Natural Language Engineering **1** (1): 1–27. Printed in the United Kingdom

© 2011 Cambridge University Press

Explorations in Lexical sample and All-words Lexical Substitution

RAVI SINHA
University of North Texas
Denton, TX
ravisinha@my.unt.edu

RADA MIHALCEA
University of North Texas
Denton, TX
rada@cs.unt.edu

(*Received May 2011; Revised March 2012; Accepted August 2012*)

Abstract

In this paper, we experiment with several techniques to solve the problem of lexical substitution, both in a *lexical sample* as well as an *all-words* setting, and compare the benefits of combining multiple lexical resources using both unsupervised and supervised approaches. Overall in the lexical sample setting, the results obtained through the combination of several resources exceed the current state-of-the-art when selecting the best substitute for a given target word, and place second when selecting the top ten substitutes, thus demonstrating the usefulness of the approach. Further, we put forth a novel exploration in all-words lexical substitution and set ground for further explorations of this more generalized setting.

1 Introduction

Word meanings are important for the semantic interpretation of texts. The understanding of the meaning of words is central to a large number of natural language processing applications, including information retrieval (Krovetz 1997; Kim et al.2004; Stokoe 2005), machine translation (Chan and Ng 2007; Carpuat and Wu 2007), knowledge acquisition (Girju et al.2003), subjectivity and sentiment analysis (Akkaya et al.2009), question answering (Beale et al.2004), and cross-language information retrieval (Monz 2005; Etzioni et al.2007).

In this paper, we experiment with lexical substitution, also referred to as 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 has been officially introduced during Semeval-2007 (McCarthy and Navigli 2007), 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.

Aside from performing a set of experiments in a lexical sample setting as proposed by (McCarthy and Navigli 2007), where a target word in context is replaced by its substitutes, and where we can compare our results with previous work, we also extend our setting to that of an all-words task, where we try to find substitutes for all open-class words in a given text.

In the experiments reported in this paper, we analyze the relative usefulness of different lexical resources – used individually or in tandem – for the purpose of lexical substitution. We experiment with several resources to determine the ones that provide the best substitutes for a given word in context. We then compare several methods for determining the fitness in context for the substitutes.

After reaching some conclusions based on the lexical sample experiments, we proceed to generate data for the all-words task, and repeat the experiments with the combinations of resources and fitness measures found to work best in the lexical sample setting.

The paper is organized as follows: we first introduce the task of lexical substitution. We then describe several lexical resources for collecting sets of substitutes (or synonyms), followed by a discussion of the methods employed to determine the goodness of fit of a synonym in a context. Next we present our experiments and evaluations in two subsections, each dedicated to the lexical sample approach and the all-words approach respectively. Further we present an overview of work related to the task and then conclude with discussions and perspectives for future work.

2 Lexical substitution

Lexical substitution, also known as contextual synonym expansion (McCarthy and Navigli 2007), involves replacing a certain word in a given context with another, suitable word, such that the overall meaning of the word and the sentence are unchanged. As an example, see the four sentences in 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 subtask 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 these synonyms can be 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*.

It is intuitive to think that 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.

Again, it is intuitive to think that the better the measure of contextual fitness, the 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.

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

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

3 Lexical resources for candidate synonym selection

For the purpose of the first step of our algorithm, namely the 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.

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.

WordNet

WordNet, as described in (Miller 1995), 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.

TransGraph

TransGraph, introduced by (Etzioni et al.2007), 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.

Lin's distributional similarity

(Lin 1998) 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, up to a maximum of twenty if available.

[1] http://www.thesaurus.com

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.

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*. Overall, 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.

The diversity of these lexical resources can be quantified by measuring their overlap. For each pair of two resources, we calculate the *percentage overlap*, defined as the number of candidate synonyms provided by a resource R_1 that are also provided by resource R_2, divided by the total number of words provided by resource R_1. Table 3 shows the percentage overlap between the five resources used in our experiments, where, using the notation above, the rows correspond to R_1 and the columns correspond to R_2.[3] Interestingly, the highest overlap between any two resources is slightly over 50%, which demonstrates the diversity of the resources, and suggests that a combination of resources could potentially improve over the use of one resource at a time.

Table 3. *Overlap among the different resources (measured on the test data of 171 target words described in Section 5.1)*

	RG	WN	TG	LN	EN
RG	100.00%	20.28%	47.01%	17.97%	36.23%
WN	37.39%	100.00%	51.34%	19.97%	40.04%
TG	25.92%	15.35%	100.00%	12.77%	22.85%
LN	42.66%	25.70%	54.99%	100.00%	41.61%
EN	46.77%	28.03%	53.51%	22.64%	100.00%

4 Methods for contextual fitness

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

[2] http://encarta.msn.com

[3] Note that for the Lin distributional similarity, we use the implementation provided by its author Dekang Lin. This implementation does not cover adverbs, and thus only a subset of the target words can be covered with this resource.

One aspect that needs to be addressed when measuring the fitness in context is the issue of morphological variations. In particular, for methods that look at substitutability in context using N-gram-based language models (descriptions ahead), 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,[4] consisting of morphological inflections for 110,000 English words. The dictionary is organized by part-of-speech, and therefore it provides different sets of inflections for words with different parts-of-speech. For instance, the inflection dictionary will list *watered, watering* and *waters* as inflections for the verb *water*, but it will only list *waters* as inflection for the noun *water.*

We describe below the three fitness algorithms used in our experiments. For the lexical sample experiments, we run exhaustive experiments using all these algorithms. For the all-words experiments, we employ the ones found to work best in the lexical sample subtask.

