Combining Lexical Resources for Contextual Synonym Expansion

Ravi Sinha and Rada Mihalcea
University of North Texas
ravisinha@my.unt.edu, rada@cs.unt.edu

Abstract

In this paper, we experiment with the task of contextual synonym expansion, and compare the benefits of combining multiple lexical resources using both unsupervised and supervised approaches. Overall, the results obtained through the combination of several resources exceed the current state-of-the-art when selecting the best synonym for a given target word, and place second when selecting the top ten synonyms, thus demonstrating the usefulness of the approach.

Keywords

lexical semantics, synonym expansion, lexical substitution

1 Introduction

Word meanings are central to the semantic interpretation of texts. The understanding of the meaning of words is important for a large number of natural language processing applications, including information retrieval [11, 10, 19], machine translation [4, 3], knowledge acquisition [7], text simplification, question answering [1], cross-language information retrieval [18, 5].

In this paper, we experiment with contextual synonym expansion as a way to represent word meanings in context. We combine the benefits of multiple lexical resources in order to define flexible word meanings that can be adapted to the context at hand. The task, also referred to as lexical substitution, has been officially introduced during SEMEVAL-2007 [16], where participating systems were asked to provide lists of synonyms that were appropriate for selected target words in a given context. Although it may sound simple at first, the task is remarkably difficult, as evidenced by the accuracies reported by the participating systems in SEMEVAL-2007.

In the experiments reported in this paper, we focus on the usefulness of different lexical resources – used individually or in tandem – for the purpose of contextual synonym expansion. We experiment with several resources to determine which ones provide the best synonyms for a given word in context.

2 Synonym expansion in context

Contextual synonym expansion, also known as lexical substitution [16], is the task of replacing a certain word in a given context with another, suitable word. See for example the four sentences from table 1, drawn from the development data from the SEMEVAL-2007 lexical substitution task. In the first sentence, for instance, assuming we choose *bright* as the target word, a suitable substitute could be *brilliant*, which would both maintain the meaning of the target word and at the same time fit the context.

Sentence	Target	Synonym
The sun was **bright**.	bright	brilliant
He was **bright** and independent.	bright	intelligent
His feature **film** debut won awards.	film	movie
The market is **tight** right now.	tight	pressured

Table 1: *Examples of synonym expansion in context*

We perform contextual synonym expansion in two steps: candidate synonym collection, followed by context-based synonym fitness scoring.

Candidate synonym collection refers to the task of collecting a set of potential synonym candidates for a given target word, starting with various resources. Note that this step does not account for the meaning of the target word. Rather, all the possible synonyms are selected, and further refined in the later step. For example, if we consider all the possible meanings of the word *bright*, it can be potentially replaced by *brilliant, smart, intelligent, vivid, luminous*.

The better the set of candidates, the higher the chance that one or more synonyms that are correct for the given context are found. Thus, one of the questions that we aim to answer in this paper is concerned with the role played by different lexical resources, used individually or combined, for the collection of good candidate synonyms.

Context-based synonym fitness scoring refers to picking the best candidates out of the several potential ones obtained as a result of the previous step. There are several ways in which fitness scoring can be performed, accounting for instance for the semantic similarity between the context and a candidate synonym, or for the substitutability of the synonym in the given context. Note that a factor that needs to be taken into account is the inflection of the words, which can influence the measures of fitness in context.

The better the measure of contextual fitness, the

International Conference RANLP 2009 - Borovets, Bulgaria, pages 404–410

higher the chance of identifying the correct synonyms from the input set of candidates. Hence, another question that we try to answer is the usefulness of different unsupervised and supervised methods in picking the best synonyms for a given target.

3 Lexical resources for candidate synonym selection

For the purpose of candidate synonym selection, we experiment with five different lexical resources, which are briefly described below. For all these resources, we perform several preprocessing steps, including removal of redundancies (i.e., making sure that all the candidates are unique), making sure that the target word itself is not included in the list, and also making sure that all the multiwords are normalized to a standard format (individual words separated by underscores). We also enforce that the part-of-speech of the candidates obtained from these resources coincide with the part-of-speech of the target word.

