Word Similarity Using Constructions as Contextual Features[1]

Nai-Lung Tsao
National Central University
No.300, Jhongda Rd. Jhongli City,
Taoyuan County 32001, Taiwan
beaktsao@stringnet.org

David Wible
National Central University
No.300, Jhongda Rd. Jhongli City,
Taoyuan County 32001, Taiwan
wible@stringnet.org

Abstract

We propose and implement an alternative source of contextual features for word similarity detection based on the notion of lexico-grammatical construction. On the assumption that selectional restrictions provide indicators of the semantic similarity of words attested in selected positions, we extend the notion of selection beyond that of single selecting heads to multiword constructions exerting selectional preferences. Our model of 92 million cross-indexed hybrid n-grams (serving as our machine-tractable proxy for constructions) extracted from BNC provides the source of contextual features. We compare results with those of a grammatical dependency approach (Lin 1998), testing both against WordNet-based similarity rankings (Lin 1998; Resnik 1995). Averaged over the entire set of target nouns and 10-best candidate similar words, Lin's approach gives overall similarity results closer to WordNet rankings than the constructional approach does, while the constructional approach overtakes Lin's in approximating WordNet similarity for target nouns with a frequency over 3000. While this suggests feature sparseness for constructions that resolves with higher frequency nouns, constructions as shared contextual features render a much higher yield in similarity performance in approximating WordNet similarity than grammatical relations do. We examine some cases in detail showing the sorts of similarity detected by a constructional approach that are undetected by a grammatical relations approach or by WordNet or both and thus overlooked in benchmark evaluations.

1. Introduction

Distributional approaches to semantics have contributed substantially to computational techniques for detecting or judging the semantic similarity of words for a wide range of applications. Such approaches work from the assumption that the distribution (or the set of contexts) of a word reflect the meaning of that word and, accordingly, that words with similar distributions have similar meanings (Harris 1954; 1968; Miller and Charles 1991; Lenci 2008, inter alia). Computational work taking such a distributional approach involves two dimensions: (1) some operationalization of the notion 'context' used in determining a word's distribution, and (2) some means of measuring similarity between or among sets of contexts that constitute a word's distribution. Such work typically involves extracting from a reference corpus the contexts of the candidate words, under some specified definition of context, and rendering these contexts as feature vectors in a vector space that can in turn be compared for (dis)similarity. In this paper we propose a novel construal of context and contextual features in determining word similarity distributionally and describe and evaluate an implementation of it.

A motivating premise for our approach is that in comparing words by comparing quantitative measures of their distributions, certain details of these distributions and the contexts that constitute them are obscured. For numerous applications, such as query expansion, document similarity judgment and document classification, this opacity may be irrelevant. There are, however, applications where the loss of some of this obscured detail comes at a cost, that is, where it may become relevant to ask for a pair or set of words not only 'How similar are they?' but 'How are they similar?' While current distributional approaches generally focus on the first question, we would like to build on those results to explore ways to further address the second.

2. The Basic Approach

Central to any implementation of distributional lexical semantics is the notion of context, or, as

[1] This research was supported by Taiwan's National Science Council through Grant #NSC 100-2511-S-008-005-MY3

Harris referred to this, a word's "environments" (1954, p. 146). Computational work on word similarity has operationalized context typically as features. These include unordered sets of co-occurent words attested within some window of proximity to the target word, i.e., bag-of-words (Dagan et al. 1993; Ng and Lee 1996; Tumuluru et al. 2012), ordered sequences of words, i.e., n-grams (Damashak 1995; Jones et al. 2006; Sahlgren et al. 2008; DeVine and Bruza 2010), ordered sequences of POS categories and collocations co-occurring with the target word (Ng and Lee 1996) and co-occurring words that stand in specified grammatical relation to the target word (Hindle 1990; Ruge 1992; Grefenstette 1994; Lin 1997, 1998; Geffet and Dagan 2009, inter alia). Distributional semantic work on word similarity over the past three decades has shown relatively little variety in how context has been operationalized, falling under one of these few types just mentioned. Probably the most linguistically sophisticated construal of context among these is the use of grammatical relations such as subject-verb, object-verb, adjective-noun as the contextual features. Crucial for us, these approaches that take grammatical relations as contextual features constitute, as Dagan (2000) points out, "a statistical alternative to traditional notions of selectional constraints and semantic preferences" (p. 3). Thus, as a feature of the noun *cell* reported in Lin (1998), the triples *cell, subject-of, absorb* and *cell, object-of, attack* indicate the selection of the noun *cell* by the verb *absorb* as its subject argument and by *attack* as its (direct) object argument. It is worth noting here that these grammatical relations (or selectional preferences) are head to head (that is, lexeme to lexeme) relations; a particular verb or preposition, for example, is seen as selecting for a particular semantic class (or set of classes) of noun.