The first two methods, latent semantic analysis and explicit semantic analysis, are measures of similarity. We use these measures to determine the fitness of a candidate synonym in context as follows. First, for the given context, we determine the corresponding vectors for the words in context, as provided by one of the methods. Next, a vector representation for the entire context is obtained by doing a component-wise sum of the vectors for the individual words that constitute the context. The similarity between a candidate synonym and the context is then calculated as the cosine similarity between the vectors corresponding to the candidate synonym and to the context. This is similar to earlier work by (Gabrilovich and Markovitch 2007; Landauer and Dumais 1997; Mitchell and Lapata 2008). Thus, given two vector representations A and B, obtained by using one of the methods: $A = [a_1, a_2, a_3, ..., a_n]$ and $B = [b_1, b_2, b_3, ..., b_n]$, the similarity is calculated using the cosine similarity: sim(A, B) = $\frac{\sum a_i b_i}{\sqrt{\sum a_i^2}\sqrt{\sum b_i^2}}$. The candidates are then ordered in reverse order of their similarity with the context. Note that the target word itself is part of this context, and thus the similarity between the target word and a candidate synonym also contributes toward this similarity score. While contexts of different lengths can be considered, in our experiments we use a context consisting of the sentence where the target word occurs.

The third method, based on Google N-grams, consists of a language model, and is used to determine the likelihood of a candidate synonym given its surrounding context. As explained below, in this method contexts of various sizes are considered.

Latent semantic analysis

One corpus-based measure of semantic similarity is latent semantic analysis (LSA) proposed by (Landauer and Dumais 1997). LSA builds semantic vector spaces,

[4] A large automatically generated inflection database (AGID) available from http://wordlist.sourceforge.net/

where the meaning of a word can be represented by a vector, which in turn depicts the presence of the word in a large corpus, obtained using statistical computations. Despite being criticized for not accounting for the various senses a polysemous word might have (thus averaging all meanings for a term), and also for the fact that it depends solely on the text representation and does not derive any information from outside knowledge or supervision, LSA is still based on powerful mathematical analysis to produce meaning representations by inferring deep relations in text, and has been widely and successfully used for various language processing similarity tasks. As documented in (Landauer and Dumais 1997), LSA scores have been shown to match those of humans on standard vocabulary and subject matter tests, as well as category judgment tests, among several other applications.

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.[5]

Explicit semantic analysis

Explicit semantic analysis (ESA) was proposed as an improvement over LSA, amounting to substantial improvements in correlation of computed relatedness scores with human judgments, as described in (Gabrilovich and Markovitch 2007). The proponents of ESA also contend that the model is easier to explain to human users since the dimensions in the model are natural concepts. Analogous to LSA, ESA also does not make use of outside world knowledge or human supervision, however it can be thought of as naturally using common sense since each dimension is a natural concept in Wikipedia (which has been chosen solely owing to the fact that a larger knowledge repository does not exist). In ESA, 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.[6] We use the ESA vectors of a target word and the context and compute the cosine similarity between them, as described above.

Google N-gram models

The Google Web 1T corpus (Brants and Franz 2006) is a collection of English N-grams, ranging from one to five N-grams, and their respective frequency counts observed on the Web. 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

[5] We use the Infomap implementation of LSA http://infomap-nlp.sourceforge.net/

[6] We use an in-house implementation of ESA kindly provided by Samer Hassan.

is important, and thus we use all the possible inflections for all the potential candidates, obtained by querying the inflection dictionary described earlier in this section.

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.[7] 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 models. In all these models, the frequencies corresponding to the different N-grams are summed up. We decided to use summations of frequencies rather than products of probabilities based on results from earlier work (Bergsma et al.2009), and also based on a small experiment on the development dataset where a frequency based approach was observed to lead to better results as compared to a probabilistic approach.

We describe below the five models used in our experiments, which use three-, four-, and five-grams independently, as well as combinations of all N-grams.

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.

For the sake of completeness, we also experimented with other combinations,

[7] To query Google N-grams, we use an in-house *B-tree search* implementation, kindly made available by Hakan Ceylan.

namely *34gramSum, 345gramSum, 3gramAny*, etc., but we are only reporting the results for the models performing best on a small development dataset provided with the SEMEVAL task (McCarthy and Navigli 2007).

5 Experiments and evaluations

5.1 Lexical sample

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. There are ten instances for each target word, for a total of 30 target words in the development dataset, and 171 target words in the test dataset. The words are split among the four parts-of-speech; for instance, the test data contains 50 nouns, 44 verbs, 47 adjectives, and 30 adverbs.

We use the same evaluation metrics as used for the lexical substitution task at SEMEVAL-2007. Although another set of arguably more meaningful metrics have been introduced in (Jabbari et al.2010), we decided to stick with the original metrics. Our justification for proceeding with these metrics is twofold. First, using the initial metrics gives us a platform for comparing our numerical results with the results published before. Second, all the results reported so far on this task correspond to systems that have been developed on these metrics, and thus we believe it would be unfair to re-score the older systems based on the newer metrics.

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 dataset 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 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 (McCarthy and Navigli 2007), are summarized below. Let us assume that $\mathtt{H}$ is the set of annotators, namely $\{\mathtt{h}_1, \mathtt{h}_2, \mathtt{h}_3, ...\}$, and $\mathtt{T}$, $\{\mathtt{t}_1, \mathtt{t}_2, \mathtt{t}_3, ...\}$ is the set of test items for which the humans provide at least two responses. For each $\mathtt{t_i}$ we calculate $\mathtt{m_i}$, which is the most frequent response for that item, if available. We also collect all $\mathtt{r}_\mathtt{i}^\mathtt{j}$, which is the set of responses for the item $\mathtt{t_i}$ from the annotator $\mathtt{h_j}$.

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 $\mathtt{TM}$, such that $\mathtt{TM} \subseteq \mathtt{T}$. Also, let $\mathtt{A} \subseteq \mathtt{T}$ 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 $\mathtt{AM}$, such that $\mathtt{AM} \subseteq \mathtt{TM}$. Let $\mathtt{a_i} \in \mathtt{A}$ be the set of system's responses for the item $\mathtt{t_i}$.

