A Study on the Contribution of WordNet in Supplementing Corpus Statistics for a Language Modeler

Kathleen L. Go
De La Salle University
Taft Avenue,
Metro Manila
go.kathleen@gmail.com

Solomon L. See
De La Salle University
Taft Avenue,
Metro Manila
sees@dlsu.edu.ph

ABSTRACT
Statistical language modeling (SLM) is a technique used to capture regularities of natural language [5]. One of the issues in SLM is the “curse of dimensionality” wherein word sequences on which the model will be used are likely to be different from those seen during training [2]. To address the issue, WordNet was incorporated to a trigram language modeler to supplement training data. In this approach, the LM makes use of WordNet to generate trigrams that are related to the unseen trigrams and uses them to reinforce the probabilities of the given. This paper focuses on: 1) the usefulness of Similarity and Relatedness measures in preventing the LM from using sequences that are neither fluent nor related to the given and 2) identifying the useful WordNet relationships in addressing the issue. Results show that the measures of Similarity and Relatedness are ineffective in their purpose while it has been confirmed that the IS-A relationship for nouns and verbs and the synonymy relationship for all parts-of-speech are the major contributors in addressing “curse of dimensionality”.

Categories and Subject Descriptors

General Terms
Algorithms, Languages

Keywords
Language Modeling, Statistical Models, WordNet, Measures of Similarity and Relatedness

1. INTRODUCTION
Statistical language modeling (SLM) is the attempt to capture regularities of natural language [5]. This is done by learning word sequences and related information from given training texts. SLM techniques are commonly used to improve performance of various natural language applications such as speech recognition (SR), machine translation (MT), and information retrieval (IR).

Even if widely researched, SLM still faces a number of issues. One of them is the “curse of dimensionality” issue wherein majority of the word sequences on which the model will be tested are likely to be different from those seen during training [2]. This leads to the assignment of low probabilities (i.e. scores) to sequences even if they are fluent.

There have already been attempts to address this problem. A number of works made use of smoothing techniques such as the one presented in [3] which made use of linear interpolation of trigram, bigram and unigram probabilities. [2] presents an approach that combines neural networks and smoothing techniques to a trigram model.

Another possible solution is to make use of linguistic resources such as WordNet. WordNet is an English lexical dictionary that covers nouns, verbs, adjectives and adverbs. Each syntactic category (i.e. part-of-speech) is organized into synonym sets (synsets), lists of synonymous word forms that are interchangeable in some syntax and have one underlying lexical concept, which are connected by different relation links [1]. By using WordNet for unseen sequences, it is possible to generate related sequences that were seen during training.

A study on the integration of WordNet to address the data sparseness of nouns in bigrams has already been conducted and is presented in [4]. The study covered the IS-A relationship of nouns and reported improvement in addressing the issue (i.e. measured by the decrease in model perplexity), although the improvement was below expectation.

A research was conducted on a trigram language modeler that considers all parts-of-speech and common relationships in WordNet to address the “curse of dimensionality” issue. The language modeler uses WordNet to supplement training data by using it to generate proxy trigrams for unseen trigrams (i.e. trigrams that do not have an exact match in the training data). A proxy trigram is a sequence that is formed by replacing at least one word in the given trigram with related words from WordNet. The scores of the proxy trigrams are then used to reinforce the score of their parent trigrams.

This paper focuses on two WordNet-related concerns. First is the effectiveness of WordNet Similarity and Relatedness measures in filtering out proxy trigrams that are neither fluent nor related to the given. The purpose of filtering out proxy trigrams is to prevent the language modeler from using them in the reinforcement of the trigram scores. This paper also focuses on
identifying the useful WordNet relationships for each POS in addressing the “curse of dimensionality”.