The work reported here shares this assumption that semantic selection is a potentially rich source for identifying similar words. We suggest, however, that semantic selection is not always head-driven. More specifically, we explore an approach to detecting semantically restricted positions that are governed by larger multiword units. In other words, we consider the possibility of positions that are selected by something more like a construction (roughly along the lines of Fillmore et al. 1988; Goldberg 2006; inter alia) rather than a lexical head. For example, taking discrete grammatical relations as a feature, standing in object relation to the transitive verb *remove* would be one feature that various nouns could share, nouns attested as object of *remove*. If, however, we expand the notion of selection beyond single heads as the selecting expression such as a single verb, we create the possibility of not simply the verb *remove* as the contextual feature of its objects, as in (1), but also of that noun slot taking the more enriched context in (2) as a feature.

(1) remove [noun]
(2) undergo surgery to remove a [noun]

While taking (2) rather than (1) as the contextual feature of the [noun] slot would of course reduce dramatically the set of nouns attested in that slot, our motivating assumption is that it offers the possibility of narrowing the semantic class of nouns we would expect to find there. At the same time, and of equal interest to us, (2) provides a more articulated, fleshed out context.

Here perhaps the relevance of constructional selection and a constructional approach to contextual features for some applications can be made a bit clearer. Thesaurus construction is a fundamental domain of word similarity application which itself feeds numerous other applications. One area of such applications for thesauri where contextual detail becomes relevant is language learning. For language learners seeking to expand their vocabulary, a decontextualized list of discrete synonyms is of limited value, as attested by the uses that learners can create when relying on traditional thesauri. What does constitute a potentially useful source of traction for mastering unknown words from known ones, however, is access to exactly which multiword patterns of behavior of the known word generalize to the unknown word(s) and which patterns do not. Such patterning may elude what can be captured even by grammatical relations. The noun *place* stands in the grammatical relation of object to the verb *take* in both *take place* (as in *occur*) and *take the place of* (as in *replace*). Of course, it could be assumed that contributions of such nuanced differences come out in the wash when taken with broader distributional trends from sufficiently large corpora. We would like to consider the alternative possibility that incorporating such nuance as part of the contextual features used in statistical approaches to distribution can contribute to word similarity research.

In what follows we describe one specific implementation of detecting constructional selection to determine word similarity and compare it to an approach that uses head to head grammati-

cal relations (subject-verb; object-verb, etc.). Since Lin (1998) is the most widely referenced approach using grammatical dependencies as a feature type for word similarity detection (Padó and Lapata 2007; Geffet and Dagan 2009; Kotlerman et al. 2009 ; inter alia), we run an implementation of Lin (1998) as our point of comparison to a grammatical relations approach. We first describe our method and then Lin's in section 3, and then in section 4 report and compare results produced from these two approaches applied to the same set of nouns.

3. Methods

3.1. An Implementation of the Constructional Approach

The challenge posed by our approach is how to automatically identify positions that are semantically selected. Since we are trying to identify selectional preferences imposed not by lexical heads but by multiword lexico-grammatical constructions, extracting head-to-head grammatical relations (e.g., subject-verb) will not suffice. That is, we need an enriched version of context and contextual features. To motivate our means of identifying constructional selection, an example in (3) can show the sort of linguistic phenomenon we aim to detect.