3.1 WordNet

WordNet [17] is a lexical knowledge base that combines the properties of a thesaurus with that of a semantic network. The basic entry in WordNet is a synset, which is defined as a set of synonyms. We use WordNet 3.0, which has over 150,000 unique words, over 110,000 synsets, and over 200,000 word-sense pairs. For each target word, we extract all the synonyms listed in the synsets where the word appears, regardless of its sense.

3.2 Roget's thesaurus

Roget is a thesaurus of the English language, with words and phrases grouped into hierarchical classes. A word class usually includes synonyms, as well as other words that are semantically related. We use the publicly available version of the Roget's thesaurus.[1] This version of Roget has 35,000 synonyms and over 250,000 cross-references. We query the online page for a target word, and gather all the potential synonyms that are listed in the same word set with the target word.

3.3 Encarta

Microsoft Encarta is an online encyclopedia and thesaurus resource, which provides a list of synonyms for each query word. We use Microsoft's online Encarta thesaurus[2] to extract direct synonyms for each target word, for a given part-of-speech.

3.4 TransGraph

TransGraph [5] is a very large multilingual graph, where each node is a word-language pair, and each edge denotes a shared sense between a pair of words. The graph has over 1,000,000 nodes and over 2,000,000 edges, and consists of data from several wiktionaries and bilingual dictionaries. Using this resource, and utilizing several "triangular connections" that place a constraint on the meaning of the words, we derive candidate synonyms for English words. Briefly, using the TransGraph triangular annotations, we collect the sets of all the words (regardless of language) that share a meaning with any of the meanings of the target word. From these sets, we keep only the English words, thus obtaining a list of words that have the property of being synonyms with the target word.

3.5 Lin's distributional similarity

Lin [14] proposes a method to identify distributionally similar words, which we use to derive corpus-based candidate synonyms. We use a version trained on the automatically parsed texts of the British National Corpus. From the ranked list of distributionally similar words, we select the top-ranked words in the ranking, up to a maximum of twenty if available.

To illustrate the diversity of the candidates that can be obtained from these resources, table 2 provides a snapshot of the potential candidates for the adjective *bright*. The average number of candidates selected from the different resources is 24, 19, 30, 48 and 15 from Encarta, Lin, Roget, TransGraph and WordNet respectively.

4 Methods for contextual fitness

Provided a set of candidate synonyms for a given target word, we need to select those synonyms that are most appropriate for the text at hand. We do this by using several methods to determine the fitness of the synonyms in context.

One aspect that needs to be addressed when measuring the fitness in context is the issue of morphological variations. For methods that look at substitutability in context using N-gram-based language models, we need to account for both the inflected as well as the non-inflected forms of a word. Instead, for methods that measure the similarity between a synonym and the input context, using the non-inflected form is often more beneficial. We use an online inflection dictionary[3] combined with a set of rules to derive all the inflected forms of the target word.

We describe below the three fitness algorithms used in our experiments.

4.1 Latent semantic analysis

One corpus-based measure of semantic similarity is latent semantic analysis (LSA) proposed by Landauer [13]. In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-document matrix $\mathbf{T}$ representing the corpus. For the experiments reported in this paper, we run the SVD operation on the entire English Wikipedia. Using

[1] http://www.thesaurus.com
[2] http://encarta.msn.com
[3] A large automatically generated inflection database (AGID) available from http://wordlist.sourceforge.net/

Resource	Candidates
WordNet (WN)	burnished sunny shiny lustrous undimmed sunshiny brilliant
Encarta (EN)	clear optimistic smart vivid dazzling brainy lively
Roget (RG)	ablaze aglow alight argent auroral beaming blazing brilliant
TransGraph (TG)	nimble ringing fine aglow keen glad light picturesque
Lin (LN)	red yellow orange pink blue brilliant green white dark

Table 2: *Subsets of the candidates provided by different lexical resources for the adjective bright*

LSA, we can calculate the similarity between a potential candidate and the words surrounding it in context. In our experiments, we consider a context consisting of the sentence where the target word occurs.