2. WORDNET

WordNet is an English lexical dictionary designed based on psycholinguistic theories of human lexical memory. It covers nouns, adjectives, adverbs and verbs as syntactic categories. Each syntactic category is organized into synonym sets (synsets), lists of synonymous word forms that are interchangeable in some syntax and have one underlying lexical concept, which are connected by different relation links. Table 1 shows the statistics of WordNet 2.1, the version used in this study. The statistics remain the same for WordNet 3.0, the latest version of WordNet.

![Table 1. WordNet 2.1 Statistics](image)

The common relationships used in WordNet are: synonymy, antonymy, hyponymy/hypernymy, and meronymy/holonymy.

Two expressions are said to be synonymous in a linguistic context C if the substitution of one for the other does not change the truth value. For example, *plank* and *board* are synonymous over the carpentry context because the two words can be used interchangeably in the said context. It must be noted that *board* has another context wherein its synonyms (e.g. *committee*) cannot be used to substitute for the carpentry context. Synonymy is symmetric such that if *x* is similar to *y*, then *y* is similar to *x*.

The antonym of a word *x* is sometimes not-*x*, but not always. For example, *heavy* and *light* are antonyms, but to say that something is not *heavy* does not imply that it must be *light*. This relationship is the main organizing principle for adjectives and adverbs in WordNet.

The hyponymy/hypernymy relationship is also called the subordination/superordination relationship, subset/superset relationship or IS-A relationship. A concept represented by synset {*x1, x2, ..., xn*} is said to be a hyponym of the concept represented by synset {*y1, y2, ..., yn*} if there are sentences constructed from such frames as An *x* is a (kind of) *y*. An example of this relationship is: *[maple]* is a hyponym of *[tree]* while *[tree]* is a hyponym of *[plant]*. On the other hand, hypernymy is the opposite of hyponymy such that *[tree]* is the hypernym of *[maple]* and *[plant]* is the hypernym of *[tree]*. This relationship is the main organizing principle for nouns and verbs in WordNet.

The meronymy/holonomy relationship is also called the part-whole relationship or HAS-A relationship. A concept represented by synset {*x1, x2, ..., xn*} is said to be a meronym of the concept represented by synset {*y1, y2, ..., yn*} if there are sentences constructed from such frames as *A y has an x (as a part)* or *An x is a part of y*. For example, *{finger}* is a meronym of *{hand}*.

3. MEASURES OF SIMILARITY AND RELATEDNESS

In order to get a numerical value of the similarity or relatedness of words from WordNet, similarity and relatedness measures were developed. These measures are found in the WordNet::Similarity package.

Measures of Similarity are the measures that are usually used to compute the similarity of nouns and verbs. It makes use of the IS-A relationship and is limited to measuring similarity of words belonging to the same POS only (i.e. noun-noun pairs and verb-verb pairs).

Aside from vertical (i.e. IS-A) relationships, concepts can also be related in other ways such as through the HAS-A and inter-POS (e.g. relatedness of a noun concept and adjective concept) relationships. These horizontal relationships are handled by Measures of Relatedness, which are applicable to any part-of-speech.

The measures of similarity are: *res*, *jen*, *lin*, *ich*, *wup*, and *path*. The first three are Information Content-Based measures wherein the specificity of the least common subsumer (LCS) of two concepts is measured. The LCS of concept A and concept B is their most specific common ancestor. The remaining are Path-based measures wherein similarity is based on the path lengths between concepts.

The measures of relatedness are: *hso*, *vector*, and *lesk*. *hso* is based on the shortest and least-changing path between two words. *vector* measures relatedness by getting the cosine of the gloss vectors of a pair of words while *lesk* scores the overlaps between their glosses.