(3) have no [noun] but [to verb]

There are 325 tokens in BNC (British National Corpus) that instantiate this pattern (e.g., *have no choice but to accept…*). Crucially, considering the [noun] slot in those 325 tokens, 323 of them are tokens of just three distinct nouns: *choice* (freq: 137), *option* (freq: 110), *alternative* (freq: 76). Clearly, these three nouns are semantically similar. This semantic similarity could be fortuitous or it could reflect that this position is subject to selectional preference. Pursuing this latter possibility, the question is what might be the source of the semantic preference. It cannot plausibly be attributed to a specific lexical head, say an argument-taking predicate; in (3) that would be the semantically uninformative light verb *have*. Hence, this sort of semantic selection will fly below the radar of grammatical dependency approaches to semantic similarity. We suggest that the noun slot in (3) is semantically selected by the entire surrounding construction: *have no ____ but [to verb]*. This surrounding construction we will take as a shared feature of the three nouns attested: *choice*, *option*, *alternative*. We call this phenomenon *constructional selection*.

The challenge now can be stated as how to automatically identify loci of constructional selection, paradigms like the noun slot in (3), which are semantically restricted yet not by a lexical head. For this, we first need a means of identifying candidate constructions from corpora. We do this using the notion of hybrid n-gram from Wible and Tsao (2010) as the machine-tractable proxy, and then identify positions within them that exhibit semantic selection. We describe these two steps in turn.

Hybrid N-grams and Semantically Selected Slots

We operationalize the class of contexts that potentially exhibit constructional selection with the notion of hybrid n-gram (Tsao and Wible 2009; Wible and Tsao 2010). Hybrid n-grams are a variation of n-gram which, in addition to lexemes or specific word forms as grams, also admit part-of-speech category labels as a gram type. Thus, in addition to a traditional tri-gram *consider yourself lucky*, a hybrid tri-gram would also include *consider yourself [adj],* a more abstract version that thereby describes the tokens *consider yourself lucky* and *consider yourself fortunate*, for example. Hybrid n-grams would also include *consider [reflx prn] [adj]*, *[verb] [reflx prn] lucky*, and so on. A requirement we impose on hybrid n-grams for our language model is that they must each include one lexical gram (at least one gram that is either a lexeme or a specific word form of a lexeme). In this sense, all hybrid n-grams are lexically anchored. (See Wible and Tsao (2010) for details on hybrid n-gram extraction.)

Our language model consists of all hybrid n-grams from 3 to 6 grams in length extracted from BNC. As with any n-gram model spanning more than one value of n, there is substantial redundancy in our first-pass model, which is magnified because of our inclusion of more abstract part-of-speech grams. To mitigate the effects of this redundancy, we prune more abstract counterparts of a more specific hybrid n-gram when the more specific version accounts for 80% or more of the tokens of the more abstract one. Thus *point [prep] view* is pruned since more than 80% of its tokens in BNC are cases of the more specific *point of view*. Likewise we prune shorter n-grams in cases where 80% of their tokens are also tokens of the n+1 counterpart hybrid n-gram. Thus, *the*

other hand is pruned because a threshold proportion of its tokens are part of the longer *on the other hand*. (See Wible and Tsao 2010 for details on extraction and pruning of hybrid n-grams.) To prevent a proliferation of unhelpful contexts such as *of the [noun]*, we further require that the hybrid n-gram must contain at least one lexical content word in addition to the target noun slot. The fully pruned version of the model contains 92 million unique hybrid n-grams.

Detecting Selectional Preferences in Hybrid N-gram Contexts

The pruned model of 92 million hybrid n-grams serves as the pool of candidate contexts we use to determine both the distribution of a word and its similarity to the distribution of other words. Two words share a context in case they are attested in the same gram or slot in a hybrid n-gram; that is, the two words share this contextual feature. Thus, *option* and *choice* have the shared feature of being occupants of the [noun] slot in *have no [noun] but [to verb]*. Put in structuralist terms, the words *option* and *choice* share a precise context as members of the same paradigmatic slot within a syntagmatic sequence.

As we noted with the pattern in (3) above, not all slots (or paradigms) in hybrid n-grams are selective. Thus, we need some further means of identifying those that are. Recall the two slots in the hybrid n-gram in (3) (repeated here) differ in selectivity and thus suggest the sort of distinction we need to make to identify selectionally restrictive slots (of the pattern's 325 tokens, only 5 different nouns account for the 325 noun tokens but 172 different verbs for the 325 tokens filling the [to verb] slot).