4.2 Explicit semantic analysis

Explicit semantic analysis (ESA) [6] is a variation on the standard vector-space model in which the dimensions of the vector are directly equivalent to abstract concepts. Each article in Wikipedia represents a concept in the ESA vector. The relatedness of a term to a Wikipedia concept is defined as the tf*idf score for the term within the Wikipedia article. The relatedness between two words is then defined as the cosine of the two concept vectors in a high-dimensional space. We can also measure the relatedness between a word and a text, computed by calculating the cosine between the vector representing the word, and the vector obtained by summing up all the vectors of the words belonging to the text. As before, we consider a context consisting of the sentence containing the target word.

4.3 Google N-gram models

The Google Web 1T corpus is a collection of English N-grams, ranging from one to five N-grams, and their respective frequency counts observed on the Web [2]. The corpus was generated from approximately 1 trillion tokens of words from the Web, predominantly English. We use the N-grams to measure the substitutability of the target word with the candidate synonyms, focusing on trigrams, four-grams, and five-grams. For this method, the inflection of the words is important, as discussed above, and thus we use all the possible inflections for all the potential candidates.

For each target instance (sentence), we collect the counts for all the possible trigrams, four-grams and five-grams that have the target word replaced by the candidate synonym and its inflections, at different locations.[4] As an example, consider the trigram counts, for which we collect the counts for all the possible sequences of three contiguous words containing the target word: two words before and the target word; one word before, the target word, and one word after; the target word and two words after.

From these counts, we build several language models, as described below:

1. 3gramSum. We only consider trigrams, and we add together the counts of all the inflections of a candidate synonym. For example, if the target word is *bright* and one candidate synonym is *smart*, then we consider all of its inflections, i.e., *smart, smarter, smartest*, put them in the sequence of trigrams at different locations, collect all the counts from the Google Web 1T corpus, and then finally add them all up. This number is used as the final count to measure the substitutability of the word *smart*. After collecting such scores for all the potential candidates, we rank them according to the decreasing order of their final counts, and choose the ones with the highest counts.

2. 4gramSum. The same as 3gramSum, but considering counts collected from four-grams.

3. 5gramSum. The same as 3gramSum and 4gramSum, but considering counts collected only for five-grams.

4. 345gramSum. We consider all the trigrams, four-grams and five-grams, and add all the counts together, for the candidate synonym and for all its inflections.

5. 345gramAny. We again consider the counts associated with all the trigrams, four-grams and five-grams for the candidate synonym along with its inflections, but this time rather than adding all the counts up, we instead select and use only the maximum count.

In all the models above, the synonyms ranking highest are used as candidate replacements for the target word.

5 Experiments and evaluations

For development and testing purposes, we use the dataset provided during the SEMEVAL-2007 Lexical Substitution task. The development set consists of 300 instances (sentences) and the test set consists of 1710 instances, where each instance includes one target word to be replaced by a synonym.

We use the same evaluation metrics as used for the lexical substitution task at SEMEVAL-2007. Specifically, we measure the precision and the recall for four subtasks: *best normal*, which measures the precision and recall obtained when the first synonym provided by the system is selected; *best mode*, which is similar to *best normal*, but it gives credit only if the first synonym returned by the system matches the synonym in the gold standard data set that was most frequently selected by the annotators; *out of ten (oot) normal*, which is similar to *best normal*, but it measures the precision and recall for the top ten synonyms suggested by the system; and *out of ten (oot) mode*, which is

[4] To query Google N-grams, we use a B-tree search implementation, kindly made available by Hakan Ceylan from University of North Texas.

similar to *best mode*, but it again considers the top ten synonyms returned by the system rather than just one. For *oot*, we do not allow our system to report duplicates in the list of best ten candidates. The metrics, detailed in [16] are summarized below.

