A Comparative Study of Conversion Aided Methods for WordNet Sentence Textual Similarity

Muhidin Mohamed
EECE University of Birmingham,
Edgbaston, Birmingham, UK
Mam256@bham.ac.uk

M. Oussalah
EECE University of Birmingham,
Edgbaston, Birmingham, UK
M.Oussalah@bham.ac.uk

Abstract

In this paper, we present a comparison of three methods for taxonomic-based sentence semantic relatedness, aided with word parts of speech (PoS) conversion. We use WordNet ontology for determining word level semantic similarity while augmenting WordNet with two other lexicographical databases; namely Categorial Variation Database (CatVar) and Morphosemantic Database in assisting the word category conversion. Using a human annotated benchmark data set, all the three approaches achieved a high positive correlation reaching up to (r = 0.881647) with comparison to human ratings and two other baselines evaluated on the same benchmark data set.

1 Introduction

Sentence textual similarity is a crucial and a prerequisite subtask for many text processing and NLP tasks including text summarization, document classification, text clustering, topic detection, automatic question answering, automatic text scoring, plagiarism detection, machine translation, conversational agents among others (Ali, Ghosh, & Al-Mamun, 2009; Gomaa & Fahmy, 2013; Haque, Naskar, Way, Costa-Jussà, & Banchs, 2010; K. O'Shea, 2012; Osman, Salim, Binwahlan, Alteeb, & Abuobieda, 2012). There are two predominant approaches for sentence similarity: corpus-based and knowledge-based. The former utilises information exclusively derived from large corpora including word frequency of occurrence, and latent semantic analysis, to infer semantic similarity. On the other hand, Knowledge-based measures employ the intrinsic structure of a semantic network including its hierarchy to derive the semantic similarity. One of the commonly used knowledge networks for semantic similarity is WordNet. It is a hierarchical lexical database for English developed at Princeton University (Miller, 1995). The state of the art WordNet sentence similarity is harvested from pairing the constituent words of the two compared sentences. This is based on the intuition that similar sentences in meaning will indeed comprise semantically related words. However, these pairings only handle nouns and verbs as other part-of-speech (PoS) attributes are not accounted for in WordNet taxonomy. Taxonomic similarity is a conceptual relatedness derived from hyponymy/hypernymy relations of lexical ontologies. In this study, we use a group of WordNet semantic relations, e.g. synonymy, hyponymy, for similarity determination and for the approximation of noun equivalents of other PoS words.

In implementing the conversion aided methods, we adapted a publicly available package (Pedersen, Patwardhan, & Michelizzi, 2004) to measure word level similarity. We computed word similarities from word senses using Wu and Palmer's measure (Wu & Palmer, 1994) as given in expression 1.

$$Sim(w_1, w_2) = \max_{c_1 \in senses(w_1), c_2 \in senses(w_2)} \left(\frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)}\right) \quad (1)$$

Where $\boldsymbol{lcs}(\boldsymbol{c_i}, \boldsymbol{c_j})$ (lowest common subsumer) stands for the synset subsuming concepts $\boldsymbol{c_i}$ and $\boldsymbol{c_j}$ while depth ($\boldsymbol{c_i}$) indicates the number of nodes from concept $\boldsymbol{c_i}$ to the root node of the hierarchy.

Next, the above word-to-word semantic similarity is extended to sentence-to-sentence semantic similarity, say $\boldsymbol{S_i}$ and $\boldsymbol{S_j}$ using (Malik, Subramaniam, & Kaushik, 2007) like approach, where pairs of the same PoS tokens from the two sentences are evaluated.

$$Sim(S_i, S_j) = \frac{1}{2}\left[\frac{\sum_{w_1 \in S_i} \max_{w_2 \in S_j} Sim(w_1, w_2)}{|S_i|} + \frac{\sum_{w_1 \in S_j} \max_{w_2 \in S_i} Sim(w_1, w_2)}{|S_j|}\right],\ PoS(w_1) = PoS(w_2) \quad (2)$$

In (2), $Sim(w_1, w_2)$ stands for word level similarity measure in (1).

Proceedings of the AHA! Workshop on Information Discovery in Text, pages 37–42,
Dublin, Ireland, August 23 2014.