4. LANGUAGE MODELER

The trigram language modeler used in this research addresses the “curse of dimensionality” issue by computing for a trigram’s reinforced score – combination of the original trigram score and the scores of its proxy trigrams. The formula used to compute for the reinforced score of a trigram *xyz* is shown below. If a trigram was seen in training (i.e. trigram has an exact match), proxy trigrams are not generated and its score automatically becomes the reinforced score.

\[
P(\text{xyz})_{\text{reinforced}} = (1 - \lambda) P(\text{xyz}) + \lambda \left( \sum_{i=1}^{l} P(\text{xyz}_i) + \sum_{j=1}^{m} P(\text{xyz}_j) + \sum_{k=1}^{n} P(\text{xyz}_k) \right)
\]

where:

- *xyz* = given trigram,
- *xyz*\(_i\) = proxy trigram with a trigram match,
- *xyz*\(_j\) = proxy trigram with a bigram match that contains a proxy word,
- *xyz*\(_k\) = proxy trigram with a unigram match that contains a proxy word,

\[
P(w_1, w_2, w_3) = P(\text{trigram}) \text{ or } P(\text{proxy trigram}) =
0.80 \times \text{frequency count of } w_1, w_2, w_3 / \text{frequency count of } w_1 +
0.14 \times \text{frequency count of } w_2, w_3 / \text{frequency count of } w_2 +
\]
(0.099 * frequency count of \(w_i\)) / (total # of words seen) + 0.001

\[
\lambda = 0.9, \quad \frac{1}{m} \sum_{i=1}^{m} P(x_i|z_i) <= 0.9,
\]

\[
\lambda = 0.01, \quad \frac{1}{n} \sum_{k=1}^{n} P(y_k|z_k) <= 0.006
\]

The purpose of \(\lambda\) is to prevent non-fluent trigrams from being reinforced too much in case it has a proxy trigram that is fluent. \(\lambda\) was derived by getting the value with the highest performance in the ranking of the fluency of parallel sentences.

As shown in the formula, the total scores for each match level are assigned different score limits (i.e. 0.9, 0.01, 0.006). This was done in order to prevent lower level matches from causing a high increase in the reinforced score. These values were derived through multiple tests conducted wherein different values were used to compute reinforced scores of corresponding fluent and non-fluent sets. The results were compared and the set of threshold values that gave a high perplexity reduction while at the same time maintained the perplexity distance was chosen. Perplexity is a metric for the actual performance of language models. A lower perplexity score indicates better generalization performance of the model. The perplexity of an \(n\)-gram language model is computed on a test set by getting the geometric average of the inverse of all the \(n\)-grams’ probabilities [5]. If the goal is to compare the performance of two language models (e.g. trigram language model vs. trigram language model with WordNet), the perplexity reduction between them must be computed. It is important to note that when comparing two models, both the training and test sets used must be the same.

5. SIMILARITY AND RELATEDNESS MEASURES AND THE “CURSE OF DIMENSIONALITY”

Using the language modeler described in the previous section, a sample set of 240 unseen trigrams was evaluated. With a training set consisting of about 1.8 million words, the language modeler had a perplexity value of 72.848. This shows a significant reduction in perplexity (i.e. 84.73%) when compared to the base language modeler (i.e. without WordNet), which had a perplexity value of 476.940 on the same sample set. However, WordNet is a large database with no specific domain. This means that some of the words taken from WordNet produced proxy trigrams that are neither fluent nor related to the original trigrams. Since there are no means of eliminating these proxy trigrams, they were used in trigram score reinforcement.

In order to have a better approximation of a trigram’s fluency, there must be a way to filter out the proxy trigrams that must not be used in trigram score reinforcement. An experiment was conducted to get a score that can be used for this purpose.