Let us assume that H is the set of annotators, namely $\{h_1, h_2, h_3, ...\}$, and T, $\{t_1, t_2, t_3, ...\}$ is the set of test items for which the humans provide at least two responses. For each t_i we calculate m_i, which is the most frequent response for that item, if available. We also collect all r_i^j, which is the set of responses for the item t_i from the annotator h_j.

Let the set of those items where two or more annotators have agreed upon a substitute (i.e. the items with a mode) be denoted by TM, such that $TM \subseteq T$. Also, let $A \subseteq T$ be the set of test items for which the system provides more than one response. Let the corresponding set for the items with modes be denoted by AM, such that $AM \subseteq TM$. Let $a_i \in A$ be the set of system's responses for the item t_i.

Thus, for all test items t_i, we have the set of guesses from the system, and the set of responses from the human annotators. As the next step, the multiset union of the human responses is calculated, and the frequencies of the unique items is noted. Therefore, for item t_i, we calculate R_i, which is $\sum r_i^j$, and the individual unique item in R_i, say res, will have a frequency associated with it, namely $freq_{res}$.

Given this setting, the precision (P) and recall (R) metrics we use are defined below.

Best measures:

$$\text{P} = \frac{\sum_{a_i:t_i \in A} \frac{\sum_{res \in a_i} freq_{res}}{|a_i|} \frac{}{|R_i|}}{|A|}$$

$$\text{R} = \frac{\sum_{a_i:t_i \in T} \frac{\sum_{res \in a_i} freq_{res}}{|a_i|} \frac{}{|R_i|}}{|T|}$$

$$\text{mode P} = \frac{\sum_{bestguess_i \in AM} 1if\_best\_guess=m_i}{|AM|}$$

$$\text{mode R} = \frac{\sum_{bestguess_i \in TM} 1if\_best\_guess=m_i}{|TM|}$$

Out of ten (oot) measures:

$$\text{P} = \frac{\sum_{a_i:t_i \in A} \frac{\sum_{res \in a_i} freq_{res}}{|R_i|}}{|A|}$$

$$\text{R} = \frac{\sum_{a_i:t_i \in T} \frac{\sum_{res \in a_i} freq_{res}}{|R_i|}}{|T|}$$

$$\text{mode P} = \frac{\sum_{a_i:t_i \in AM} 1if\_any\_guess \in a_i=m_i}{|AM|}$$

$$\text{mode R} = \frac{\sum_{a_i:t_i \in TM} 1if\_any\_guess \in a_i=m_i}{|TM|}$$

For each setting, we calculate and report the F-measure, defined as the harmonic mean of the precision and recall figures.

5.1 Experiment 1: Individual knowledge sources

The first set of experiments is concerned with the performance that can be obtained on the task of synonym expansion by using the individual lexical resources: Roget (RG), WordNet (WN), TransGraph (TG), Lin (LN), Encarta (EN). Table 3 shows the results obtained on the development data for the four evaluation metrics for each lexical resource when using the LSA, ESA and N-gram models.

As a general trend, Encarta and WordNet seem to provide the best performance, followed by TransGraph, Roget and Lin. Overall, the performance obtained with knowledge-based resources such as WordNet normally tend to exceed that of corpus-based resources such as Lin's distributional similarity or TransGraph.