Nevertheless, for common natural language texts, it remains biased if only verbs and nouns are used to measure semantic relatedness ignoring other word categories such as adjectives, adverbs and named entities. To elaborate that, consider the following pair of semantically identical sentences with different word surface forms and classes.

S_1: He stated that the construction of the house is complete.

S_2: He said in a statement that the house is completely constructed.

Initial preprocessing tasks including tokenization, normalization, and stop-words removal reduce sentences to their semantic words with S_1 yielding (*state, construction, house, complete*) and (*statement, house, completely, construct*) for S_2. To optimize the semantic similarity of the two sentences, their scores from the word pairings need to be maximized regardless their associated part of speech. For S_1 and S_2, this is only achievable when words are paired as *(statement, state), (house, house), (construction, construct) and (complete, completely).* However, using quantification (2) yields a $Sim(S_1,S_2)$ score of 0.543. This is justifiable as computing the similarity of the above first, third and fourth pairs, is out of reach using conventional WordNet measures due to each word pair falling in different PoS. To handle the above limitation, the idea advocated in this paper is to turn all non-noun PoS terms into corresponding noun expressions in order to enhance the pairing tasks.

The rationale behind the migration to noun category instead of other PoS categories relies on the inherent well elaborated properties of noun category in the taxonomical hierarchy, e.g., number of nouns is much more important than other attributes in most lexical databases, which increases the chance of finding noun-counterpart; WordNet 3 has a depth of 20 for nouns and 14 for verbs, which allows for much more elaborated hyponym/hypernym relations for instance. It is also the case that words in the lower layers of the deeper hierarchical taxonomy have more specific concepts which consequently yield a high semantic similarity (Li, McLean, Bandar, O'shea, & Crockett, 2006). This is again supported by the argument presented in (Bawakid & Oussalah, 2010).

The reasons stated above and WordNet limitation of parts of speech boundary motived the current study of word PoS conversion in an attempt to improve the measurement of taxonomic-based short text semantic similarity. In this respect, transforming all other primary word categories[1] of the previous example to nouns using CatVar (Habash & Dorr, 2003) aided conversion has raised the similarity from 0.543 to 0.86. Since the two sentences of the previous example are intuitively highly semantically related, the noun-conversion brings the sentence similarity closer to human judgement. This again highlights the importance of word PoS conversion to move freely beyond the barrier of PoS restriction. This paper aims to investigate three distinct word conversion schemes. Although, all the three approaches use WordNet for measuring the term level similarity, each stands on a distinct external lexical resource in converting word's category; namely, WordNet 3.0, the Categorial Variation Database (CatVar), and the Morphosemantic Database (Fellbaum, Osherson, & Clark, 2009).

CatVar is a lexical database containing word categorial variations for English lexemes sharing a common stem, e.g. $research_V$, $researcher_N$, $researchable_{AJ}$,. Likewise, Morphosematic Database is a WordNet-related linguistic resource that links morphologically related nouns and verbs in WordNet. Both aforementioned databases are solely utilized to aid the PoS conversion of three primary word classes to nouns. Contributions of this paper are two folded. First, we improved traditional WordNet sentence similarity by converting poorly or non-hierarchized word categories (e.g. verbs, adverbs and adjectives) to a class with well-structured and deep taxonomy (nouns) using WordNet relations, CatVar and Morphosemantic databases. Second, we have performed a comparison among the three PoS conversion techniques to discover the most appropriate supplementary database to WordNet.

2 Word Parts of Speech Conversion Methods

The two conversion methods aided with CatVar and Morphosemantics were performed by looking up the word to be converted from the corresponding database and replacing it with target category word. For example to convert the verb ***arouse***, a simple look-up database matching yields **arousal** as an equivalent noun to ***arouse*** in both databases (***arouse ⇒ arousal***). However, WordNet aided conversion cannot be accomplished with a simple look up and replacement strategy due to the nature of its lexical organization that emphasises word semantics rather than their morphology. For this purpose, to con-