(3) have no [noun] but [to verb]

To identify the selective slots, we require that a word must account for at least 10% of the tokens attested in that specific slot of that hybrid n-gram in order for that hyrbrid n-gram to qualify as a contextual feature of that word. Accordingly, for two words to share a contextual feature, they must each account for 10% of the tokens attested in the same slot in the same hybrid n-gram. Thus, *trouble* and *problem* share a contextual feature by virtue of each accounting for minimally 10% of the tokens attested in the [noun] slot of the hybrid n-gram: *have a lot of [noun] with*. *Trouble* occurs in 12 of the 32 tokens of this construction and *problem* in 4 of the 32.

Recall that we further require shared contexts contain, in addition to the target noun slot, at least one lexical content word to avoid a massive proliferation of uninformative shared contexts such as: *and the [noun]*.

It is worth noting here that our means of identifying contexts that have selectionally restrictive slots makes no reference to semantic knowledge sources such as WordNet (Miller 1995) or other thesauri, but relies simply on frequency distribution profile of words attested in a paradigm slot. Note also that there could be a variety of ways to identify selective slots within hybrid n-grams, and our use of the 10% occupancy threshold is a first and basic approximation.

We measure similarity between two words by simply determining the number of shared contextual features, operationalized as shared membership in the same selective slots within the same hybrid n-gram. The set of nouns we consider are all and only the nouns found in WordNet and that have a frequency in BNC ≥ 100. We exclude from consideration compound nouns found in WordNet. This leaves us with 12,061 nouns. For every pair of such nouns, we calculate a similarity score for a target word *t* as follows:

$$\frac{\log(|P|) * \log(|C|)}{\log(f(w))}$$

, where $|P|$ is the number of unique shared contexts or hybrid n-grams between two words, $|C|$ is the number of unique shared lexical collocates occurring in the set of shared contexts and *w* is the frequency of the candidate similar word.

The reason we take into account $|C|$, the number of unique shared collocates, is basically to reward lexical diversity across shared contexts on the assumption that greater diversity within the circle of 'mutual friends' for two words indicates greater similarity of those two words. Consider the target noun *wealth* and two of its candidate similar nouns—*range* and *lack*—which have the same value of $|P|$, the same number of shared contexts with *wealth*; (11 contexts each). There are seven different collocates in the eleven contexts shared by *wealth* and *range* (e.g., *draw* in *draw on a [wealth/range/…] of; available* in *the [wealth/range/…] of [noun] available from*), but there are only three distinct collocates in the eleven contexts shared by *wealth* and *lack* (e.g. *experience* in *his [wealth/lack…] of experience*). Including $|C|$ in

our equation is a means of differentiating these otherwise indistinguishable cases.

Using similarity scores calculated with the above equation, we can generate for each of the 12061 target nouns a ranked list of similar nouns. In this paper we consider only the 10-best similar nouns created by these rankings. While Lin 1998 uses 200-best, and 10-best will certainly yield lower recall and hurt evaluation scores against benchmarks, we find little motivation for considering more than 10 similar nouns in light of the fact that, for example, WordNet averages under 2 words per synset for all its nouns, even for high frequency nouns.

3.2. Lin's Approach

To compare our constructional selection results with a head-driven selectional approach that uses grammatical dependency, we implement Lin (1998) using BNC as the reference corpus as a representative of the latter.

Lin's version requires a parsed corpus in order to extract the grammatical relations as contextual features. For this we use Link parser (Sleator and Temperley 1993) to parse BNC and extract all head-to-head dependency relations as triples: word 1, rel, word 2. Lexical categories of the words extracted for dependency relations were noun, verb, adj, adv, prep. From these triples we retain only those that include a noun and filter out redundancies (for example, for a token dependency 'brown dog' Link parser extracts two triples 'brown modif dog' and 'dog noun-mod brown' but we retain only the latter). About 78 million such triples are extracted and retained. We measure word association strength between the two words in each triple using the following MI measure from Lin (1998).

$$I(w,r,c) = \log \frac{\|w,r,c\| \times \|*,r,*\|}{\|w,r,*\| \times \|*,r,c\|}$$

where w is the target word, r is the dependency relation (subject of; object of, etc.), and c is a collocate standing in relation r to word w. $\|w,r,c\|$ denotes the frequency of the relational triple in parsed BNC. When w, r, or c is replaced by the wild card(*), the frequency of the relational triples that match the rest of the pattern is summed up. For example, $\|\, cell, subject-of, * \,\|$ is the total number of occurrences of *cell-subject* relationships for any c in parsed BNC.