	RG	WN	TG	LN	EN
Best, normal					
LSA	1.55%	4.85%	2.40%	1.43%	3.80%
ESA	0.44%	3.40%	1.49%	2.42%	5.30%
3gramSum	3.04%	**9.09**%	8.63%	1.82%	7.64%
4gramSum	3.13%	8.02%	7.01%	2.95%	8.27%
5gramSum	2.97%	5.41%	4.06%	2.92%	5.07%
345gramSum	3.04%	**9.09**%	8.73%	1.82%	7.64%
345gramAny	3.04%	8.79%	7.78%	1.88%	7.44%
Best, mode					
LSA	1.50%	4.50%	4.00%	1.99%	5.45%
ESA	0.50%	3.50%	0.50%	3.50%	6.99%
3gramSum	3.54%	13.08%	12.58%	1.99%	11.59%
4gramSum	4.68%	11.90%	9.26%	3.63%	12.45%
5gramSum	4.77%	7.94%	5.80%	4.26%	7.94%
345gramSum	3.54%	13.08%	12.58%	1.99%	11.59%
345gramAny	3.54%	**13.58**%	11.59%	1.99%	11.59%
Oot, normal					
LSA	16.67%	21.39%	18.22%	14.93%	30.68%
ESA	15.77%	21.19%	17.47%	15.68%	26.73%
3gramSum	20.20%	21.62%	23.24%	15.90%	**32.86**%
4gramSum	15.26%	19.48%	20.98%	14.67%	30.45%
5gramSum	12.38%	17.45%	16.30%	12.59%	24.51%
345gramSum	20.50%	21.78%	23.68%	15.90%	**32.86**%
345gramAny	20.20%	21.68%	22.89%	15.80%	32.76%
Oot, mode					
LSA	19.98%	26.48%	21.53%	16.48%	36.02%
ESA	17.49%	25.98%	23.98%	19.48%	36.02%
3gramSum	25.71%	27.21%	29.71%	18.67%	**41.84**%
4gramSum	20.12%	23.75%	27.38%	19.12%	37.25%
5gramSum	16.36%	22.77%	22.22%	17.45%	29.66%
345gramSum	26.16%	27.21%	30.71%	18.67%	**41.84**%
345gramAny	25.71%	27.21%	29.26%	18.67%	41.29%

Table 3: *F-measures for the four scoring schemes for individual lexical resources (development data)*

Based on the results obtained on development data, we select the lexical resources and contextual fitness models that perform best for each evaluation metric. We then use these optimal combinations and evaluate their performance on the test data. Table 4 shows the F-measure obtained for these combinations of resources and models on the test set. Note that, in this experiment and also in experiment 2 below, adding four-grams and five-grams to three-grams either increases the performance, albeit slightly, or keeps it the same. However, in our experiments the absolute best performances occur in cases where the four-grams and five-grams do not really contribute much and hence the score after adding them is the same as that of only using three-grams. We only depict the three-grams scores in Table 4 and in Table 6 because it shows that less computation is enough for this particular problem and the extra processing to collect the higher order N-grams is not necessarily required.

5.2 Experiment 2: Unsupervised combination of knowledge sources

In the next set of experiments, we use unsupervised combinations of lexical resources, to see if they yield

Metric	Resource	Model	F-Measure
best, normal	WN	3gramSum	10.15%
best, mode	WN	345gramAny	16.05%
oot, normal	EN	3gramSum	43.23%
oot, mode	EN	3gramSum	55.28%

Table 4: *F-measure for the four scoring schemes for individual lexical resources (test data)*

improvements over the use of individual resources. We consider the following combinations of resources:

- Encarta and WordNet. All the candidate synonyms returned by both Encarta and WordNet for a target word.
- Encarta or WordNet. The candidate synonyms that are present in either WordNet or Encarta. This combination leads to increased coverage in terms of number of potential synonyms for a target word.
- Any Two. All the candidate synonyms that are included in at least two lexical resources.
- Any Three. All the candidate synonyms that are included in at least three lexical resources.

The results obtained on development data using these unsupervised resource combinations are shown in Table 5. Overall, the combined resources tend to perform better than the individual resources.

