, tack2#6, switch#7, alternate2#4, flip#10, flip-flop#1	reverse (a direction, attitude, or course of action)
mailman#1, postman#1, mail carrier#1, letter carrier#1, carrier2#7	a man who delivers the mail	carrier1#8, newsboy#1	a boy who delivers newspapers

NAP false negatives

Sense 1	Sense 1 gloss	Sense 2	Sense 2 gloss	NAP angle
focus#2	cause to converge on or toward a central point; "Focus the light on this image"	focus1#5, focalize1#4, focalise1#4, sharpen3#4	put (an image) into focus; "Please focus the image; we cannot enjoy the movie"	3.01
agreement1#3, accord#1	harmony of people's opinions or actions or characters; "the two parties were in agreement"	agreement#6	the verbal act of agreeing	2.765
situate#2, fix1#10, posit#1, deposit1#3	put (something somewhere) firmly; "She posited her hand on his shoulder"; "deposit the suitcase on the bench"; "fix your eyes on this spot"	fixate1#3, fix#8	make fixed, stable or stationary; "let's fix the picture to the frame"	2.702
agreement#2, correspondence2#2	compatibility of observations; "there was no agreement between theory and measurement"; "the results of two tests were in correspondence"	agreement1#3, accord#1	harmony of people's opinions or actions or characters; "the two parties were in agreement"	2.667
fasten1#1, fix#2, secure1#2	cause to be firmly attached; "fasten the lock onto the door"; "she fixed her gaze on the man"	situate#2, fix1#10, posit#1, deposit1#3	put (something somewhere) firmly; "She posited her hand on his shoulder"; "deposit the suitcase on the bench"; "fix your eyes on this spot"	2.47
integrate1#3	become one; become integrated; "The students at this school integrate immediately, despite their different backgrounds"	integrate#1, incorporate#1	make into a whole or make part of a whole; "She incorporated his suggestions into her proposal"	2.447
assume#2, adopt1#3, take on1#2, take over1#2	take on titles, offices, duties, responsibilities; "When will the new President assume office?"	assume#3, acquire#2, adopt1#4, take on#1, take1#5	take on a certain form, attribute, or aspect; "His voice took on a sad tone"; "The story took a new turn"; "he adopted an air of superiority"; "She assumed strange manners"; "The gods assume human or animal form in these fables"	2.417
peace#1	the state prevailing during the absence of war	peace#5, peace treaty#1, pacification#2	a treaty to cease hostilities; "peace came on November 11th"	2.394
peace#1	the state prevailing during the absence of war	peace1#2	harmonious relations; freedom from disputes; "the roommates lived in peace together"	2.389

NAP true negatives

Sense 1	Sense 1 gloss	Sense 2	Sense 2 gloss	NAP angle
head4#8	form a head or come or grow to a head; "The wheat	head1#4, head up#1	be the first or leading member of (a group) and excel; "This student	3.002

	headed early this year"		heads the class"	
close6#8, shut6#2	become closed; "The windows closed with a loud bang"	close1#13	be priced or listed when trading stops; "The stock market closed high this Friday"; "My new stocks closed at $59 last night"	2.99
rise#4, lift1#12, rear#3	rise up; "The building rose before them"	move up1#2, rise6#8	be promoted, move to a better position	2.99
rise9#11, jump1#7, climb up2#2	rise in rank or status; "Her new novel jumped high on the bestseller list"	move up1#2, rise6#8	be promoted, move to a better position	2.99
scale#1	measure by or as if by a scale; "This bike scales only 25 pounds"	scale1#8	size or measure according to a scale; "This model must be scaled down"	2.99
rise#4, lift1#12, rear#3	rise up; "The building rose before them"	arise#3, rise5#3, uprise3#4, get up1#1, stand up#1	rise to one's feet; "The audience got up and applauded"	2.99
sharpen8#7, taper#2, point#12	give a point to; "The candles are tapered"	point2#10	be positionable in a specified manner; "The gun points with ease"	2.98
break#42, wear#7, wear out#2, bust#4, fall apart#2	go to pieces; "The lawn mower finally broke"; "The gears wore out"; "The old chair finally fell apart completely"	wear#6, hold out1#3, endure1#5	last and be usable; "This dress wore well for almost ten years"	2.97
come11#16, add up1#1, amount#3	develop into; "This idea will never amount to anything"; "nothing came of his grandiose plans"	total#1, number#1, add up3#3, come12#15, amount1#2	add up in number or quantity; "The bills amounted to $2,000"; "The bill came to $2,000"	2.97