Taking all nouns found in WordNet with frequency in BNC ≥ 100 (compound nouns excluded), for each pair of such nouns we calculate a similarity score following Lin (1998) with the following equation:

$$\frac{\sum_{(r,c)\in T(w_1)\cap T(w_2)} \left(I(w_1,r,c) + I(w_2,r,c)\right)}{\sum_{(r,c)\in T(w_1)} I(w_1,r,c) + \sum_{(r,c)\in T(w_2)} I(w_2,r,c)}$$

where $T(w)$ is the set of pairs (r, c) such that $I(w, r, c)$ is positive.

Using similarity scores calculated accordingly, we can generate for each target noun a ranked list of similar nouns.

4. Evaluation and Comparison [2]

We first consider here the extent of overlap in the 10-best results produced by the constructional and relational approaches, then compare both constructional and relational approaches as they approximate word similarity scores derived from WordNet, and finally elaborate on specific illustrative cases.

4.1. Comparison of Overlap in Results: Constructional and Relational Approaches

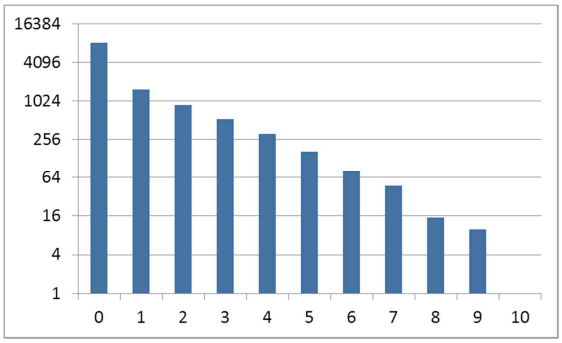

Figure 1. Overlap between 10-best lists of similar words by Lin (1998) and construction approach

For each of the two approaches, we generated rankings of similar words for all 12061 target nouns found in WordNet (compounds excluded) and with a minimum frequency of 100 in BNC. Figure 1 shows the comparison for overlap of the 10-best lists, with the x axis showing the number of similar nouns out of the two 10-best lists with increasing overlap from left to right (from 0 to 10 overlapping similar words from the two methods) and the y axis representing the number of target nouns whose 10-best similar words show that amount of overlap. As the figure makes apparent, the two approaches yield widely

[2] Similarity rankings available at http://www.stringnet.org

divergent results, with well over half of the 12061 nouns tested showing no overlapping similar words from the two 10-best lists.

We should note that our purpose for comparing results of our approach with Lin's here is not to use Lin's as a benchmark for our method to aspire to. Rather, we are interested in the differences in that come of using head-to-head grammatical dependencies as in Lin's method compared to using constructional selection as the contextual feature type that reflects word similarity as in ours. Before discussing these differences, we first compare the performances of the two approaches to similarity results based on WordNet.

4.2. Comparisons with WordNet-based Similarity Results

Method of Comparison

Here we compare the automatically generated results of the constructional approach (cxnl) and the relational approach (rlnl) each to similarity results based on the handcrafted resource, WordNet (wn). We first need similarity results from WordNet. For this, we use WordNet 3.0 (Miller 1995) and the following word similarity measure applied to WordNet from Lin (1997):

$$\mathrm{sim}_{\mathrm{wnc}}(c_1,c_2) = \max_{c \in \mathrm{super}(c_1) \cap c \in \mathrm{super}(c_2)} \frac{2\log P(c)}{\log P(c_1) + \log P(c_2)}$$

$$\mathrm{sim}_{\mathrm{wn}}(w_1,w_2) = \max_{c_1 \in S(w_1) \cap c_2 \in S(w_2)} (\mathrm{sim}_{\mathrm{wnc}}(c_1,c_2))$$

where $S(w)$ is the set of senses of word w in WordNet, super(c) is the set of super-ordinate classes of concept c in WordNet. The probability of a concept is estimated by the sense tag count information in WordNet. We use Resink's approach (1995) to estimate the probabilities. The probability of a concept subsumes all probabilities of its descendants in WordNet.

