Evaluating Lexical Substitution: Analysis and New Measures

Sanaz Jabbari, Mark Hepple, Louise Guthrie

Department of Computer Science, University of Sheffield
S.Jabbari@dcs.shef.ac.uk, M.Hepple@dcs.shef.ac.uk, L.Guthrie@dcs.shef.ac.uk

Abstract

Lexical substitution is the task of finding a replacement for a target word in a sentence so as to preserve, as closely as possible, the meaning of the original sentence. It has been proposed that lexical substitution be used as a basis for assessing the performance of word sense disambiguation systems, an idea realised in the English Lexical Substitution Task of SemEval-2007. In this paper, we examine the evaluation metrics used for the English Lexical Substitution Task and identify some problems that arise for them. We go on to propose some alternative measures for this purpose, that avoid these problems, and which in turn can be seen as redefining the key tasks that lexical substitution systems should be expected to perform. We hope that these new metrics will better serve to guide the development of lexical substitution systems in future work. One of the new metrics addresses how effective systems are in ranking substitution candidates, a key ability for lexical substitution systems, and we report some results concerning the assessment of systems produced by this measure as compared to the relevant measure from SemEval-2007.

1. Introduction

Lexical substitution is the task of finding a replacement for a target word in a sentence so as to preserve, as closely as possible, the meaning of the original sentence. For example, we might replace the target word *match* with *game* in the sentence *they lost the <u>match</u>*. Since target words may be *sense ambiguous* (as is the word *match* in the above example), correct lexical substitution will in general require that *word sense disambiguation* (WSD) is implicitly achieved, i.e. that amongst the word's alternative senses, the sense that is needed in the *given sentential context* be identified. McCarthy (2002) proposed that lexical substitution be used as a basis for evaluating WSD systems, as it is a fairly narrow task where performance will directly reflect correct WSD, and also a task which, if done effectively, could in turn contribute to performance on broader tasks (e.g. sentence paraphrase). Crucially, in contrast to standard WSD evaluations, which use sense-tagged gold standard data, lexical substitution as an approach to evaluation side-steps the divisive issue of what is the appropriate *sense inventory* that should be used, an issue which ultimately may have no 'true' answer. Related issues such as identifying an appropriate granularity of sense are likewise side-stepped. This idea was eventually realised at SemEval-2007 as the English Lexical Substitution Task (here called ELS07), as described by McCarthy & Navigli (2007).

For any competitive exercise, such as any of the SemEval tasks, the scoring metrics that are used to evaluate system outputs form a crucial part of what *defines* the subtasks that systems are asked to perform. In this paper, we examine the evaluation metrics used in ELS07, and argue that they have some significant problems. In the light of this analysis, we then propose some alternative measures that avoid these problems, which we believe will better serve to guide the development of lexical substitution systems in future work. In what follows, we begin by introducing the English Lexical Substitution Task of SemEval-2007, and consider the dataset that was created for it, for distribution to participants to aid system development, and for evaluating the performance of the final competing systems. We then present and analyse the evaluation metrics used by ELS07, and then propose some alternative measures that avoid the problems identified.

Ideally, we would apply the newly proposed measures to the outputs of the systems that took part in ELS07. As we will see, however, this cannot be done, as the new measures significantly *change* the tasks that form the basis of evaluation, and so we would be scoring systems on their performance at tasks that they were not created, or ever intended, to perform. However, we do present some preliminary comparative analysis between one of the ELS07 metrics and a newly proposed measure of the ability of systems to rank alternative candidates for lexical substitution.

2. The English Lexical Substitution Task

The English Lexical Substitution Task of SemEval-2007 requires systems to provide substitution candidates for an identified target word appearing in a specified sentence (McCarthy & Navigli, 2007: M&N). To reiterate the example above, a possible test sentence might be *they lost the <u>match</u>*, in which the target word is identified to be *match*. A plausible substitution for this example is *game*, which is clearly also sense ambiguous, but which when substituted into the sentence is itself disambiguated by the context to yield an appropriate sense and an overall sentence interpretation that is close to the original. Note that identifying an appropriate lexical substitute is not just a matter of finding a term that shares a suitable sense with the target term. Various factors interact to affect the suitability of a term to a given context, like stylistic considerations such as formality (e.g. *inebriate* vs. *drunk* vs. *pissed*), or collocational factors (e.g. *strong tea* vs. *powerful tea*).