Thus, for all test items $\mathtt{t_i}$, 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 $\mathtt{t_i}$, we calculate $\mathtt{R_i}$, which is $\sum \mathtt{r_i^j}$, and the individual unique item in $\mathtt{R_i}$, say `res`, will have a frequency associated with it, namely $\mathtt{freq_{res}}$.

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

Best measures:

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

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

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

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

Out of ten (oot) measures:

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

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

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

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

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

5.1.1 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 4 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.

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

	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 5 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 5 and in Table 10 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.

Table 5. *F-measure for the four scoring schemes for individual lexical resources (test data)*

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%

Additionally, we also run a separate evaluation for each part-of-speech, shown in Table 6. Not surprisingly, verbs appear to be the most difficult part-of-speech, which is inline with previous findings on word sense disambiguation (Mihalcea and Edmonds 2004; Agirre and Edmonds 2006). The best results are obtained for adverbs, which is also justified by the fact that words with this part-of-speech tend to have a smaller polysemy on average.

Table 6. *F-measures for the four scoring schemes for individual lexical resources, separated by part-of-speech (test data)*

Metric	Resource	Model	Noun	Verb	Adj	Adv
best, normal	WN	3gramSum	9·04%	7·65%	7·99%	19·24%
best, mode	WN	345gramAny	13·24%	11·95%	13·89%	28·79%
oot, normal	EN	3gramSum	41·69%	37·19%	45·04%	52·03%
oot, mode	EN	3gramSum	51·06%	46·09%	62·04%	64·63%

To have a better understanding of these results, we perform two analyses on the individual lexical resources used in the experiments. First, we try to quantify the "usefulness" of each of these resources by counting the number of synonyms that can be extracted for a given set of target words. Considering the test set of 171 target words, Table 7 shows the number of target words that have exactly one synonym, between two and ten synonyms, and more than ten synonyms.[8] As illustrated in this table, with one exception (two words that have only one synonym provided by the Lin distributional similarity resource), all resources provide more than one synonym. For the large majority of words, ten synonyms or more are provided by any given resource, although there are also a number of words for which the number

[8] As mentioned before, Lin distributional similarity does not cover adverbs, which is the reason why the numbers in the column corresponding to the Lin resource do not add up to 171.

of synonyms is less than ten. This suggests that the contextual filters are in fact needed for generating the *best* answer for all five lexical resources, and they are also needed for a large majority of the target words to generate the *oot* answer (all the words with more than ten synonyms need a contextual filter).

Table 7. *Number of synonyms identified in different lexical resources (test data)*

	RG	WN	TG	LN	EN
1	0	0	0	2	0
2-10	29	75	17	0	33
>10	142	96	154	139	138

The second analysis that we performed was concerned with a measurement of the upper bound that can be achieved by using these five lexical resources. Assuming an oracle that can pick all the gold standard answers from among the candidate synonyms provided by a lexical resource, we determine the highest result that can be obtained in this way. Table 8 shows these upper bound figures calculated for the test data of 171 words. Note that we did not duplicate the mode gold standard answers in order to obtain ten answers for the oot mode. Instead, we chose to include in oot mode only the one mode from the gold standard, if present in the lexical resource, and thus the upper bound obtained with best and oot mode are identical. These upper bound figures show once again that there are large differences between the synonyms provided by these lexical resources, and thus they are likely to complement each other – an observation that is in fact supported by our findings in the following section, where we use a combination of resources to identify candidate synonyms. Interestingly, Encarta has the highest upper bound, which suggests that resources that are more encyclopedic in nature are more useful for this task as compared to other resources that emphasize more the lexicographic and dictionary aspects such as WordNet or Roget. Finally, these upper bound figures are also very useful to place results in perspective, by providing the means to compare the results of our system with the highest achievable results given the individual lexical resources. For instance, a comparison between the results obtained with our system and this upper bound for the best measure shows that this subtask is more difficult than the oot subtask, where the gap between our system and the upper bound is smaller. This is not surprising, since in the best subtask, a system has to find one best synonym, which is harder than finding ten words that are likely to be synonyms, as required in the oot subtask.

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

Table 8. *Upper bound F-measure scores (test data)*

Resource	best, normal	best, mode	oot, normal	oot, mode
RG	25.21%	39.60%	43.09%	39.60%
WN	28.01%	43.10%	38.21%	43.10%
TG	35.29%	59.00%	53.31%	59.00%
LN	20.67%	33.29%	33.66%	33.29%
EN	36.63%	65.59%	59.88%	65.59%

1. Encarta and WordNet. All the candidate synonyms returned by both Encarta and WordNet for a target word.
2. 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.
3. Any Two. All the candidate synonyms that are included in at least two lexical resources.
4. 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 9. Overall, the combined resources tend to perform better than the individual resources.

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 10 shows the results obtained on the test set for the selected combinations of lexical resources. The results separated by part-of-speech are shown in Table 11, which suggest the same difficulty pattern as observed on the individual lexical resources, with the best results being obtained for adverbs, and the lowest results being obtained for verbs.

5.1.3 Experiment 3: Supervised combination of knowledge sources

As a final set of experiments for the lexical sample setting, 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. It is important to note that this is a global classifier that learns how to combine the invidual knowledge-sources, rather than a per-word classifier.

We build a feature vector for each candidate synonym, and for each instance in the training and the test data. The features include an identifier 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

Table 9. *F-measures for the four scoring schemes for combined lexical resources (development data)*

	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%

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 dataset. We use an under-sampling technique described in (Kubat and Matwin 1997), and randomly eliminate

Table 10. *F-measures for the four scoring schemes for combined lexical resources (test data)*

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 11. *F-measures for the four scoring schemes for combined lexical resources, separated by part-of-speech (test data)*

Metric	Resource	Model	Noun	Verb	Adj	Adv
best, normal	EN and WN	3gramSum	11·68%	9·45%	12·02%	20·29%
best, mode	AnyThree	345gramAny	17·59%	11·95%	17·29%	28·79%
oot, normal	EN or WN	3gramSum	41·37%	36·44%	44·44%	57·18%
oot, mode	EN or WN	3gramSum	53·11%	47·39%	62·99%	74·28%

negative examples until we reach a balance of almost two negative examples for each positive example. The final training dataset contains a total of 700 positive examples and 1,500 negative examples. The under-sampling is applied only to the training set.