With the WordNet-based similarity, we have word similarity results on the same noun set for three different approaches: construction-based (cxnl), grammatical relation–based (rlnl), and WordNet-based (wn). We use Lin's approach (1998) to measure two pair-wise correlations of results: cxnl-wn; rlnl-wn. The correlation for a pair of methods is arrived at following Lin (1998). For a target word, two similar word lists based on two methods are represented as follows:

method 1: $(w_1, s_1), (w_2, s_2), ..., (w_n, s_n)$

method 2: $(w'_1, s'_1), (w'_2, s'_2), ..., (w'_n, s'_n)$

where w is a candidate similar word and s is the similarity score between the target word and w.

The set of similar words and similarity scores for each target word schematized above can be taken as a vector, the features of that vector being the pairings of similar word and similarity score $(w_1, s_1)\ldots(w_n, s_n)$. The similarity between the results of two methods is taken as the cosine of these two vectors for each target word averaged across all target words, as defined in the following equation:

$$\frac{\sum_{w_i = w'_j} s_i s'_j}{\sqrt{(\sum_{i=1}^{n} s_i^{\,2})(\sum_{j=1}^{n} s'^{\,2}_j)}}$$

We apply this equation to two pairings of methods for comparison: constructional:WordNet (cxnl:wn) and relational:WordNet (rlnl:wn).

Results and Discussion of WordNet Comparisons

The overall similarity scores for the pairings of approaches (see below) show the grammatical relations approach approximating WordNet-based similarity results more closely than the constructional approach does.

cxnl-wn: 0.0411
rlnl-wn: 0.0565

Figure 2 represents the similarity to WordNet results of the constructional and relational methods broken down into frequency bands for target words (frequency in BNC). The y axis represents cosine averages of constructional:WordNet results and relational:WordNet results, i.e., the similarity of these two approaches to WordNet-based results, and the x axis is the frequency of the target words receiving these similarity scores. What is worth noting in Figure 2 and not apparent from the overall scores is that the constructional approach performance catches up to the relational approach at a frequency of 3000 and overtakes it for frequencies above that.

This raises the question of how the trend here would play out with higher frequencies from a larger corpus. In this regard, we also consider the average number of features responsible for these scores under the two different methods. This is shown in Figure 3.

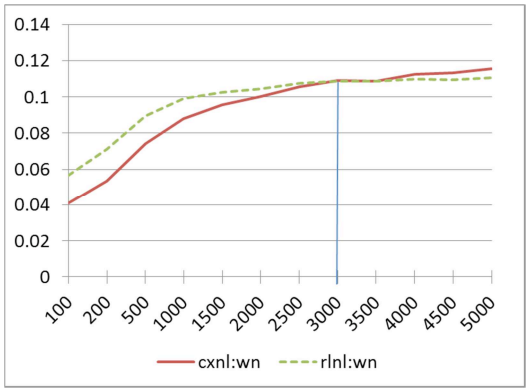

Figure 2. similarity score (y axis) with WN and frequency of target nouns (x axis)

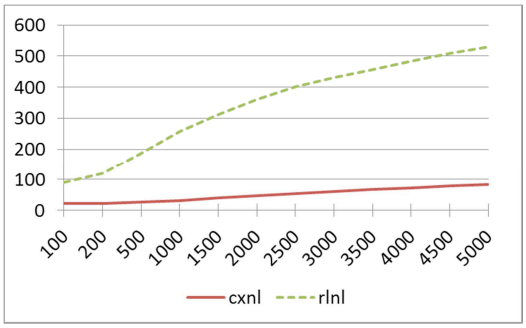

Figure 3. x axis: average number of shared features of 10-best sim nouns; y axis: frequency of target noun.

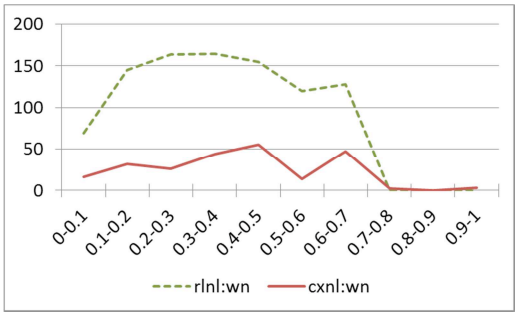

Figure 4. x axis: score of approximation to WN similarity results; y axis: number of shared features for 10-best sim nouns

While Figure 2 might suggest that the constructional approach is relatively data-hungry and suffers from feature sparseness at the lower frequency levels, another perspective on this is suggested by Figure 3 and Figure 4, which show a notable difference in the "yield" of similarity performance by the two different sorts of features; i.e., constructions compared to grammatical relations as features. Notably, Figure 3 shows a comparatively sharp rise in the number of features used by the relational approach, reaching over 500 for the high frequent words, whereas the number of constructional features rises gradually and remains well under 100 for all levels of frequency. This suggests a relatively healthy 'return on investment' (ROI) or what we might call 'feature yield' for constructions as contextual features.