The terms selected for use a target terms for ELS07 were all required to be sense ambiguous and have at least one synonym (serving as an indicator that the term was likely to be substitutable). The overall set of target terms was required to include a reasonable representation of different parts of speech, covering nouns, verbs, adjectives and adverbs. A total of 201 target words were chosen, and then an overall dataset of around 2000 sentences selected, providing 10 test sentences for each target word. System outputs are evaluated against a set of candidate substitutes proposed by human subjects for the test items. Five human annotators were asked to suggest one or more (up to three) substitutes for the target word of each test sentence, and their collected suggestions serve as the gold standard against which system outputs are scored. Around 300 of the sentences were distributed as development data, and the remainder retained for the final evaluation.

For each test item, the gold standard records not just the *set* of substitutes suggested by the human annotators, but also the *count* of annotators that proposed each candidate, i.e. since a term proposed by five appears a stronger substitution candidate than one proposed by just one annotator. This count information feeds into the scoring process, with more credit being assigned for correctly returning a high-count gold standard term than a low-count one.

3. Notation

To assist the definition of the scoring metrics, we formally characterise the data set as follows. For each sentence t_i in the test data ($1 \leq i \leq N$, N the number of test items), let H_i denote the set of human proposed substitutes. For each t_i, there is a function freq_i which returns this count for each term within H_i (and 0 for any other term), and a value $\mathit{maxfreq}_i$ corresponding to the maximal count for any term in H_i. The pairing of H_i and freq_i might be seen as providing a *multiset* representation of the human answer set. (However, standard multiset operations, such as multiset intersection, do not yield correct definitions for the metrics that are to be defined, so the definitions should be taken as given.) Note that we use $|S|^i$ in what follows to denote the *multiset cardinality* of S according to freq_i, i.e. $\Sigma_{a \in S} \mathit{freq}_i(a)$. Some of the ELS07 metrics use a notion of *mode* answer m_i, which exists only for test items that have a single most-frequent human response, i.e. a *unique* $a \in H_i$ such that $\mathit{freq}_i(a) = \mathit{maxfreq}_i$.

To adapt an example from M&N for use later in the paper, an item with target word *happy* (adj) might have human answers $\{\mathit{glad}, \mathit{merry}, \mathit{sunny}, \mathit{jovial}, \mathit{cheerful}\}$ with associated counts (3,3,2,1,1) respectively. For convenience, we will abbreviate this answer set as $H_i = \{$*G:3,M:3,S:2,J:1,Ch:1*$\}$, as a short-hand reminder of its content, where it is used later in the paper.

4. Best Answer Measures

Two of the ELS07 tasks address how well systems are able to find a 'best' substitute for a test item, for which the scoring of individual test items is as follows:[1]

$$\mathit{best}(i) = \frac{\sum_{a \in A_i} \mathit{freq}_i(a)}{|H_i|^i \times |A_i|}$$

$$\mathit{mode}(i) = \begin{cases} 1 & \text{if } \mathit{bg}_i = m_i \\ 0 & \text{otherwise} \end{cases}$$

For the first task, a system can return a *set* of answers A_i (the answer set for item i), but since the score achieved is divided by $|A_i|$, returning multiple answers only serves to allow a system to 'hedge its bets' if it is uncertain which of its candidate responses really is the best. The optimal score on a test item is achieved by returning a single answer whose count is $\mathit{maxfreq}_i$, with proportionately lesser credit being received for any answer in H_i with a lesser count. For the second task, which uses the *mode* metric, only a single system answer – its 'best guess' bg_i – is allowed, and the score is simply 0 or 1 depending on whether the best guess is the mode. (In practice, the single 'best guess' answer is taken as the first amongst the answers returned for the first task.)

Overall performance is computed by averaging across a broader set of test items (which for the second task includes only items that have a mode value). M&N distinguish two overall performance measures: *Recall*, which averages over all relevant items, and *Precision*, which averages only over the relevant items *for which the system gave a non-empty response*.