[1] Verbs, adjectives, adverbs

vert verb category into noun category, we designed a systematic four level conversion procedure starting with a verb surface form where the verb itself is checked for having noun form. If the latter fails, the second level investigates the synonyms of the verb senses, where each synset is checked whether a noun-form exists. If a noun member is found a replacement is issued, otherwise, another subsequent reasoning is applied. The third level differs from the previous two in that it goes down one level to the child node in the WordNet taxonomy following the hyponymy relation in which case the verb is converted by replacing it by the first encountered node containing the target category. Last but not least, the fourth level is based on moving one parent node up the taxonomy through the hypernymy relation where the first obtained noun is used as an approximate noun counterpart. Fig. 1 illustrates the WordNet aided conversion levels indicating an example of word conversion achieved at each level (see underneath the figure). On the other hand, derivation rules in WorldNet allow us to convert advert/adjective categories into their noun counterparts if available.

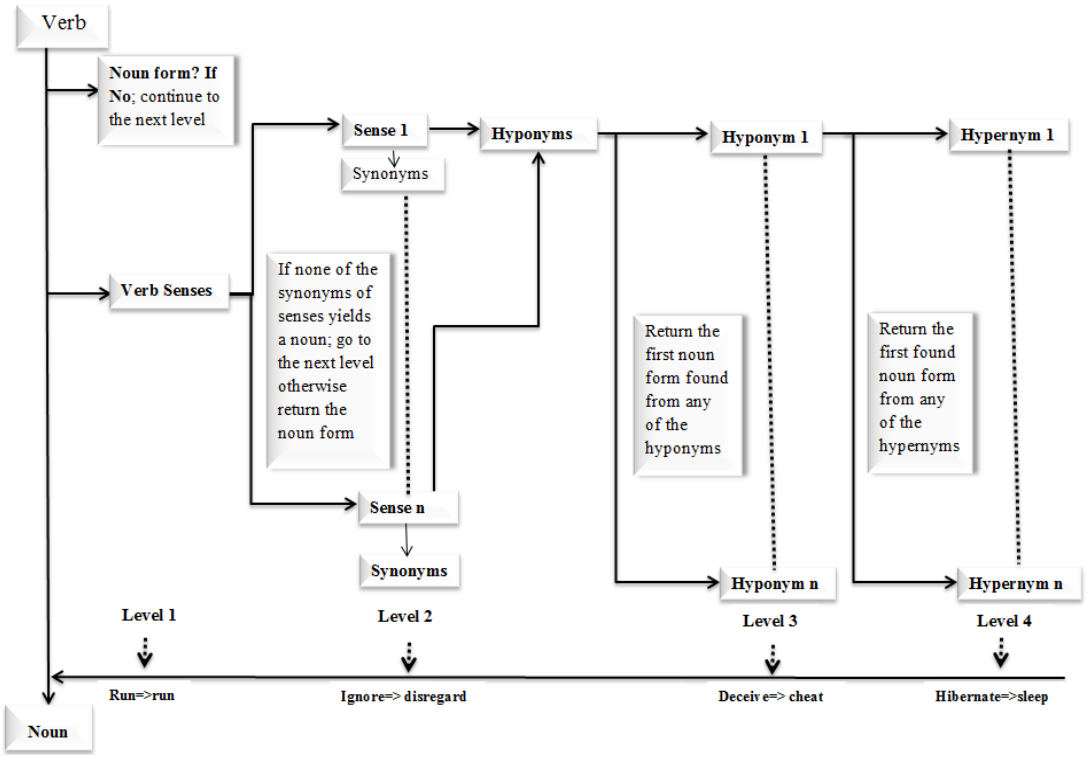


Fig. 1: The 4-level WordNet Aided Parts of Speech (PoS) Conversion

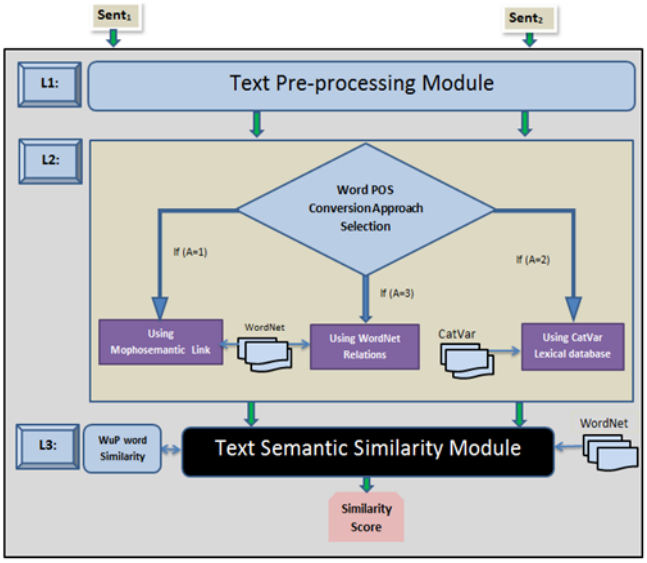
3 Implementation and Experiments

Figure 2 (a) depicts our layered implementation of the multiple conversion aided sentence semantic similarity. For every two sentences, we determine how closely the two are semantically related using scores between 1 and 0 with 1 indicating identical texts. Fig 1 (b) highlights a functional algorithm that summarizes the word category conversion process. The ***convert(w)*** function in the same algorithm performs the parts of speech conversion from the selected database depending on the active approach (A in Fig.2 (a)). All text pre-processing tasks including tokenization, parts of speech tagging, and stop words removal are implemented in layer 1. The second layer houses the three main word category conversion approaches in discussion. In each experimental run, only one approach is used depending on the choice of internally hardcoded system logic. The generated output from layer 2 is sentence text vectors having the same part of speech. These vectors are then fed into the Text Semantic Similarity Module to measure the similarity score using Wu and Palmer measure (Wu & Palmer, 1994) for word level similarity and WordNet taxonomy as an information source according to equations (1-2).

3.1 Data set

We conducted system experiments on a pilot benchmark data set created for measuring short-text semantic similarity (O'Shea, Bandar, Crockett, & McLean, 2008). It contains 65 sentence pairs with hu-