For each POS, 30 unseen trigrams, together with their proxy trigrams, were gathered. Due to the large number of proxy trigrams, it was decided that only the 10 most similar proxy trigrams for each trigram will be included in the experiment. The top proxy trigrams were identified using the similarity scores of their proxy words and corresponding words in the given trigrams. Since there are several measures applicable for each word pair and each measure has a different range of scores, the similarity scores of each word pair were normalized and averaged. For example, given the trigram “create 17 direct” and its proxy trigram “produce 17 direct”, the similarity scores of the word pair “create - produce” using all applicable measures for verbs are computed. The similarity scores, shown in Table 2, have an average normalized similarity score of 0.7949. This process is repeated for all proxy trigrams of a trigram in order to identify the top proxy trigrams (i.e. proxy trigrams having the highest average normalized similarity score). Table 3 shows the summary of the sample set.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Similarity Score</th>
<th>Normalized Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>hso</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>lesk</td>
<td>2410</td>
<td>0.0771</td>
</tr>
<tr>
<td>lin</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>path</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>res</td>
<td>6.0974</td>
<td>0.4872</td>
</tr>
<tr>
<td>vector</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>wup</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th># of Trigrams</th>
<th># of Proxy Trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>27</td>
<td>253</td>
</tr>
<tr>
<td>Adjective</td>
<td>28</td>
<td>235</td>
</tr>
<tr>
<td>Adverb</td>
<td>30</td>
<td>180</td>
</tr>
<tr>
<td>Verb</td>
<td>30</td>
<td>294</td>
</tr>
<tr>
<td>Total</td>
<td>115</td>
<td>962</td>
</tr>
</tbody>
</table>

Using the proxy trigrams, proxy sentences where generated by replacing the given trigrams in the original sentences. For example, the proxy trigram “produce 17 direct” can replace “create 17 direct” in the sentence “The expansion is expected to create 46 direct jobs.” giving the proxy sentence “The expansion is expected to produce 46 direct jobs.”.

Each of the proxy sentences were then manually evaluated by the authors for fluency. In evaluating the fluency of a sentence, the authors focused only on the trigrams containing the proxy words (i.e. the segment containing the proxy words). The criteria for fluency, which was based on [3], are 1) no strange word orderings and 2) no uncommon vocabulary used. When applied to the experiment, the first criterion means that the segment containing the proxy words must be structurally correct and coherent while the second criterion means that the proxy words used must be those commonly used in that context. For example, using these criteria, we can say that the sample proxy sentence presented above is fluent since the proxy word “produce” maintained the coherence and structure of the affected segment (i.e. “expected to produce 46 direct”) and is also commonly used in the same context.
In order to get the similarity score limit for each POS, a frequency distribution table was built using the average normalized similarity scores of proxy words found in proxy trigrams that produced fluent sentences. The lower bound of the range is the similarity score limit for that POS wherein proxy trigrams whose proxy words do not satisfy this limit will be filtered out. The similarity score limits for nouns, adjectives, adverbs, and verbs are 0.4, 0.6, 0.45, and 0.4, respectively. For example, the proxy trigram “produce 17 direct” will not be filtered out since the average normalized similarity score of “produce” is above the limit (i.e. 0.7949 > 0.4) for verbs.

To see if the similarity score limits are effective in filtering out proxy trigrams that are neither fluent nor related to the given, another experiment was conducted. This was done by comparing the amount of perplexity reduction when similarity score limits are used (i.e. 84.73%) and when they are not used. By doing so, we will be able to see how much was filtered out.

The experiment was conducted on the same sample set described at the start of this section. The sample set containing 240 trigrams had 19,149 proxy trigrams. Four sets of reinforced scores were computed such that for each set, only the similarity score limit for one POS is activated. By doing this, we will be able to see the full effects of the similarity score limit of each POS. The next step was to compute the perplexity and perplexity reduction of each set of reinforced score. Table 4 shows the results. Each column in the table indicates which of the similarity score limits is activated.