We next discuss these measures and make an alternative proposal. The task for the first measure seems a reasonable one to include, i.e. assessing the ability of systems to provide a 'best' answer for a test item, but allowing them to offer multiple candidates (to 'hedge their bets'). However, the metric is unsatisfactory in that a system that performs optimally in terms of this task (i.e. which, for every test item, returns a single correct 'most frequent' response) will get a score that is well below 1, because the score is also divided by $|H_i|^i$, the multiset cardinality of H_i, whose size will vary between test items (being a consequence of the number of alternatives suggested by the human annotators, each offering between 1 and 3 substitutes), but which will typically be larger than the numerator value $\mathit{maxfreq}_i$ of an optimal answer (unless H_i is singleton). This problem is fixed in the following modified metric definition, by dividing instead by $\mathit{maxfreq}_i$, as then a response containing a single optimal answer will score 1.

$$\text{(new) } \mathit{best}(i) = \frac{\sum_{a \in A_i} \mathit{freq}_i(a)}{\mathit{maxfreq}_i \times |A_i|}$$

$$\mathit{best}_1(i) = \frac{\mathit{freq}_i(\mathit{bg}_i)}{\mathit{maxfreq}_i}$$

For example, with human answer set $H_i = \{$*G:3,M:3,S:2,J:1,Ch:1*$\}$, an optimal response $A_i = \{M\}$ receives score 1, where the original metric gives score 0.3. Singleton responses containing a correct but non-optimal answer receive proportionately lower credit, e.g. for $A_i = \{S\}$ we score 0.66, as compared to 0.2 for the original metric. For a non-singleton answer set including, say, a correct answer and an incorrect one, the credit for the correct answer will be halved, e.g. for $A_i = \{S, X\}$ we score 0.33.

Regarding the second task, we think it reasonable to have a task where systems may offer only a single 'best guess' response, but argue that the *mode* metric used has two key failings: it is too *brittle* in being applicable only to items that have a mode answer, and it *loses information* valuable to system ranking, in assigning no credit to a response that

[1] We have here somewhat notationally restated the ELS07 metrics, whilst preserving their essential content, for reasons of notational consistency with the new metrics that are to be proposed.

might be good but not optimal. We propose instead the $best_1$ metric above, which assigns score 1 to a best guess answer whose count is at the $maxfreq_i$ value, but applies to all test items irrespective of whether or not they have a unique mode. For answers having lesser counts, proportionately less credit is assigned. Note that this metric is equivalent to the new *best* metric shown beside it, for the case where $|A_i| = 1$.

For assessing overall performance, we suggest just taking the average of scores across *all* test items (M&N's Recall measure). M&N's Precision metric is presumably intended to favour a system that can tell whether it does or does not have any reasonable answers to return. However, the ability to draw a boundary between good vs. poor candidates will be reflected widely in a system's performance and captured elsewhere (not least by the coverage metrics discussed later) and so, we believe, does not need to be separately assessed in this way. Furthermore, the fact that a system does not return any answers may have other causes, e.g. that its lexical resources have failed to yield *any* substitution candidates for a target word.

5. Measures of Coverage

A third task of ELS07 assesses the ability of systems to field a wider set of good substitution candidates for a target, rather than just a 'best' candidate. This 'out of ten' (oot) task allows systems to offer a set A_i of *up to 10* guesses per item i, which is scored as below. Since the score is *not* divided by the answer set size $|A_i|$, no benefit derives from offering less than 10 candidates.[2]

$$oot(i) = \frac{\sum_{a \in A_i} freq_i(a)}{|H_i|^i}$$

When systems are asked to field a broader set of candidates, we suggest that evaluation should assess if the response set is *good* in containing as many correct answers as possible, whilst containing as few incorrect answers as possible. In general, systems will tackle this problem by combining a means of ranking candidates (drawn from some lexical resource) with a means of drawing a boundary between good and bad candidates, e.g. threshold setting. Since the *oot* metric does not penalise incorrect answers, it does not encourage systems to develop a means of distinguishing good and bad answers, even though this is important to their ultimate practical utility.