The results obtained when applying the supervised classifier on the test data are shown in Table 12. 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[9] package.

Table 12. *F-measure for a supervised combination of lexical resources (test data). NN=nearest neighbor; LR=logistic regression; DL=decision lists; SVM=support vector machines*

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%

In Table 13, we also report the results obtained for individual parts-of-speech, this time only for the decision list classifier, which is the classifier that led to the

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

best results. The same trend as before is observed, with verbs having the lowest results, and adverbs the highest.

In fact, in a more in-depth analysis of the output of this supervised system, we calculated the score obtained by each individual target word, which allowed us to determine the "difficult" and "easy" words for this task. Among the most difficult words for the best measure, we found words like *clear (adj), shed (verb), clear (verb)*, whereas other words such as *external (adj), often (adv), around (adv)* were among the easiest. For the oot measure, the most difficult words were *return (verb), pass (verb), jam (noun)*, while the easiest were *therefore (adv), worldwide (adj), entirely (adv).* These appear to be correlated to the difficulty associated with different parts-of-speech, as observed before, and also to some extent with the polysemy of the various words.

Table 13. *F-measure for a supervised combination of lexical resources using decision lists, separated by part-of-speech (test data)*

Metric	Noun	Verb	Adj	Adv
best, normal	11·7%	9·5%	14·9%	21·0%
best, mode	15·7%	14·6%	25·3%	33·0%
oot, normal	43·8%	43·3%	52·4%	62·7%
oot, mode	57·9%	55·4%	70·1%	79·8%

To gain further insights, we also look at the information gain weight as assigned by Weka to each feature in the dataset, in order to determine the role played by each feature. 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.

Table 14 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.

5.1.4 Comparison with previous work

There are several systems for synonym expansion that participated in the Semeval-2007 lexical substitution task (McCarthy and Navigli 2007). In this section, we present a comparison between the results obtained with our approach and those reported by the teams participating in the Semeval task.

Most of the Semeval systems used only one lexical resource, although two systems also experimented with two different lexical resources. Also, several systems

Table 14. *Information gain feature weight*

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	0·0110
345gramSum	0·0109
3gramSum	0·0106
345gramAny	0·0104

used Web queries or Google N-gram data to obtain counts for contextual fitness. We describe below the top five performing systems.

KU (Yuret 2007) is the highest ranking system for the *best normal* metric. It uses a statistical language model based on the Google Web 1T 5-grams dataset to calculate the probabilities of all the synonyms. In the development phase, it compares two of the resources that we also use in our work, namely WordNet and Roget's Thesaurus. In the test phase, it only uses the Roget resource.

UNT (Hassan et al.2007) 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.

IRST2 (Giuliano et al.2007) 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 N-grams corpus, just like most other high-performing systems participating in the task.

HIT (Zhao et al.2007) 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 (Martinez et al.2007), 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 15 shows a comparison between the results obtained with our system and the ones reported by the systems participating in the SEMEVAL-2007 task. Results obtained by SEMEVAL-2007 systems that are smaller than our supervised system

Table 15. *Comparison between our systems and the* Semeval-2007 *systems*

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%*

by a statistically significant margin are denoted with a star (statistical significance was computed by using a paired t-test, p=0.01). Our system outperforms all the other systems for the *best normal* and *best mode* metrics, even if not always by a significant margin, and ranks the second for the *oot normal* and *oot mode* metrics.

In these comparative results, it is interesting to note that for the *oot normal* measure, there is one Semeval-2007 system that exceeds our results by a large margin (IRST2). A closer look at the results of all the participating systems in the Semeval-2007 lexical substitution task (McCarthy and Navigli 2007) reveals the fact that this system is an outlier for this measure, as it exceeds all the other systems by 20% or more. Since this system was also heavily based on the use of the Google N-grams as a contextual filter method, we believe the difference may be due to the use of a different lexical resource, namely the Oxford dictionary, which is not readily available and it was not used by any of the other participants.

5.2 All-words

In our endeavor toward extending our work to an all-words lexical substitution setting, our first task was to generate data for evaluating the algorithms with respect to human annotations. This was needed because the work done so far on lexical substitution has mostly been focused on a single target word in context and all the data available for this task is thus suitable only for a lexical sample setting.

Unlike lexical sample, in an all-words setting we expect a more varied set of words, following the distribution that typically occurs in language. We chose to develop our evaluation dataset starting with the texts used for trial and test in the all-words word sense disambiguation task at Semeval-2007 (Pradhan et al.2007).

The original source data consists of around 550 head words, spread unevenly over multiple sentences. The dataset contains 164 nouns, 377 verbs, 8 adjectives and 1 adverb. In order to make the data more suitable for annotation as well for use with the scorer we used for the lexical sample task, we transform the data into a format similar to the one used for the lexical sample. While the original data consists of multiple head words in a given sentence, we convert it into a format where we repeat every sentence as many times as there are head words in that sentence, but each time only one of the words is marked as a head word.

For the annotations, we use three human annotators – one of them a native speaker of English and two of them with a high level of English proficiency. The pairwise inter-annotator agreement, calculated based on the technique discussed in (McCarthy and Navigli 2009), was determined as 15.53%. This is a relatively lower figure compared to the agreement of 27.75% calculated for the lexical sample (McCarthy and Navigli 2009), which may be due to the increased difficulty and higher word diversity in the all-words subtask, as well as the fact that the lexical sample annotations used only English native speakers.