### Table 4. Comparison of Perplexity Reduction Caused by Similarity Score Limit Per POS

<table>
<thead>
<tr>
<th>POS</th>
<th>Noun</th>
<th>Adjective</th>
<th>Adverb</th>
<th>Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perplexity</td>
<td>72.939</td>
<td>76.798</td>
<td>72.848</td>
<td>80.145</td>
</tr>
<tr>
<td>% Perplexity Reduction</td>
<td>84.707</td>
<td>83.898</td>
<td>84.726</td>
<td>83.196</td>
</tr>
</tbody>
</table>

Since the difference in perplexity reduction when similarity score limits were used and when similarity score limits were not used is insignificant (i.e. about 1% for all similarity score limits), it can be concluded that WordNet Similarity and Relatedness measures are ineffective in filtering out proxy trigrams that are neither related to the given nor fluent for all parts-of-speech. For example, the proxy trigram “cope with exchange” of the trigram “cope with change” would not be filtered out even if it is not fluent since the word pair “exchange - change” has an average normalized similarity score of 0.5588 (i.e. greater than the 0.4 limit for nouns). It can also be concluded that these limits are not successful in aiding the language modeler compute for a better approximation of a trigram’s fluency.

### 6. WORDNET RELATIONSHIPS AND THE “CURSE OF DIMENSIONALITY”

In this section, we explore the contribution of each WordNet relationship in addressing the “curse of dimensionality” issue.

Table 5 shows the results of the manual rating of the fluency of proxy sentences described in the previous section. The columns display: 1) the total number of trigrams with proxy trigrams formed using a particular WordNet relationship, 2) total number of trigrams having at least one proxy trigram that produced a fluent sentence, 3) total number of proxy trigrams formed using a particular WordNet relationship, and 4) the total number of proxy trigrams that produced a fluent sentence.

### Table 5. Results of the Rating of Proxy Sentence Fluency Classified by POS and WordNet Relationship

<table>
<thead>
<tr>
<th>POS</th>
<th>Trigrams Total</th>
<th>with a Fluent Proxy Sentence Total</th>
<th>Fluent Proxy Sentence Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>84.707</td>
<td>72.939</td>
<td>66</td>
</tr>
<tr>
<td>Adjective</td>
<td>27</td>
<td>27 (77.78%)</td>
<td>232</td>
</tr>
<tr>
<td>Adverb</td>
<td>22 (75.86%)</td>
<td>21 (70%)</td>
<td>21 (70%)</td>
</tr>
<tr>
<td>Verb</td>
<td>22 (75.86%)</td>
<td>21 (70%)</td>
<td>21 (70%)</td>
</tr>
</tbody>
</table>

By looking at the table, we can learn which of the relationships for each POS are effective in producing fluent proxy sentences, and therefore, fluent proxy trigrams.

For all parts-of-speech, the most useful relationship is synonymy. The number of trigrams having at least one fluent proxy sentence produced by using the said relationship is the highest for each POS: 1) 16 out of 27 (59.26%) for nouns, 2) 21 out of 28 (75%) for adjectives, 3) 21 out of 30 (70%) for adverbs, and 4) 22 out of 30 (73.33%) for verbs. The said relationship also produced the highest number of fluent proxy sentences for each POS. For nouns and verbs, the IS-A relationship also proved to be useful. Another observation is that since the number of trigrams and proxy trigrams falling under the antonymy and HAS-A relationships are small compared to the others, they can be considered as not useful.

Table 6 shows a list of sample fluent proxy sentences while Table 7 shows a list of sample non-fluent proxy sentences. Proxy trigrams are underlined while the proxy words are also in bold.
The results in Table 8 agree with most of the results in the previous experiment. For nouns and verbs, the more useful relationships are IS-A and synonymy since these relationships produce the most number of trigrams with at least one exact-matching proxy trigram. For adjectives and adverbs, the synonymy relationship is more useful but due to the small number of matches for both parts-of-speech, the contribution of the antonymy relationship may be significant. This must be explored in a study to see the effects of the antonymy relationship in trigram score reinforcement. The said study was already conducted on the HAS-A relationship and it was shown that its contribution to trigram score reinforcement is insignificant giving only about 1% reduction in perplexity when proxy trigrams produced using the said relationship are included.