The view of a 'good' answer set described above suggests a comparison of A_i to H_i using versions of 'recall' and 'precision' metrics, that incorporate the 'weighting' of human answers via $freq_i$. For purposes of comparison, let us begin by noting the obvious definitions for recall and precision metrics *without* count-weighting (which are *not* our proposed metrics):

$$R(i) = \frac{|H_i \cap A_i|}{|H_i|}$$

$$P(i) = \frac{|H_i \cap A_i|}{|A_i|}$$

Our definitions for these metrics *do* include count-weighting, and are given below. Note that the numerator of our recall definition is $|A_i|^i$ not $|H_i \cap A_i|^i$ as $|A_i|^i = |H_i \cap A_i|^i$, i.e. because $freq_i$ assigns 0 to any term not in H_i, a fact which also affects the numerator of our P definition. Regarding the latter's denominator, merely dividing by $|A_i|^i$ would not penalise incorrect terms, again because $freq_i(a) = 0$ for any $a \notin H_i$. Hence, this penalty is imposed directly, by adding the component $k|A_i - H_i|$, where $|A_i - H_i|$ is the number of incorrect answers, and k some weighting that is applied to them. This penalty weighting might be $k = 1$ in the simplest case, but other weightings are possible, e.g. setting k to the *average* count weight of terms in the answer set. For assessing overall performance, we can average P and R across all test items, and combine them to an overall F-score as the harmonic mean of these averages (i.e. $F = 2PR/(P + R)$). Note that, although stated somewhat differently, our weighted R metric is in fact equivalent to the *oot* definition given above.

$$R(i) = \frac{|A_i|^i}{|H_i|^i}$$

$$P(i) = \frac{|A_i|^i}{|A_i|^i + k|A_i - H_i|}$$

$$F(i) = \frac{2P(i)R(i)}{P(i) + R(i)}$$

For example, with H_i = {*G:3,M:3,S:2,J:1,Ch:1*}, the perfect response set $A_i = \{G, M, S, J, Ch\}$ gives P and R scores of 1. The response $A_i = \{G, M, S, J, Ch, X, Y, Z, V, W\}$, containing all correct answers plus 5 incorrect ones, gets $R = 1$, but only $P = 0.66$ (assuming $k = 1$, giving $10/(10 + 5)$). The response $A_i = \{G, S, J, X, Y\}$, with 3 out of 5 correct answers, plus 2 incorrect ones, gets $R = 0.6$ $(6/10)$ and $P = 0.75$ $(6/6 + 2)$)

6. Measures of Ranking

As noted, systems will in general tackle the coverage task by combining a method to rank guesses, with a means for selecting some *top N* to return. The possibility arises that we may separately assess the performance of systems at the first of these stages, i.e. their ability to generate candidates and rank them effectively. In terms of our running example (with H_i = {*G:3,M:3,S:2,J:1,Ch:1*}, we might get responses A_i = {*G,M,X,Y*} and A'_i = {*X,Y,G,M*}, which are *equivalent as sets*, but where A_i is better than A'_i *if we regard them as ranked lists*.[3] Clearly, M&N's *oot* metric does not distinguish these two responses in terms of their merit at ranking (and is not intended to).

[2] We do not consider here a related ELS07 task which assesses whether the *mode* answer m_i for an item is found within an answer set of up to 10 guesses. We do not favour the use of this metric for reasons parallel to those discussed for the *mode* metric of the previous section, i.e. *brittleness* and *information loss*.

[3] The fact that answer sets are really ranked lists is implicit in the fact that, for the ELS07 'best' tasks, a system's *first* candidate is treated as its 'best guess' for mode-based scoring.

We propose a metric to evaluate ranked answer lists, which (like *oot*) allows systems to offer up to 10 guesses, with no benefit for offering fewer. Our approach is to assess, at each rank from 1 to 10, what (count-weighted) proportion of *optimal* performance an answer lists achieves, as compared to the gold standard answer set. Thus, at rank 1, we consider the first answer and compare its frequency value to that of the best human answer. At rank 2, we compare the top *two* answers to the two best human answers, and so on. The performance at the 10 ranks is then averaged to give an overall score. We shall explain the approach in terms of our running example. To compute the optimal performance at each rank, we first extract the frequency counts for the terms in the human answer set, sort them into *descending order*, and then map them into a table with columns for the 10 ranks (padding any unfilled cells with 0s), and then compute a *cumulative frequency* value from left-to-right, as in the following table for our example.