As done in the original lexical sample annotation task, for each of the 550 target words, the annotators were asked to provide comma separated synonyms: single word alternatives as much as possible, although phrases were also acceptable. The annotators were free to use any dictionaries, lexicons or other sources deemed worthwhile. They were also asked to manually verify the correctness of the lemma and part-of-speech associated with each target word.

Some of the guidelines that the annotators followed are listed below:

1. Try to find the best possible one word that can substitute the target word, while preserving the meaning of the sentence.
2. If a very good multi-word phrase can be found that can bring out the meaning of the sentence in a better way than a single-word substitute could, then add the multi-word phrase as a possible substitute.
3. In general, you should try to use multi-words only if a single-word substitute is not at all possible or does not clearly preserve the meaning of the sentence.
4. Where there is a phrasal verb (a multi-word verb) e.g. *take up*, and synonyms for a part of that multi-word verb are requested, e.g. *take*, find words that substitute for that single-word, *take* with the meaning of the multi-word verb, *take up*, but do not have to necessarily fit into the sentence without changing its structure
5. It is acceptable to provide only one synonym because others could not be found, or because other synonyms do not really bring out the context in the right *meaning*. In other words, prefer quality over quantity. Even though it is expected that you provide as many synonyms as you possibly can, if there are only one or two synonyms that fit well in the context and others do not, then only provide those few synonyms

There were some mismatches in the lemmas that the annotators provided. The mismatches happened because of certain cases where the lemmatization caused a change in the part-of-speech of the word. For example, *detailed* – which ideally should be marked as *(detailed, adjective)* – was changed by some annotators to *detail*, which is either a verb or a noun. This kind of confusion occurs because of the fact that lemmatization sometimes changes the part-of-speech and therefore the meaning of the word.

Overall, there were an average of 6.4 annotations per word, and there were only 11 instances where the total number of annotations by the three annotators was less than 2.

After obtaining the annotations for the new dataset, we create a gold-standard by compiling all the annotations from the three annotators into one document and recording their corresponding frequency, and ranking the synonyms for each instance according to the descending order of their frequencies in order to match the format of the gold standard of the lexical substitution task.

Next, we use the dataset to run several evaluations. As before, we use both an unsupervised and a supervised setting. However, instead of running several comparative experiments on a development dataset, we use the findings from the lexical sample subtask and run experiments only with those settings that were found to work best on that subtask. For the unsupervised setting, we report results obtained with several individual lexical resources, as well as results obtained with combinations of resources. For the supervised setting, we use a 2 : 1 training/test split of the dataset.

Similar to the lexical sample evaluations, the results are reported using the F-measure, by comparing the synonyms suggested by the different algorithms against the gold standard. As before, the classifier makes binary decisions, and the training data is balanced using under-sampling (see Experiment 3 in the previous section).

Table 16 shows the results obtained when using individual resources and Table 18 shows the results obtained when using combinations of the lexical resources. The results of the supervised experiments are shown in Table 20. Results separated by part-of-speech are also shown for each of these experiments in Table 17 (individual lexical resources), Table 19 (combined lexical resources), and Table 21 (supervised combination).

Table 16. *F-measure for the four scoring schemes for individual lexical resources (all-words data)*

Metric	Resource	Model	F-Measure
best, normal	WN	3gramSum	7·13%
best, mode	WN	345gramAny	13·18%
oot, normal	EN	3gramSum	25·56%
oot, mode	EN	3gramSum	32·78%

Not surprisingly, machine learning seems to provide the best results, with an F-measure as high as 73.7% obtained by using decision trees. Interestingly, the results of the supervised setting exceed the results obtained with a similar setting for the lexical sample, although the unsupervised results are significantly below those obtained in the lexical sample. This may be explained by the fact that we have a larger and more varied training set in the all-words setting, as compared

Table 17. *F-measures for the four scoring schemes for individual lexical resources, separated by part-of-speech (all-words data)*

Metric	Resource	Model	Noun	Verb	Adj	Adv
best, normal	WN	3gramSum	6·58%	7·23%	3·73%	44·42%
best, mode	WN	345gramAny	11·24%	13·99%	0·00%	66·67%
oot, normal	EN	3gramSum	31·66%	23·18%	17·58%	0·00%
oot, mode	EN	3gramSum	38·23%	30·63%	25·00%	0·00%

Table 18. *F-measures for the four scoring schemes for combined lexical resources (all-words data)*

Metric	Resource	Model	F-Measure
best, normal	EN and WN	3gramSum	7·44%
best, mode	AnyThree	345gramAny	17·44%
oot, normal	EN or WN	3gramSum	30·02%
oot, mode	EN or WN	3gramSum	38·52%

to the development set used for the lexical sample experiments. Note that we are not using the prediction for one word to facilitate the prediction for other words in context, so we do not think the errors are getting multiplied or propagated in the all-words setting.

The results separated by part-of-speech reveal an unexpected finding: unlike the lexical sample data, where verbs were clearly the most difficult part-of-speech, in this dataset nouns and verbs seem to have a similar level of difficulty, which suggests that in running text, verbs are not necessarily more difficult to handle.[10] Unfortunately no conclusive results could be obtained on this dataset for adjectives and adverbs, given the small number of words in these categories (8 adjectives, 1 adverb).[11]

For a deeper understanding of the results, as done in the lexical sample setting, we perform several additional analyses. First, we analyse the individual resources, to determine the number of candidate synonyms that are provided by each of the two resources used in these experiments. Table 22 shows the number of target words that have exactly one synonym, between two and ten synonyms, and more than ten synonyms. In this case, as it would be expected in an all-words task, there are more words that have only one meaning and one synonym. This effect is even

[10] It is important to keep in mind that these findings are based on relatively small datasets: the lexical sample contains 500 nouns and 440 verbs, and the all-words dataset contains 164 nouns and 377 verbs.

[11] Because there is always one adverb, one may think that the F-measure should be either 0 or 1, but this is not necessarily the case, as the scorer uses some weighting in how it takes into account individual items, and thus partial scores are also possible.