	EN and WN	EN or WN	Any2	Any3
Best, normal				
LSA	6.36%	3.25%	3.60%	7.09%
ESA	7.45%	3.30%	4.55%	7.83%
3gramSum	**10.08**%	8.59%	6.94%	8.93%
4gramSum	8.59%	8.33%	7.82%	9.00%
5gramSum	5.24%	5.96%	5.92%	9.07%
345gramSum	**10.08**%	8.59%	6.94%	8.93%
345gramAny	10.02%	7.44%	7.14%	9.27%
Best, mode				
LSA	5.99%	5.05%	4.50%	8.99%
ESA	9.99%	3.50%	5.99%	12.49%
3gramSum	13.08%	**14.13**%	8.59%	13.08%
4gramSum	11.09%	13.44%	11.40%	13.44%
5gramSum	6.34%	10.02%	9.03%	12.20%
345gramSum	13.08%	**14.13**%	8.59%	13.08%
345gramAny	**14.13**%	12.13%	9.04%	**14.13**%
Oot, normal				
LSA	20.27%	29.83%	32.88%	30.75%
ESA	20.23%	26.53%	29.28%	30.95%
3gramSum	19.15%	36.16%	32.66%	30.42%
4gramSum	18.02%	32.65%	30.25%	28.19%
5gramSum	17.64%	23.32%	24.31%	27.60%
345gramSum	19.15%	**36.21**%	32.76%	30.42%
345gramAny	19.15%	36.06%	33.16%	30.42%
Oot, mode				
LSA	25.03%	34.02%	38.02%	42.51%
ESA	25.53%	35.52%	37.51%	44.01%
3gramSum	23.67%	**45.84**%	41.84%	43.29%
4gramSum	22.26%	40.33%	38.24%	40.78%
5gramSum	21.68%	29.11%	31.19%	39.68%
345gramSum	23.67%	**45.84**%	41.84%	43.29%
345gramAny	23.67%	45.34%	42.34%	43.29%

Table 5: *F-measures for the four scoring schemes for combined lexical resources (development data)*

Based on the development data, we select the best combinations of unsupervised resources for each of the four scoring metrics, and evaluate them on the test data. Table 6 shows the results obtained on the test set for the selected combinations of lexical resources.

Metric	Resource	Model	F-Measure
best, normal	EN and WN	3gramSum	12.81%
best, mode	AnyThree	345gramAny	19.74%
oot, normal	EN or WN	3gramSum	43.74%
oot, mode	EN or WN	3gramSum	58.38%

Table 6: *F-measures for the four scoring schemes for combined lexical resources (test data)*

5.3 Experiment 3: Supervised combination of knowledge sources

As a final set of experiments, we also evaluate a supervised approach, where we train a classifier to automatically learn which combination of resources and models is best suited for this task. In this case, we use the development data for training, and we apply the learned classifier on the test data.

We build a feature vector for each candidate synonym, and for each instance in the training and the test data. The features include the id of the candidate; a set of features reflecting whether the candidate synonym appears in any of the individual lexical resources or in any of the combined resources; and a set of features corresponding to the numerical scores assigned by each of the contextual fitness models. For this later set of features, we use real numbers for the fitness measured with LSA and ESA (corresponding to the similarity between the candidate synonym with the context), and integers for the Google N-gram models (corresponding to the N-gram counts). The classification assigned to each feature vector in the training data is either 1, if the candidate is included in the gold standard, or 0 otherwise.

One problem that we encounter in this supervised formulation is the large number of negative examples, which leads to a highly unbalanced data set. We use an undersampling technique [12], and randomly eliminate negative examples until we reach a balance of almost two negative examples for each positive example. The final training data set contains a total of 700 positive examples and 1,500 negative examples. The undersampling is applied only to the training set.

The results obtained when applying the supervised classifier on the test data are shown in Table 7. We report the results obtained with four classifiers, selected for the diversity of their learning methodology. For all these classifiers, we use the implementation available in the Weka[5] package.