$H_i = \{$*G:3,M:3,S:2,J:1,Ch:1*$\} \mapsto$

rank	1	2	3	4	5	6	7	8	9	10
freq	3	3	2	1	1	0	0	0	0	0
cum.freq	3	6	8	9	10	10	10	10	10	10

For an answer list such as $A_i = (S, Ch, M, J, G, X, Y, Z, V)$, we construct a corresponding table, mapping frequency counts into the table for the answer terms *in their given order*, and again compute cumulative values across these counts:

$A_i = (S, Ch, M, J, G, X, Y, Z, V) \mapsto$

rank	1	2	3	4	5	6	7	8	9	10
freq	2	1	3	1	3	0	0	0	0	0
cum.freq	2	3	6	7	10	10	10	10	10	10

The score at each rank is the result of dividing the *cum.freq* value from the human answer table by that from the system answer table, and then these ten values are averaged, which for this example is:

$$\begin{aligned} rank(i) &= (\tfrac{2}{3}+\tfrac{3}{6}+\tfrac{6}{8}+\tfrac{7}{9}+\tfrac{10}{10}+\tfrac{10}{10}+\tfrac{10}{10}+\tfrac{10}{10}+\tfrac{10}{10}+\tfrac{10}{10})/10 \\ &= 0.87 \end{aligned}$$

In A_i, the correct answers appear above all incorrect answers, but are sub-optimally ordered w.r.t. each other. In our next example A'_i, the same answers appear, but with a poorer ranking:

$A'_i = (X, Y, S, Ch, M, Z, J, V, G) \mapsto$

rank	1	2	3	4	5	6	7	8	9	10
freq	0	0	2	1	3	0	1	0	3	0
cum.freq	0	0	2	3	6	6	7	7	10	10

This response yields the score:

$$\begin{aligned} rank(i) &= (\tfrac{0}{3}+\tfrac{0}{6}+\tfrac{2}{8}+\tfrac{3}{9}+\tfrac{6}{10}+\tfrac{6}{10}+\tfrac{7}{10}+\tfrac{7}{10}+\tfrac{10}{10}+\tfrac{10}{10})/10 \\ &= 0.52 \end{aligned}$$

The even more poorly ordered answer list $(X, Y, Z, V, W, G, M, S, J, Ch)$, for example, gives a score of 0.36. Note that the score is also reduced if any human answers are missing from the answer list, e.g. for $(X, Y, Z, V, W, G, S, J, Ch)$, we get score 0.28. An optimal answer list, such as $(M, G, S, Ch, J, X, Y, Z, V, W)$, which ideally orders the correct answers w.r.t. each other, and above all incorrect answers, scores 1.

We are not aware of any other evaluation metrics that fulfil the needs of ranking evaluation for this task. Thus, our metric favours the finding of *all* correct answers, in a way that accommodates answer weighting (in our case by count), assigning a score of 1 where (and only where) all correct answers are found and optimally ordered, with the lesser scores (ultimately down to 0) for responses which are less well-ordered and/or have missing answers. An example of another metric that favours high ranking of correct answers is the *mean reciprocal rank* (MRR) metric used, for example, in work on Automatic Question Answering. Note that MRR is not a viable alternative for the current purpose, as MRR looks only at the rank of the *first* correct answer in the response list, and so will give the same score for alternative responses that have the same first correct answer appearing at the same rank position, but which differ widely in whether other correct answers are well-ranked or not, or even returned at all.

7. Discussion of the New Measures

In the above sections, we have proposed metrics for three groups of measures, i.e. measures for coverage, measures for best answers, and a measure of the quality of candidate ranking. Given this multiplicity of measures, and hence also of tasks, it is perhaps worthwhile to make clear our view of the relative importance of these alternative tasks and measures.

We would argue that the *core* task for lexical substitution should be that addressed under "coverage", i.e. given a human/gold standard set of substitutes for a given test example, a good system is one that can return as many of these correct substitutes as possible, whilst returning as few false additional terms as possible. This reduces quite intuitively to a comparison of the system answer set to the human set in terms of precision and recall, even if the need to accommodate count weighting makes our final statement of the metrics appear slightly more obscure than this simple intuition might suggest. Performance on this task requires that systems can both field and rank promising candidates, and also find a means of discriminating between the candidates that are likely to be correct (and should be returned) and those that are likely to be false. It is regarding this latter aspect of performance, drawing a boundary between good and bad candidates, that the existing metrics most obviously fail to give adequate guidance to research effort.