Table 19. *F-measures for the four scoring schemes for combined lexical resources, separated by part-of-speech (all-words data)*

Metric	Resource	Model	Noun	Verb	Adj	Adv
best, normal	EN and WN	3gramSum	7·43	7·58	0·00%	0·00%
best, mode	AnyThree	345gramAny	15·74%	18·79%	0·00%	0·00%
oot, normal	EN or WN	3gramSum	30·99%	29·78%	15·92%	44·42%
oot, mode	EN or WN	3gramSum	37·08%	39·22%	25·00%	66·67%

Table 20. *F-measure for a supervised combination of lexical resources (all–words data). NN=nearest neighbor; LR=logistic regression; DL=decision lists; SVM=support vector machines*

Metric	NN	LR	DL	SVM
best, normal	21·8%	21·2%	**21·5**%	20·2%
best, mode	29·8%	36·9%	**29·9**%	26·2%
oot, normal	50·3%	46·3%	**63·9**%	46·0%
oot, mode	61·1%	57·5%	**73·7**%	56·6%

stronger in WordNet, where as many as 4% of the words have only one synonym. These are words for which the contextual filter is not necessary neither for the *best* evaluation nor for the *oot*. There is also a large number of words that have more than two synonyms but less than ten, and these represent words for which the contextual filter is not needed for the *oot* evaluation. As many as 57% of the words (WordNet) and 37% of the words (Encarta) fall under this category. Finally, the dataset also includes the more difficult words, with more than ten synonyms, for which a contextual filter is necessary for both evaluations (39% of the words in WordNet and 63% of the words in Encarta).

Second, we also calculate the upper bound that can be achieved on this dataset with the two resources used in the evaluations. Table 23 shows the upper bound figures for the all-words dataset of 550 words, calculated as the highest results that can be obtained by selecting all the gold standard answers provided by a given lexical resource. Consistent with our observation on the lexical sample dataset, Encarta leads to a higher overall upper bound. Overall, on the all-words dataset, the gap between the results obtained using our system with individual lexical resources (Table 16) and these upper bounds is somehow smaller as compared to the gap observed on the lexical sample dataset between the results obtained with the individual lexical resources (Table 5) and the corresponding upper bounds (Table 8). We believe this is due to the different aspects covered by the two datasets: the lexical sample set is mainly addressing difficult, ambiguous words, with a constant number of examples for each target word, whereas the all-words dataset covers running text, which in-

Table 21. *F-measure for a supervised combination of lexical resources using decision lists (all-words data), separated by part-of-speech*

Metric	Noun	Verb	Adj	Adv
best, normal	28·50%	21·10%	12·50%	0·00%
best, mode	33·33%	32·70%	0·00%	0·00%
oot, normal	78·50%	65·50%	25·50%	0·00%
oot, mode	85·70%	78·80%	0·00%	0·00%

Table 22. *Number of synonyms identified in different lexical resources (test data, all-words)*

	WN	EN
1	23	0
2-10	315	203
>10	212	347

cludes a larger spectrum of words ranging from easier monosemous words, all the way to highly polysemous words.

The evaluation closest to ours is the one reported in (Inkpen 2007) on work done on large scale lexical substitution. However, we cannot directly compare our results, because our settings and goals are different. In that work, although the reported raw numbers are higher, they are not compared against human judgment. Instead, in their evaluation, they attempt to match the original target word in the sentences by choosing from a large set of potential synonyms including the original target word. Further, they only evaluate nouns and adjectives, while in our experiments we also consider verbs and adverbs.

6 Related work

In this section we present an overview of recent work in the area of lexical substitution, and also attempt to synthesize the current state-of-the-art and the current research projects concerned with this task.

Lexical substitution has been an interesting topic for research for a long time, but has most recently received a lot of attention under the Semeval monolingual (McCarthy and Navigli 2007) and cross-lingual (Mihalcea et al.2010) lexical substitution tasks. The work that is closest to ours consists of the lexical substitution systems participating in the Semeval monolingual task. These systems have been briefly described in Section 5.1.4, along with a direct comparison with our own system.

In addition to the Semeval task, there are also a number of projects concerned with applications of lexical substitution. For instance, (Chang and Clark 2010)

Table 23. *Upper bound F-measure scores (test data, all-words)*

Resource	best, normal	best, mode	oot, normal	oot, mode
WN	16.62%	32.34%	27.94%	32.29%
EN	19.24%	39.64%	39.60%	39.90%

investigate the use of synonym expansion in context for information assurance and security. The accuracy of lexical substitution is directly reflected in the success of information hiding because, ideally, changing the text to hide information should not result in ungrammatical or unnatural text, rendering the camouflaged text inconspicuous. (Chang and Clark 2010) use the WEB1T Google N-gram corpus to check how applicable a synonym is in context, starting with synonym candidates as suggested by WordNet. The fitness of each candidate synonym is calculated by summing up the logarithms of the counts of all the N-grams containing that synonym, with N ranging between two and five.

(Yatskar et al.2010) takes lexical substitution in a different direction, highlighting another potential use of the task – that of *lexical simplification*. Rather than focusing on syntactic transformations as done in the past, the authors explore data-driven methods to learn lexical simplifications, e.g., *collaborate → work together*. The authors focus on the edit history information in the Simple English Wikipedia[12], and use both a probabilistic distribution model as well as meta-data to learn potential simplifications. Also related is the lexical simplification work targetting domain-specific data, such as medical texts (Elhadad and Sutaria 2007; Deléger and Zweigenbaum 2009).