man similarity judgements assigned to each pair. During this data set creation, 32 graduate native speakers were assigned to score the degree of similarity using scores from 0 to 4 and following a guideline of semantic anchor (Charles, 2000) included in Table 2. To make the semantic anchors comply with our system generated scores (0 to 1), the scale points have been linearly transformed as indicated in the second column of the same table.



(a)

Algorithm1: Word Parts Of Speech Conversion

Input: sentence with different word classes;
Output: sentence with same word class(CWS);

```
S ← Sentence;
CWS← { };
C ← { }
W ← tokenize(S)
for each w_i ϵ W do
    If w_i = inflected then
      w_i ← baseform(w_i)
    endif
    If (w_i not in targetcategory)
      cw← convert (w_i)
    endif
    CWS ← CWS {cw}
 end for
 return CWS
```

(b)

Fig. 2: (a) Word POS conversion aided semantic similarity system; (b) Word parts of speech conversion Algorithm

Table 1: Semantic Anchors

Scale Points	Transformed Scale Points*	Semantic Anchor
0.0	0.0	The sentences are unrelated in meaning
1.0	0.25	The sentences are vaguely similar in meaning
2.0	0.5	The sentences are very much a like in meaning
3.0	0.75	The sentences are strongly related in meaning
4.0	1.0	The sentences are identical in meaning

3.2 Results and Evaluation

Our evaluation for all three conversion assisted systems is centered around the human judgements. Human ratings reflect the extent to which every two sentences are semantically related from the human perception. A comparison of our conversion aided methods (*TW, CwW, CwM, CwC*) and the findings of two baseline methods (*STASIS, LSA*) is presented in Table 2. The notations *TW, CwW, CwM, CwC* stand for, traditional WordNet, conversion with WordNet, conversion with Morphosemantics and conversion with CatVar respectively. We selected the baselines because of their fitness for purpose and their evaluation on the same benchmark data. *STASIS*, thoroughly described in (Li, et al., 2006), is a textual similarity measure combining taxonomy and word order information to compute the semantic relatedness for two sentences. While LSA (latent sematic analysis) (Deerwester et. al, 1990) is a corpus-based measure developed for indexing and retrieval of text documents but later adapted for tasks including sentence similarity. In LSA, texts are represented as a matrix, of high dimensional semantic vectors, which is then transformed using Singular Value Decomposition (SVD); namely, $A = TSD^T$, where A is a term-document matrix, S is the diagonal matrix of the Singular Value Decomposition, while T and D are left and right singular vectors with orthogonal columns. As pointed out, the results obtained in (J. O'Shea, Bandar, Crockett, & McLean, 2008) have been compared to our experimental results. Due to the space limitation, results of only 10 randomly selected sentence pairs from the benchmark data set are listed in Table 2 with the second column being the human ratings.