To gain further insights, we also carried out an experiment to determine the role played by each feature, by using the information gain weight as assigned by Weka to each feature in the data set. Note that ablation studies are not appropriate in our case, since the features are not orthogonal (e.g., there is high redundancy between the features reflecting the individual and the combined lexical resources), and thus we cannot entirely eliminate a feature from the classifier.

[5] www.cs.waikato.ac.nz/ml/weka/

Metric	NN	LR	DL	SVM
best, normal	1.6%	9.90%	**13.60**%	3.10%
best, mode	1.5%	14.80%	**21.30**%	4.30%
oot, normal	21.8%	43.10%	**49.40**%	32.80%
oot, mode	21.6%	56.50%	**64.70**%	40.90%

Table 7: *F-measure for a supervised combination of lexical resources (test data). NN=nearest neighbor; LR=logistic regression; DL=decision lists; SVM=support vector machines*

Feature	Weight
AnyTwo	0.1862
AnyThree	0.1298
EN and WN	0.1231
EN	0.1105
EN or WN	0.0655
LSA	0.0472
WN	0.0458
4gramSum	0.0446
5gramSum	0.0258
TG	0.0245
ESA	0.0233
RG	0.0112
LN	0.011
345gramSum	0.0109
3gramSum	0.0106
345gramAny	0.0104

Table 8: *Information gain feature weight*

Table 8 shows the weight associated with each feature. Perhaps not surprisingly, the features corresponding to the combinations of lexical resources have the highest weight, which agrees with the results obtained in the previous experiment. Unlike the previous experiments however, the 4gramSum and 5gramSum have a weight higher than 3gramSum, which suggests that when used in combination, the higher order N-grams are more informative.

6 Related work

There are several systems for synonym expansion that participated in the SEMEVAL-2007 lexical substitution task [16]. Most of the systems used only one lexical resource, although two systems also experimented with two different lexical resources. Also, several systems used Web queries or Google N-gram data to obtain counts for contextual fitness. We describe below the top five performing systems.

KU [20] is the highest ranking system for the *best normal* metric. It uses a statistical language model based on the Google Web 1T five-grams dataset to calculate the probabilities of all the synonyms. In the development phase, it compares two of the resources that we use in our work, namely WordNet and Roget's Thesaurus. In the test phase, it only uses the Roget resource.

UNT [9] is the best system for both the *best mode* and the *oot mode* mode. As lexical resources, it uses WordNet and Encarta, along with back-and-forth translations collected from commercial translation engines, and N-gram-based models calculated on the Google Web 1T corpus.

System	*best, normal*	*best, mode*	*oot, normal*	*oot, mode*
Our systems				
Unsup.indiv.	10.15%	16.05%	43.23%	55.28%
Unsup.comb.	12.81%	19.74%	43.74%	58.38%
Sup.comb.	**13.60**%	**21.30**%	*49.40%*	*64.70%*
SEMEVAL 2007 lexical substitution systems				
KU	12.90%	20.65%	46.15%	61.30%
UNT	12.77%	20.73%	49.19%	**66.26%**
MELB	12.68%	20.41%	N/A	N/A
HIT	11.35%	18.86%	33.88%	46.91%
IRST2	6.95%	20.33%	**68.96%**	58.54%

Table 9: *Comparison between our systems and the* SEMEVAL-2007 *systems*

IRST2 [8] ranks first for the *oot normal* metric. They use synonyms from WordNet and the Oxford American Writer Thesaurus, which are then ranked using either LSA or a model based on the Google Web 1T five-grams corpus.

HIT [21] uses WordNet to extract the synonyms. For the candidate fitness scoring, they construct Google queries to collect the counts. In order to collect the queries they only look at words close to the target word in context, with the intention of keeping noise at a low level.

MELB [15], which only participated in the *best* task, also relied on WordNet and Google queries. It is similar to the other systems described above, except that for the ranking of the candidates, they also take into account the length of the query and the distance between the target word and the synonym inside the lexical resource.