(Biemann 2010) examines how features arising from co-occurrence clusters can be used for lexical substitution. Starting in an unsupervised fashion the author first clusters the words from the local neighborhood of a target word, and then uses the clusters as features in a supervised word sense disambiguation setting. Part of the motivation behind this work is *word sense induction*, i.e., the automatic identification of word senses using clustering. Several word graphs around each target word are constructed based on several parameters and using sentence-based co-occurrence statistics from a large corpus. The edges of the graphs have weights corresponding to the log-likelihood significance of the co-occurrence between pairs of words. The clustering is performed using *Chinese Whispers*, as described in (Biemann 2006). After having obtained the features, a classification algorithm is used, which suggests and ranks substitutes.

(Dagan et al.2006) present work done on lexical substitution from another point of view, namely *sense matching*. Given two synonyms obtained from a sense lexicon, such as WordNet, or a database constructed using statistical word similarities, the authors focus on verifying whether the meanings of those synonyms indeed match

[12] http://simple.wikipedia.org/wiki/Main_Page

in a given context. The traditional approach of doing this is using word sense disambiguation, but the authors present a novel approach without WSD, which they claim might be an unnecessary (and probably harder) problem, and suggest that a binary classification is probably a more feasible approach. A database of pairs of synonyms is created using WordNet, and a classification instance is created with the synonym pair and the context of the target word. A positive instance is where the sense of the target word matches one of the senses of the source word in that context, and the instance is negative otherwise. The gold standard is derived automatically from the data. Several experiments (both supervised and unsupervised) are performed, and the conclusion is that the results obtained when an intermediate WSD step is used and those obtained without the WSD step are almost identical.

(Preiss et al.2009) present three approaches for lexical substitution, using Hidden Markov Models, grammatical relations, and n-gram language models, which are applied on a candidate list built from WordNet and Encarta. They also demonstrate the language independence of the approach by applying their methods on a lexical substitution dataset built for Czech.

A closely related line of work is the one concerned with lexical choice or synonym expansion. For instance, (Davidov and Rappoport 2009) propose a method to extend a given concept by using translations into intermediate languages and disambiguating translations using Web counts. First, a concept is represented as a set of individual words (e.g., all the synonyms in a WordNet synset), and then translations are done using bilingual dictionaries. The disambiguation using Web counts follows the conjecture that words belonging to the same concepts tend to appear together. The disambiguated translations are then translated back into the original language, scored, and the source concepts are therefore extended by adding those translations that had high scores. The method can be useful for extending the overall set of potential candidates for lexical substitution.

(Wang and Hirst 2010) approach the related task of near-synonym lexical choice with a word-space model built on co-occurrences. The evaluation is run using a fill-in-the-blank task. Although the problem they address is somehow more involved than lexical substitution because the emphasis of the task is to distinguish synonyms on a very fine-grained level, we believe that this is a methodology that can be used on top of our systems to fine-tune the reported substitutes and to also determine the *best* outcome out of the several outcomes in the out-of-ten output. The authors argue that the seemingly intuitive problem of "choosing the right word for the right context" is far from trivial because every dimension of variation amongst the synonyms introduces differences in style, connotation and truth conditions into the discourse. The non-triviality is valid for humans as well, as shown in an earlier fill-in-the-blank evaluation where two humans achieved an agreement of only 80% (Inkpen 2007).

(Yu et al.2010) consider the near-synonym substitution task as a classification task and construct classifiers for each near-synonym set. Further, to improve the classifiers, the authors propose *discriminative training*, which distinguishes between positive and negative features for each synonym set. Focusing on how suitable a potential candidate is in a given surrounding context, there are a lot of statistical measures that have been proposed. (Edmonds 1997) approaches this issue by summing the *t-scores* of the co-occurrences of a candidate word with the context words, while (Bangalore and Rambow 2000) use conditional probability. (Inkpen 2007) uses pointwise mutual information (PMI) to measure the associations between the candidate and context words, thereby improving upon the method used in (Edmonds 1997).

Also related to our work is the broader task of paraphrase generation and detection, which is often concerned with larger spans of text such as sentences or even entire documents. Methods proposed in this direction are either targetting the identification of closely related texts by aggregating methods of word relatedness into metrics that work at text level (Mihalcea et al.2006; Islam and Inkpen 2009; Hassan and Mihalcea 2011), or they address the acquisition of paraphrase corpora (Dolan et al.2004; Shinyama et al.2002) or the generation of paraphrased text (Barzilay and Lee 2003; Quirk et al.2004). The acquisition of synonym words from such large paraphrase corpora is a potentially promising avenue to explore in future work for our candidate synonym extraction step, as described in Section 3.

It is interesting to note that although lexical substitution can be regarded as a special case of paraphrase generation (Androutsopoulos and Malakasiotis 2010), the task of lexical substitution has additional constraints: the paraphrasing is limited to a pair of words, and the contexts must be exactly the same. This further means that the two expressions must be usable interchangeably in grammatically correct sentences, and thus a system for paraphrase generation may not be directly applicable to the task of lexical substitution.

7 Conclusions

In this paper, we experimented with the task of lexical substitution, and compared the benefits of combining multiple lexical resources, by using several contextual fitness models integrated into both unsupervised and supervised approaches. Further, we extended the experiments to the broader realm of all-words lexical substitution, by introducing a new dataset and running comparative evaluations for all the words in running text.

The experiments provided us with several insights into the most useful resources and models for the task of lexical substitution. 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 models, 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, we believe the main contribution of this paper stands in providing an in-depth analysis of the role played by different lexical resources and different contextual fitness algorithms for the task of lexical substitution, on two datasets covering a lexical sample and an all-words setting. We think this may prove useful for future researchers working on the lexical substitution task, as they will have a big picture overview of the resources and tools that are useful for this task, along with comparative evaluations.

As future work, we intend to connect the monolingual and cross-lingual lexical substitution tasks by jointly exploiting several monolingual and multilingual resources.

Acknowledgments

We would like to thank the reviewers for their thoughtful comments, which helped improve an earlier version of this paper. 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

E. Agirre and P. Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

C. Akkaya, J. Wiebe, and R. Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 190–199, Singapore.

I. Androutsopoulos and P. Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38(1).

S. Bangalore and O. Rambow. 2000. Corpus-based lexical choice in natural language generation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 464–471, Morristown, NJ, USA. Association for Computational Linguistics.