The 'best guess' task is a lesser indicator of the overall quality of lexical substitution systems, but its results are of interest, since they are suggestive of the likely utility of incorporating a lexical substitution system as a component within a broader practical application, i.e. where a best guess is often what will be needed.

The ranking measure is somewhat different to both of the above groups of measures, in that it does not realise any straightforward intuition as to what constitutes good performance at lexical substitution, and what it evaluates is at least one step removed from such actual performance. However, the ability to successfully rank candidates is so critical to the means by which most lexical substitution systems work that the availability of a measure of good ranking will, we believe, contribute to the development of better systems in future work.

Given that we have argued that the *core* task for lexical substitution is that addressed by the *coverage* metrics, we would ideally here like to provide some assessment of the measures when applied the outputs of the systems participating in ELS07, as compared to use of the original *oot* measure. However, this is not possible, because the ELS07 tasks do not require systems to do what our measures require (and which we believe *should* be required), namely that systems should decide a boundary between good and bad candidate answer terms. Hence, scoring systems in these terms would be inappropriate, as the systems were not developed to fulfil this requirement. In the next section, we report some preliminary results that compare use of our new ranking measure to the *oot* measure, when used as a basis for comparing the performance of alternative systems in generating *oot* type answer sets.

8. Applying the Ranking Metric

In Jabbari (2010), we have evaluated three systems using the *oot* (out-of-ten) measure — one of the standard measures of the English Lexical Substitution Task. These three systems (which were created *after* ELS07) are as follows:

1. bow: a system that ranks the subsitution candidates using a bag-of-words model
2. lm: a system that ranks the candidates using a language model
3. cmlc: a system which uses a model that combines both bow and lm systems (short for *combined model of lexical context*)

According to the out-of-ten measure, the system using the combined model (cmlc) outperforms either of its submodels. Table 1 shows the recall figures, for each part-of-speech category of the target item.

	part-of-speech			
model	nouns	adj	verb	adv
bow	0.343	0.334	0.205	0.461
lm	0.371	0.442	0.252	0.561
cmlc	0.405	0.447	0.319	0.533

Table 1: Out-of-ten recall scores for the three systems (subdivided by *pos* of target item)

The three systems use the same component to generate the initial set of candidate substitutions (which are drawn from lexical resource, such as WordNet). The systems then use their model to *rank* the candidates, and discard any candidates that are not within the top ten. The *oot* measure does not consider the relative ranking of correct answers within an answer set, i.e. it will assign the same score to answer sets that contain the same correct answers irrespective of ordering. Hence, the performance differences shown in Table 1 will come from examples where the different ranking behaviour of the systems has resulted in a different number of correct answers making it into the top ten. All examples where the same correct answers have made it into the top ten will be equivalently scored, irrespective of the quality of ranking within the answer set. Where the candidate generation phase has yielded no more than ten candidates in the first place, identical scores will be assigned to the systems by necessity.

Applying the new ranking measure to the outputs of the three systems gives the results shown in Table 2. Although these results show a similar trend to those in Table 1, the scores assigned in this case *will* differentiate the behaviour of systems for test examples where the same correct answers make it into the top ten, but where there is differently effective ranking of those candidates. Overall, however, it is not surprising that the results in the two tables show a similar trend: it is still differences in the quality of ranking achieved by the three systems that will drive the differences in *oot* scores, even if that metric is less effective in assessing that quality of ranking than the new measure.

	part-of-speech			
model	nouns	adj	verb	adv
bow	0.239	0.219	0.128	0.312
lm	0.275	0.325	0.176	0.448
cmlc	0.303	0.343	0.186	0.413

Table 2: Scores from the new ranking measure for the three systems (subdivided by *pos* of target item)

9. Conclusion

We have proposed some new measures for evaluating the performance of lexical substitution systems, which address problems identified for the metrics previously used. We believe that these new measures express clear intuitions of what constitutes good performance in lexical substitution (or for the ranking measure, a key aspect of good performance). We hope that these measures will better serve to guide the development of lexical substitution systems in future work.

References

D. McCarthy, 2002. Lexical Substitution as a Task for WSD Evaluation. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, USA.

D. McCarthy and R. Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.

S. Jabbari. 2010. *A Statistical Model of Lexical Context*, PhD Thesis, University of Sheffield.