Some Specific Suggestive Cases

In considering the results above, it is important to remember that we are not aspiring to superiority to previous distributional approaches on some single linear scale of performance, though this impression is hard to avoid under the need to offer some comparative evaluation. What we would like to suggest, rather, is that a constructional approach of the sort we propose shows sensitivity to similarities between (among) words that current distributional approaches have not, similarities worth trying to capture. This latter purpose raises difficulties since, we will argue here, the traditional benchmarks for evaluating word similarity results (i.e., traditional thesauri or WordNet) are also less attuned to some of the dimensions of semantic similarity that our approach seems able to capture.

To shed some light on what these different approaches contribute, we consider results for two different target nouns: *deal* and *ground*

Rank	Constructional Method	Grammatical Relation Method
1	*floor	land
2	reason	field
3	basis	site
4	fact	area
5	cause	surface
6	term	*floor
7	way	water
8	bed	building
9	garden	space
10	issue	path

Table 1. Ranked 10-best similar nouns for *ground* from constructional vs grammatical relation methods

Rank	Constructional Method	Grammatical Relation Method
1	*amount	*agreement
2	*lot	contract
3	bit	arrangement
4	*agreement	*lot
5	degree	proposal
6	source	move
7	lack	plan
8	thing	scheme
9	sense	offer
10	range	*amount

Table 2. Ranked 10-best similar nouns for *deal* from constructional vs grammatical relation methods

For the target noun *ground,* the 10-best lists of our construction method and Lin's grammatical dependency method, shown in Table 1, have only one similar word in common: *floor*. But note the complementarity of the two lists. What we would call true positives from Lin's list that we miss include: *land*, *field*, *site*, *area*, *surface*. On the other hand, what we would consider true positives from the constructional list includes: *reason*, *basis*, *cause*. These are apparently similar in more figurative, metaphorical senses missing from the grammatical dependency list in this case. While WordNet's ranks *reason* and *basis* as the two top similar nouns for *ground*, *cause* is missed by WordNet, its similarity to *ground* receiving a score of 0.

For the target noun *deal*, the ranked list of 10-best similar words generated by Lin and the list generated by our constructional method have only 3 nouns in common, as shown in Table 2.

Focusing on where results of the two methods diverge, it is worth noticing the constructional contexts that *deal* shared with some of the words from its 10-best list that did not appear on the dependency relation or WordNet list. The noun *bit* ranks 3rd in similarity to *deal* under the construction approach but 142nd under Lin and 84th under WordNet. A few of the 92 hybrid n-grams that are shared features of *deal* and *bit* (accounting for more than 10% each of the tokens in the [noun] slot), are given in (4-10):

(4) take a [adj] [noun] of time
(5) make a [adj] [noun] of difference
(6) have a [adj] [noun] of money
(7) under a [adj] [noun] of pressure
(8) be a [adj] [noun] older than
(9) not make a [adj][noun] of
(10) get a fair [noun] of

To see the potential contribution of hybrid n-grams as a feature type for detecting similar words, we can ask whether these instances of shared contexts in (4-10) would be detectable under context construed as, say, n-grams or head-to-head grammatical dependencies or collocation. We consider only (4) in some detail.