R. Barzilay and L. Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*.

S. Beale, B. Lavoie, M. McShane, S. Nirenburg, and Tanya Korelsky. 2004. Question answering using ontological semantics. In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, TextMean '04, pages 41–48, Morristown, NJ, USA. Association for Computational Linguistics.

S. Bergsma, D. Lin, and R. Goebel. 2009. Web-scale n-gram models for lexical disambiguation. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

C. Biemann. 2006. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 73–80, Morristown, NJ, USA. Association for Computational Linguistics.

C. Biemann. 2010. Co-occurrence cluster features for lexical substitutions in context. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-5, pages 55–59, Morristown, NJ, USA. Association for Computational Linguistics.

T. Brants and A. Franz. 2006. Web 1T 5-gram version 1. Linguistic Data Consortium.

M. Carpuat and D. Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing*, pages 61–72.

Y. S. Chan and H. T. Ng. 2007. Word sense disambiguation improves statistical machine translation. In *In 45th Annual Meeting of the Association for Computational Linguistics (ACL-07*, pages 33–40.

C. Chang and S. Clark. 2010. Practical linguistic steganography using contextual synonym substitution and vertex colour coding. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1194–1203, Cambridge, MA, October. Association for Computational Linguistics.

I. Dagan, O. Glickman, A. Gliozzo, E. Marmorshtein, and Carlo Strapparava. 2006. Direct word sense matching for lexical substitution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 449–456, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. Davidov and A. Rappoport. 2009. Enhancement of lexical concepts using cross-lingual web mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 852–861, Singapore, August. Association for Computational Linguistics.

L. Deléger and P. Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 2–10, Singapore, August.

W. Dolan, C. Quirk, and C. Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

P. Edmonds. 1997. Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 507–509.

N. Elhadad and K. Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 49–56, Morristown, NJ, USA. Association for Computational Linguistics.

O. Etzioni, K. Reiter, S. Soderl, and M. Sammer. 2007. Lexical translation with application to image search on the web. In *Proceedings of the Machine Translation Summit*.

E. Gabrilovich and S. Markovitch. 2007. 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.

R. Girju, A. Badulescu, and D. I. Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics.*

C. Giuliano, A. Gliozzo, and C. Strapparava. 2007. 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. Association for Computational Linguistics.

S. Hassan and R. Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *Proceedings of the Conference of the American Association for Artificial Intelligence*, San Francisco.

S. Hassan, A. Csomai, C. Banea, R. Sinha, and R. Mihalcea. 2007. 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. Association for Computational Linguistics.

D. Inkpen. 2007. A statistical model for near-synonym choice. *ACM Transactions on Speech and Language Processing*, 4:2:1–2:17, February.

A. Islam and D. Inkpen. 2009. Semantic Similarity of Short Texts. In *Recent Advances in Natural Language Processing V*, volume 309 of *Current Issues in Linguistic Theory*, pages 227–236. John Benjamins, Amsterdam & Philadelphia.

S. Jabbari, M. Hepple, and L. Guthrie. 2010. Evaluation metrics for the lexical substitution task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 289–292, Los Angeles, California, June. Association for Computational Linguistics.

S. Kim, Hee cheol Seo, and Hae chang Rim. 2004. Information retrieval using word senses: root sense tagging approach. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 258–265.

R. Krovetz. 1997. Homonymy and polysemy in information retrieval. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97*, pages 72–79.

M. Kubat and S. Matwin. 1997. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann.

T. K. Landauer and S. T. Dumais. 1997. Solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psych. Rev.*, 104(2):211–240. Cognitive view on LSA.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

D. Martinez, S. N. Kim, and T. Baldwin. 2007. Melb-mkb: lexical substitution system based on relatives in context. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 237–240, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. McCarthy and R. Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007*, pages 48–53.

D. McCarthy and R. Navigli. 2009. The english lexical substitution task. *Language Resources and Evaluation*, 43:139–159.

R. Mihalcea and P. Edmonds, editors. 2004. *Proceedings of SENSEVAL-3, Association for Computational Linguistics Workshop*, Barcelona, Spain.

R. Mihalcea, C. Corley, and C. Strapparava. 2006. Corpus-based and knowledge-based approaches to text semantic similarity. In *Proceedings of the American Association for Artificial Intelligence*, Boston, MA.

R. Mihalcea, R. Sinha, and D. McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden, July.

G. A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38:39–41.

J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of Association of Computational Linguistics*, Columbus, Ohio.

C. Monz. 2005. 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*, pages 520–527. ACM Press.

S. Pradhan, E. Loper, D. Dligach, and M. Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June. Association for Computational Linguistics.

J. Preiss, A. Coonce, and B. Baker. 2009. Hmms, grs, and n-grams as lexical substitution techniques: are they portable to other languages? In *Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning*, MCTLLL '09, pages 21–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

C. Quirk, C. Brockett, and W. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Process*, Barcelona, Spain.

Y. Shinyama, S. Sekine, and K. Sudo. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of the Second International Conference on Human Language Technology Research*.

C. Stokoe. 2005. Differentiating homonymy and polysemy in information retrieval. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 403–410, Morristown, NJ, USA. Association for Computational Linguistics.

T. Wang and G. Hirst. 2010. Near-synonym lexical choice in latent semantic space. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1182–1190, Morristown, NJ, USA. Association for Computational Linguistics.

M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 365–368.

L. Yu, H. Shih, Y. Lai, J. Yeh, and C. Wu. 2010. Discriminative training for near-synonym substitution. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1254–1262, Morristown, NJ, USA. Association for Computational Linguistics.

D. Yuret. 2007. 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. Association for Computational Linguistics.

S. Zhao, L. Zhao, Y. Zhang, T. Liu, and S. Li. 2007. 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. Association for Computational Linguistics.