Table 2. Human, STASIS, LSA, TW, CwW, CwM and CwC similarity scores for 10 sentence pairs

Sentence Pair	Human	STASIS	LSA	TW	CwW	CwM	CwC
1.cord:smile	0.01	0.329	0.51	0.362	0.49	0.57	0.667
9.asylum:fruit	0.005	0.209	0.505	0.430	0.43	0.506	0.522
17.coast:forest	0.063	0.356	0.575	0.616	0.738	0.80	0.791
29.bird:woodland	0.013	0.335	0.505	0.465	0.583	0.665	0.665
33.hill:woodland	0.145	0.59	0.81	0.826	0.826	0.826	0.826
57.forest:woodland	0.628	0.7	0.75	0.709	0.804	0.867	0.867
58.implement:tool	0.59	0.753	0.83	0.781	0.744	0.905	0.885
59.cock:rooster	0.863	1	0.985	1	1	1	1
61.cushion:pillow	0.523	0.662	0.63	0.636	0.637	0.723	0.842
65.gem: jewel	0.653	0.831	0.86	0.717	0.745	0.793	0.778

To measure the strength of the linear association measured in terms of the correlation coefficients *r*, between the score of each conversion aided method and the human judgements, are computed and presented in Table 3 using equation 3 where $\boldsymbol{n}$ is the number of sentence pairs while $\boldsymbol{m_i}$ and $\boldsymbol{h_i}$ represent machine and human scores, respectively, for the i[th] pair.

$$r = \frac{\mathrm{n}\sum_i h_i m_i - \sum_i h_i \sum_i m_i}{\sqrt{\left(n\sum_i h_i^2 - (\sum_i h_i)^2\right)}\sqrt{\left(n\sum_i m_i^2 - (\sum_i m_i)^2\right)}} \qquad (3)$$

The performances of all the three methods gradually excel with an increasing shared semantic strength between the sentence pairs. However, for the less related sentence pairs, it is evident that the human perception of similarity is more strict than the loose definition of similarity based on lexical concepts and hierarchical taxonomy. Table 2 shows that all the three conversion aided methods considerably improve semantic scores over the traditional WordNet (TW). Out of the three schemes, CatVar-aided conversion establishes the highest semantic correlation between the sentence pairs corroborating the hypothesis that CatVar can be used as a supplementary resource to WordNet. Overall, scores of correlation coefficients of the developed approaches with the baseline methods; STASIS and LSA and human judgements indicate that CatVar-based conversion provides best performance. On the other hand, the correlation coefficients (expression 3) between our conversions aided schemes and the two compared benchmark methods along with the human judgements, summarized in Table 3, shows that statistically speaking, latent semantic analysis (LSA) provides the best consistency with WordNet-based similarity measures.

Table 3: Correlations Coefficients (r) between machine and human scores

	CwW	*CwM*	*CwC*	*STASIS*	*LSA*
Human	0.729826	0.830984	0.881647	0.816	0.838
STASIS	0.771874	0.851675	0.872939	--	0.76
LSA	0.804518	0.875024	0.822453	0.76	--

In order to visualize the effect of correlation coefficient across sentence pairs, Fig. 3 illustrates the association between the human ratings and each of the achieved results. It is evident that all the three relationships follow a positive linear trend with slightly varying but strong correlation with the human judgements and without outliers. For those sentence pairs which are either strongly related or identical in meaning, there is a high agreement between the human evaluation and machine assessment for semantic similarity. The results also confirm that CatVar aided conversion yields a strong positive correlation with the human rating.