The noun slot in (4) selects for both *deal* and *bit*. This hybrid n-gram is instantiated by 53 tokens in BNC; 22 of them with the noun *deal*, 7 of them with *bit* (and 19 of them with the noun *amount*—a conspicuous clue to the sense that *bit* and *deal* share in common here). But would that slot select for these same nouns if we reduced the contextual features to one single selecting head or collocate? The noun slot heads the object NP of the verb *take* in (4), so *take* would be the candidate verb selecting *bit* or *deal* as its object. But the light verb *take* does not select either of these nouns as object.*Take* is in fact part of a V-N collocation here, the N of the collocation being *time* in *take…time*, not the intervening [noun] slot where *bit* and *deal* occur. This excludes selection by or collocation with the verb as responsible for the selection here. Nor does the [adj] slot serve as collocate. Neither *bit* or *deal* is selected by the adjective; it is not a specific adjective here but an open adjective slot, and crucially, there is virtually no overlap in the adjectives that co-occur with *bit* and with *deal* in this context (the only shared adjective is 'good', one token each co-occurring with *bit* (freq = 7) and *deal* (freq = 22)).

Note that a version of this context in (4) rendered as a traditional n-gram made of only lexical grams and no POS slots would not select *bit* and *deal* here in the same slot and therefore detect no shared distribution for them. It requires the abstract POS slot of the hybrid n-gram to capture this portion of their shared distribution.

This covers the relations that could be captured by head to head grammatical dependencies, collocations, and n-grams. Similar considerations would show the contribution of the hybrid n-grams in (5-10) as a sampling.

5. Conclusion

An alternative construal of context in terms of the notion of construction could enrich the sorts of semantic similarity susceptible to detection. Lin's grammatical dependency approach yields substantial results that our approach misses and for which we have no straightforward means of emulating. Nor is it our intention to attempt that. Rather, and on the other hand, our results suggest that construing contextual features as multiword lexico-grammatical wholes can uncover loci of semantic selection that attract similar words. Evaluation against WordNet-based results shows also that despite an appearance of feature sparseness, constructions are comparatively potent indicators of similarity, requiring fewer features to yield similarity results approximating benchmarks. Future work could determine whether constructions reward the use of larger corpora with increased yield in similarity judgments.

References

Ido Dagan. 2000. Contextual word similarity. *Handbook of Natural Language Processing*, 459-475.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. 1993 Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pp. 164-171. Association for Computational Linguistics, 1993.

Marc Damashek. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, *267*(5199), 843-848.

Lance De Vine and Peter Bruza. 2010. Semantic oscillations: Encoding context and structure in complex valued holographic vectors. In *Proceedings of AAAI Fall Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes*.

Charles J. Fillmore, Paul Kay, and Mary Katherine O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: the Case of *Let Alone*. *Language* 64: 501–538.

Maayan Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, *35*(3), 435-461.

Adele Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston, MA.

Zellig Sabbettai Harris. 1954. Distributional structure. *Word* 10:146–162.

Zellig Sabbettai Harris. 1968. *Mathematical structures of Language*. New York: Wiley.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics* (pp. 268-275). Association for Computational Linguistics.

Michael N. Jones, Walter Kintsch, and Douglas J.K. Mewhort. 2006. High-dimensional semantic space accounts of priming. *Journal of memory and language*, 55(4), 534-552.

Lili Kotlerman, Ido Dagan, Idan Szpektor and Maayan Zhitomirsky-Geffet. 2009. Directional distributional similarity for lexical expansion. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 69-72)

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. in Lenci Alessandro. (ed.), From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, special issue of the *Italian Journal of Linguistics*, 20/1: 1-31.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL/EACL-97*, pages 64-71, Madrid, Spain, July.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pp. 768-774.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, No. 11: 39-41

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6 (1): 1–28.

Hwee Tou Ng, and Hian Beng Lee. 1996 Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pp. 40-47.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161-199.

Phil Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence(IJCAI-95)*.

Gerda Ruge. 1992. Experiments on linguistically-based term associations.*Information Processing & Management*, *28*(3), 317-332.

Magnus Sahlgren, Anders Host, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 1300–1305).

Daniel Sleator and Davy Temperley. 1993. Parsing English with a Link Grammar. *Third International Workshop on Parsing Technologies*.

Nai-Lung Tsao and David Wible. 2009. A method for unsupervised broad-coverage lexical error detection and correction. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 51-54)

Anand Karthik Tumuluru, Chi-Kin Lo and Dekai Wu. 2012. Accuracy and robustness in measuring the lexical similarity of semantic role fillers for automatic semantic MT evaluation. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation* (pp. 574-581)

David Wible and Nai-Lung Tsao. 2010. StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics* (pp. 25-31)