### Table 6. Sample Fluent Proxy Sentences for Each POS

<table>
<thead>
<tr>
<th>POS - WordNet Relationship</th>
<th>Replaced Word</th>
<th>Proxy Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun - Hyponymy</td>
<td>women</td>
<td>Employing <em>girls</em> mostly from Los Baños and Laguna, he has stores in Laguna and Metro Manila.</td>
</tr>
<tr>
<td>Adjective - Synonymy</td>
<td>hefty</td>
<td>Our baseline scenario is that the presidential camp will obtain a <em>sizeable</em> majority at the House of Representatives.</td>
</tr>
<tr>
<td>Adverb - Synonymy</td>
<td>just</td>
<td>“The joint venture with the family will only be the tip of the iceberg”, he said.</td>
</tr>
<tr>
<td>Verb - Hyponymy</td>
<td>create</td>
<td>The expansion is expected to <em>make</em> 46 direct jobs and 117 indirect jobs.</td>
</tr>
</tbody>
</table>

### Table 7. Sample Non-Fluent Proxy Sentences for Each POS

<table>
<thead>
<tr>
<th>POS - WordNet Relationship</th>
<th>Replaced Word</th>
<th>Proxy Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun - Hyponymy</td>
<td>women</td>
<td>Employing <em>miss</em> mostly from Los Baños and Laguna, he has stores in Laguna and Metro Manila.</td>
</tr>
<tr>
<td>Adjective - Synonymy</td>
<td>hefty</td>
<td>Our baseline scenario is that the presidential camp will obtain a <em>goodish</em> majority at the House of Representatives.</td>
</tr>
<tr>
<td>Adverb - Synonymy</td>
<td>just</td>
<td>“The joint venture with the family will <em>scarcely</em> be the tip of the iceberg”, he said.</td>
</tr>
<tr>
<td>Verb - Hyponymy</td>
<td>create</td>
<td>The expansion is expected to <em>refashion</em> 46 direct jobs and 117 indirect jobs.</td>
</tr>
</tbody>
</table>

In order to confirm the results of the manual experiment, the authors looked at the results of another experiment. The sample set consisted of 10,134 unseen trigrams with a total of 2,649,179 proxy trigrams. Since the aim is to check for the contribution of each WordNet relationship in addressing the “curse of dimensionality” issue, the total number of trigrams having at least one proxy trigram with an exact match was gathered for each POS and WordNet relationship. For example, the unseen trigram “sharp improvement in” is included in the count for the antonymy relationship under nouns because its proxy trigram “sharp decline in”, generated using the said relationship, had an exact match in the training data. Table 8 shows the summary of the results.

7. CONCLUSION

An approach that integrates WordNet features to a trigram language modeler to address the “curse of dimensionality” issue has been developed.

In this paper, we find that WordNet Similarity and Relatedness measures are not useful in eliminating proxy trigrams that are neither related to the given nor fluent. It can be concluded that the similarity score limits derived from these measures do not aid the language modeler in approximating the fluency of trigrams. One possible extension is to use an ontology specific to the domain the language modeler is intended to be used on. By using an ontology, the applicable proxy trigrams will be reduced and a new set of similarity score limit can be derived. These changes may prove to be more effective in aiding the language modeler.

We were also able to confirm that among the WordNet relationships, the synonymy and IS-A relationships are the major contributors in addressing the “curse of dimensionality” issue. This was first shown in a manual experiment but was later confirmed by automatic means showing the significant increase in the number of exact-matching trigrams through the proxy trigrams generated using these relationships.
Aside from having a small number of fluent proxy sentences and exact-matching proxy trigrams, it has been further confirmed that the proxy trigrams generated using the HAS-A relationship does not contribute in addressing the issue. One future work is to conduct the same study on each POS and WordNet relationship to see their actual effects in trigram score reinforcement. For example, even if the results show that proxy trigrams generated through the antonymy relationship have little contribution when compared to the others, they may still have a significant contribution in trigram score reinforcement for adjectives and adverbs. These additional experiments may also show that using only nouns and the IS-A relationship, as done in [4], is enough to address the “curse of dimensionality”.

8. REFERENCES