Table 9 shows a comparison between the results obtained with our system and those reported by the systems participating in the SEMEVAL-2007 task. Our system outperforms all the other systems for the *best normal* and *best mode* metrics, and ranks the second for the *oot normal* and *oot mode* metrics, demonstrating the usefulness of our combined approach.

7 Conclusions

In this paper, we experimented with the task of synonym expansion, and compared the benefits of combining multiple lexical resources, by using several contextual fitness models integrated into both unsupervised and supervised approaches.

The experiments provided us with several insights into the most useful resources and models for the task of synonym expansion. First, in terms of individual resource performance, WordNet and Encarta seem to lead to the best results.

Second, in terms of performance of the contextual fitness models, methods that measure substitutability in context seem to exceed the performance of methods that measure the similarity between a candidate synonym and the input context. Moreover, for the Web N-gram substitutability models, when used individually, the trigram models seem to perform as well as higher order N-gram model, which can be perhaps explained by their increased coverage as compared to the sparser four-grams or five-grams. The increased accuracy of the four-gram and five-gram models seems instead to be more useful, and thus more heavily weighted, when used in combination inside a supervised system.

Finally, a combination of several lexical resources provides the best results, exceeding significantly the performance obtained with one lexical resource at a time. This suggests that different lexical resources have different strengths in terms of representing word synonyms, and using these resources in tandem succeeds in combining their strengths into one improved synonym representation.

Overall, the results obtained through the combination of resources exceed the current state-of-the-art when selecting the best synonym for a given target word, and place second when selecting the top ten synonyms, which demonstrates the usefulness of combining lexical resources for the task of contextual synonym expansion.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation CAREER award #0747340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] S. Beale, B. Lavoie, M. McShane, S. Nirenburg, and T. Korelsky. Question answering using ontological semantics. In *Proceedings of the ACL Workshop on Text Meaning and Interpretation*, Barcelona, Spain, 2004.

[2] T. Brants and A. Franz. Web 1t 5-gram version 1. 2006.

[3] M. Carpuat and D. Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, 2007.

[4] Y. Chan, H. Ng, and D. Chiang. Word sense disambiguation improves statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, Prague, Czech Republic, 2007.

[5] O. Etzioni, K. Reiter, S. Soderland, and M. Sammer. Lexical translation with application to image search on the web. 2007.

[6] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence*, pages 1606–1611, Hyderabad, India, 2007.

[7] R. Girju, A. Badulescu, and D. Moldovan. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 80–87, Edmonton, Canada, 2003.

[8] C. Giuliano, A. Gliozzo, and C. Strapparava. Fbk-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 145–148, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[9] S. Hassan, A. Csomai, C. Banea, R. Sinha, and R. Mihalcea. Unt: Subfinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 410–413, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[10] S. B. Kim, H. Seo, and H. Rim. Information retrieval using word senses: root sense tagging approach. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, July 2004.

[11] R. Krovetz. Homonymy and polysemy in information retrieval. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)*, pages 72–79, 1997.

[12] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of the 14th International Conference on Machine Learning*, 1997.

[13] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 1997.

[14] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI, 1998.

[15] D. Martinez, S. N. Kim, and T. Baldwin. Melb-mkb: Lexical substitution system based on relatives in context. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 237–240, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[16] D. McCarthy and R. Navigli. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[17] G. Miller. Wordnet: A lexical database. *Communication of the ACM*, 38(11):39–41, 1995.

[18] C. Monz and B. Dorr. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005.

[19] C. Stokoe. Differentiating homonymy and polysemy in information retrieval. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, 2005.

[20] D. Yuret. Ku: Word sense disambiguation by substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 207–214, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[21] S. Zhao, L. Zhao, Y. Zhang, T. Liu, and S. Li. Hit: Web based scoring method for english lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 173–176, Prague, Czech Republic, June 2007. Association for Computational Linguistics.