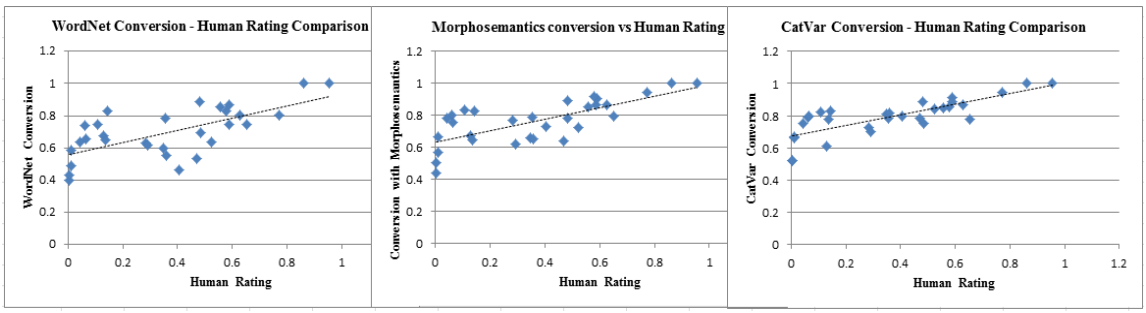


Fig. 3: Relationships between the obtained results and human judgements for the benchmark data set

4 Conclusion

To improve the accuracy of capturing semantic textual relatedness, we carried out word parts of speech conversion by augmenting two lexical databases; CatVar and Morphosemantics to traditional WordNet similarity. Our comparative analysis with human judgements and two baseline systems found that WordNet taxonomy can be supplemented with other linguistic resources, such as CatVar, to enhance the measurement of sentence semantic similarity. The findings revealed that the word parts of speech conversion captures the sematic correlation between two pieces of text in a way that brings closer to human perception. As a future work, we plan to improve the suggested conversion aided similarity measures and apply them on various large scale data set.

References

Ali, M., Ghosh, M. K., & Al-Mamun, A. (2009). *Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation.* Paper presented at the Future Computer and Communication, 2009. ICFCC 2009. International Conference on.

Bawakid, A., & Oussalah, M. (2010). *A semantic-based text classification system.* Paper presented at the Cybernetic Intelligent Systems (CIS), 2010 IEEE 9th International Conference on.

Charles, W. G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics, 21*(04), 505-524.

Deerwester et. al, S. C. (1990). Indexing by latent semantic analysis. *JASIS, 41*(6), 391-407.

Fellbaum, C., Osherson, A., & Clark, P. E. (2009). Putting semantics into WordNet's" morphosemantic" links *Human Language Technology. Challenges of the Information Society* (pp. 350-358): Springer.

Gomaa, W. H., & Fahmy, A. A. (2013). A Survey of text similarity approaches. *International Journal of Computer Applications, 68*(13), 13-18.

Habash, N., & Dorr, B. (2003). *A categorial variation database for English.* Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.

Haque, R., Naskar, S. K., Way, A., Costa-Jussà, M. R., & Banchs, R. E. (2010). *Sentence similarity-based source context modelling in pbsmt.* Paper presented at the Asian Language Processing (IALP), 2010 International Conference on.

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on, 18*(8), 1138-1150.

Malik, R., Subramaniam, L. V., & Kaushik, S. (2007). *Automatically Selecting Answer Templates to Respond to Customer Emails.* Paper presented at the IJCAI.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

O'Shea, J., Bandar, Z., Crockett, K., & McLean, D. (2008). Pilot short text semantic similarity benchmark data set: Full listing and description. *Computing*.

O'Shea, J., Bandar, Z., Crockett, K., & McLean, D. (2008). A comparative study of two short text semantic similarity measures *Agent and Multi-Agent Systems: Technologies and Applications* (pp. 172-181): Springer.

O'Shea, K. (2012). An approach to conversational agent design using semantic sentence similarity. *Applied Intelligence, 37*(4), 558-568.

Osman, A. H., Salim, N., Binwahlan, M. S., Alteeb, R., & Abuobieda, A. (2012). An improved plagiarism detection scheme based on semantic role labeling. *Applied Soft Computing, 12*(5), 1493-1502.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). *WordNet:: Similarity: measuring the relatedness of concepts.* Paper presented at the Demonstration Papers at HLT-NAACL 2004.

Wu, Z., & Palmer, M. (1994). *Verbs semantics and lexical selection.* Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics.