Proceedings of the Australasian Language Technology Workshop 2007, pages 117-124

An Empirical Investigation into Grammatically Constrained Contexts in Predicting Distributional Similarity

Dongqiang Yang | David M.W. Powers
School of Informatics and Engineering
Flinders University of South Australia
Po Box 2100, Adelaide 5001, South Australia
{Dongqiang.Yang|David.Powers@flinders.edu.au}

Abstract

The utility of syntactic dependencies in computing distributional similarity has not yet been fully investigated. Most research based on syntactically conditioned co-occurrences simply ignores the salience of grammatical relations and effectively merges syntactic dependencies into one 'context'. Through calculating distributional similarity, we design two experiments to explore and evaluate the four major types of contexts that are conditioned on grammatical relations. The consistent results show that the head-modifier dependency plays an important role in predicting the semantic features of nouns and verbs, in contrast to other dependencies.

1 Introduction

The roles of grammatical relations in predicting semantic similarity via distributional similarity have not been fully analysed. Most approaches simply chained these syntactic dependencies into one unified context representation for computing distributional similarity, such as in Word Sense Disambiguation (WSD) (Yarowsky, 1993; Lin, 1997; Resnik, 1997), word sense induction (Pantel and Lin, 2002), automatic thesaurus construction (Grefenstette, 1992; Lin, 1998; Curran, 2003), finding the predominant sense (McCarthy et al., 2004), etc.

It is clear that these approaches weighed each dependency through its frequency statistics, e.g. in the weighted (Grefenstette, 1992) or mutual information based (Lin, 1998) Jaccard's coefficient. Although they proposed to replace the unordered context with the syntactically conditioned one, the linguistic specificity of grammatical relations in semantics prediction is often overlooked. Except for the extraction of syntactically conditioned contexts, they in fact make no differentiation between grammatical relations, which work analogously as computing distributional similarity with unordered contexts. Without distinguishing the linguistic specificity of grammatical relations, the advantage of using the syntactic constrained context has not yet been fully exploited when yielding statistical semantics from word distributions. Our goal is thereof to study the salience of these syntactic dependencies in regulating statistical semantics, which can improve the acquisition of semantic knowledge in the Vector Space Model (VSM).

2 Related work

Padó and Lapata (2007) attempted to investigate the role of each single type of syntactic dependency in their syntactically conditioned VSM. They assumed a direct dependency as an undirected path (with a length of 1) in the graph of syntactic dependencies. In addition to this, they experimented a predefined (oblique) weighting scheme (Keenan and Comrie, 1977) in ranking dependencies, i.e. subject to verb: 5, object to verb: 4, prepositional phrase to verb: 3, etc. The optimal VSM they derived was equipped with inversely weighting dependencies within the path length less than 3, rather than this predefined scheme.

Although they investigated a commonly adopted case of syntactic dependencies with the path length equal to 1, the mapping function for reducing data sparseness and dimensionality of their VSM, e.g. congregating any paths ending with the same word, has obscured distinguishing the dependences in predicting semantic similarity. Their work has not completely shown to what extent one single type of syntactic dependency can contribute to statical semantics.

Another similar work was conducted by Plas and Bouma (2005) in enriching Dutch EuroWordNet through clustering distributionally similar words. They investigated the major types of grammatical relationships for nouns in Dutch, and found the predicate-object relation performing best against others such as subject-predicate and adjective-noun. Hoverer, the dependencies exposed to verbs has not been explored.

The goal of our work is to explore the utility of the major types of grammatical relations in predicting semantic similarity. Accordingly, distributional similarity is computed directly from each individual syntactic set rather than on a subtractive or additive fusion. To derive German semantic verb classes with distributional grammatical relations, Schulte im Walde (2006) uses additive fusion to merge syntactic and semantic features including pure verb subcategorization frames, prepositional preferences, and selectional preferences one-by-one into a final verb representation (on the condition that the features have been thoroughly studied in verb semantics). Since the distributional features of individual dependency set has not yet been fully explored, we will not go to seeking for the prime word representation through the subtractive or additive fusion, which could be the next phase of our work.

In the following, we first describe how to give rise to word representation using syntactic dependencies. In the two 'gold-standard' datasets, we evaluate each single type of dependency straight through correlating distributional similarity with human judgements. Without the 'gold-standard' data, we then employ automatic thesaurus construction to evaluate these dependencies in lexical acquisition.

3 Syntactic dependency

Word meaning can be represented as a function of co-occurrence frequencies within different contexts, and similar words share similar contexts (Harris, 1985). In a VSM, the dimensionality of a semantic space can be syntactically conditioned (i.e. syntactic dependencies) or unconditioned (i.e. a bag of words). Different methodologies of distributional similarity under these two context settings, have been systematically surveyed, e.g. for a bag of words (Sahlgren, 2006) and for syntactic dependencies (Curran, 2003; Weeds, 2003). Moreover, the difference between the two kinds of contexts is also contrasted in a framework (Padó and Lapata, 2007), with a preliminary conclusion that the syntactically conditioned VSM outperformed the unconditioned one.

Instead of arguing the states and advantages of these context representations in applications, we focuses on the roles of major types of grammatical relations in the syntactic constrained VSM.
The major types of these relations mainly embodied either in head-modifier, i.e. adjective to noun (**AN**) and adverb or the nominal head in a prepositional phrase to verb (**RV**) or in grammatical roles of verb-object (**VO**) and subject-verb (**SV**). The premises mainly rely on the following: (1) the meaning of a noun could depend on its modifiers such as adjectives, nouns, and the nominal head in a prepositional phrase as well as the grammatical role of a noun in a sentence as a subject or object; and (2) the meaning of a verb could be determined by its direct object, subject, or the head of a prepositional phrase.

3.1 Classification and parsing

To capture these relations accurately we employ a widely used and freely available parser based on link grammar (Sleator and Temperley, 1991).

In Link Grammar each word is equipped with 'left-pointing' and/or 'right-pointing' connectors. Based on the crafted rules of the connectors in validating word usages, a link between two words can be formed in reflecting a dependency relation. Apart from these word rules, 'crossing-links' and 'connectivity' are the two global rules working on interlinks, which respectively restrict a link from starting or ending in the middle of pre-existed links and force all the words of a sentence to be traced along links.

There are in total 107 major link types in the Link Grammar parser (ver. 4.1), whereas there are also various sub-link types that specify special cases of dependencies.

Using this parser, we extracted and classified the following link types into the four main types of dependencies:

- **RV**

1. *E*: verbs and their adverb pre-modifiers
2. *EE*: adverbs and their adverb pre-modifiers
3. *MV*: verbs and their post-modifiers such as adverbs, prepositional phrase

- **AN**

1. *A*: nouns and their adjective pre-modifiers
2. *AN*: nouns and their noun pre-modifiers
3. *GN*: common nouns and their proper nouns e.g. Prime Minister Howard.
4. *M*: nouns and their various post-modifiers such as prepositional phrases, adjectives, and participles

- **SV**

1. *S*: subject-nouns/gerunds and their finite verbs. There are also some sub-link types under S, for example, *Ss*g* stands for gerunds and their predicates, and *Sp* plural nouns and their plural verbs
2. *SI*: the inversion of subjects and their verbs in questions

- **VO**

1. *O*: verbs and their direct/indirect objects
2. *OT*: verbs and their time objects
3. *P*: verbs and their complements such as adjectives and passive participles

Note that except for **RV**, we define the **AN**, **SV**, and **VO** dependencies almost identically to shallow parsers (Grefenstette, 1992; Curran, 2003), or a full parser of MINIPAR (Lin, 1998) but we retrieve them instead through the Link Grammar parser.

Given different methodologies to implementing parsing, it is hardly fair to justify a syntactic parser. Molla and Hutchinson (2003) compared the Link Grammar (LG) parser and the Conexor Functional Dependency Grammar (CFDG) parser with respect to intrinsic and extrinsic evaluations. In the intrinsic evaluation the performance of the two parsers was compared and measured in terms of the precision and recall of extracting four types of dependencies, including subject-verb, verb-object, head-modifier, and head-complement. In the extrinsic evaluation a question-answering application was used to contrast the two parsers. Although the LG parser is inferior to the CFDG parser in locating the four types of dependencies, they are not significantly different when applied in question answering. Given that our main task is to study the difference of the syntactic dependencies: **RV**, **AN**, **SV**, and **VO**, acquired with the same LG parser, in predicting semantics, it is appropriate to use the LG parser to extract these dependencies.

3.2 Matrix construction

After parsing the 100 million-word British National Corpus (BNC) and filtering out non-content words and morphology analysis, we separately extracted and clustered the relationships to construct 4 parallel raw matrixes Xs (co-occurrence sets) in terms of the 4 syntactic dependencies above (hereafter the syntactically conditioned co-occurrences, denoted as R_X: $\mathbf{RV}_X$, $\mathbf{AN}_X$, $\mathbf{SV}_X$, and $\mathbf{VO}_X$). The row vectors of R_X denoted respectively $\mathbf{Rv}_X$, $\mathbf{An}_X$, $\mathbf{Sv}_X$, and $\mathbf{Vo}_X$, whereas the column vectors of R_X are denoted as $\mathbf{rV}_X$, $\mathbf{aN}_X$, $\mathbf{sV}_X$, and $\mathbf{vO}_X$ respectively.

The four matrices treat contexts with semantic contents in the frame of the syntactic dependencies. These additional constraints yield rarer events than word co-occurrences in a bag of words. The four syntactic matrices are extremely sparse with nulls in over 95% of the cells. However, they impose more accurate or meaningful (grammatical) relationships between words providing the parser is reasonable accurate. Instead of eliminating the triples with lower frequencies, we kept all co-occurrences to avoid worsening data sparseness.

3.3 Dimensionality reduction

We first substituted the frequency of cell $X_{i,j}$—freq($X_{i,j}$) with its information form using log(freq($X_{i,j}$)+1) to retain sparsity ($0 \rightarrow 0$). It can produce "a kind of space effect" (Landauer and Dumais, 1997) that can lessen the gradient of the frequency-rank curve in Zipf's law (1965), reducing the gap between rarer events and frequent ones.

We then applied Single Value Decomposition (SVD) to smooth the matrices and reduce their dimensionalities to 250, commonly adopted in NLP or LSA (on the word by document matrix). We do not normalize the documents by document entropy as we are not dealing with whole documents but small contexts.

In effect, we map a word-by-word matrix into two word-by-concept (uncorrelated component) matrices after SVD. Consider $\mathbf{SV}_X$ a *m* by *n* matrix representing subject-verb dependencies between *m* subjects and *n* verbs. The **SV** relation can be demonstrated by either using the rows ($\mathbf{Sv}_X$ or $\{X_{i,*}\}$) of $\mathbf{SV}_X$ corresponding to nouns conditioned as subjects of verbs in sentences, or the columns ($\mathbf{sV}_X$ or $\{X_{*,j}\}$) to verbs conditioned by nouns as subjects. The cell $X_{i,j}$ shows the frequency of the *i*th subject with the *j*th verb. The *i*th row $X_{i,*}$ of $\mathbf{SV}_X$ is a profile of the *i*th subject in terms of its all verbs and the *j*th column $X_{*,j}$ of **SV**$_X$ profiles the *j*th verb versus its subjects. We represent the **SV** relation respectively using the rows (**Sv**$_X$ or $\{X_{i,*}\}$) of **SV**$_X$ corresponding to nouns conditioned as subjects of verbs in sentences (*m* by 250 after SVD), and the columns (**sV**$_X$ or $\{X_{*,j}\}$) to verbs conditioned by nouns as subjects (*n* by 250 after SVD).

With respect to the mutual effect of the dependencies on words, the distributional features of nouns mainly focus on **aN**$_X$, **An**$_X$, **vO**$_X$, and **Sv**$_X$, whereas the verbs focus on **Vo**$_X$, **rV**$_X$, and **sV**$_X$. Distributional similarity can be evaluated on these dependency sets.

Note that we also concatenated these dependency sets into one united set (denoted as **All**$_X$) respectively for nouns and verbs, which indicates the common case of combining all dependencies in computing distributional similarity. **All**$_X$ also functioned as a baseline in the following evaluations.

We consistently employed the cosine similarity of word vectors as used in LSA and commonly adopted in assessing distributional similarity. Our contribution is to explore and contrast the semantic features of different syntactic dependencies consistently with one similarity method—the *cosine*, rather than to compare different distributional similarity measures with one united syntactic structure that combines all the dependencies together. Although taking into account more similarity measures in the evaluations can solidify conclusions, this would take us beyond the scope of the work.

4 Human similarity judgement

Rubenstein and Goodenough , in an experiment of investigating distributional similarity, constructed an evaluation dataset with word pairs and their semantic similarity scores. They hired 51 college undergraduates divided into two groups to measure 65 pairs of nouns with the similarity score ranging from 0 to 4. The higher the similarity number, the more similar the nouns were in their meanings. Many researchers (cf. Budantisky and Hirst (2006) and Pedersen et al. (2004) for some popular taxonomy similarity methods) validated semantic similarity methods using the human *group* similarity judgments on the standard dataset of the 65 noun-pairs.

Another source available is provided by Yang and Powers (2006) in their verb similarity work, where 130 pairs of verbs were scored by 6 subjects with a Likert scale from 0 to 4 (from non-similar to nearly synonymous). This dataset was acquired through the analogous instruction in the 65 noun-pairs similarity judgement.

Instead of answering if two words are synonymous or not, we compare to what extent distributional similarity derived from each dependency set correlate well with the human judgements on these 65 (noun) and 130 (verb) pairs. Finkelstein et al. (2002) created another dataset—a large volume of 353 word pairs. But these pairs are not strictly rated with semantic similarity rather than with word association strength, for example there are many word associations such as *Maradonna-football* and *FBI-investigation*. Therefore, we did not attempt to include it to evaluate distributional similarity.

4.1 Results

In this task, we tested distributional similarity (the *cosine*) respectively on the 65 noun-pairs with four sub-syntactic sets: **aN**$_X$, **An**$_X$, **vO**$_X$, and **Sv**$_X$ where nouns are mainly represented, as well as on the 130 verb pairs with the three sub-syntactic sets: **Vo**$_X$, **rV**$_X$, and **sV**$_X$ where verbs as row vectors can be represented with their objects, modifiers, and subjects.

	aN$_X$	**An**$_X$	**vO**$_X$	**Sv**$_X$	**All**$_X$	*Sim*$_{WN}$
r	**0.73**	0.63	0.47	0.41	0.62	0.90
ρ	**0.72**	0.63	0.43	0.38	0.68	0.85

(a) The correlations on '65 nouns'

	rV$_X$	**Vo**$_X$	**sV**$_X$	**All**$_X$	*Sim*$_{WN}$
r	**0.59**	0.49	0.41	0.57	0.84
P	**0.51**	0.44	0.38	0.53	0.77

(b) The correlations on '130 verbs'

Table 1: The value/rank correlation (r/ρ) on the syntactically conditioned dependencies

After calculating the cosine similarity of two word vectors in each subset, we then computed Pearson's correlation (r) and Spearman's correlation (ρ) between human average scores and distributional similarity (the *cosine*) scores. The results in different sub-synsets are shown in Table 1. Note that in Table 1 we also listed the taxonomy-based similarity measures proposed by Yang and Powers (2005; 2006), shortened for *Sim*$_{WN}$ that is based on a lexical knowledge base (WordNet) and can be referred in the next section. *Sim*$_{WN}$ can be taken as the upper bands for the 2 tasks, because Yang and Powers results on both ‘130 verbs’ and ‘65 nouns’ were competitive against others popular methods coded in the WordNet similarity package (Pedersen et al., 2004).

4.2 Discussion on the noun task

Note that unless otherwise specified we ran the paired T-test at the significance level of $\alpha = 0.05$ in the following sections. As to the ‘65’ dataset in Table 1-(a), distributional similarity in **aN**$_X$ with correlations over 72% predicted more accurate semantic similarity than the other three subsets: **An**$_X$, **Sv**$_X$, and **vO**$_X$. Nonetheless, **aN**$_X$ only significantly outperformed **Sv**$_X$ and **vO**$_X$ rather than **An**$_X$. Note that **An**$_X$ was significantly better in correlating with human judgments than **vO**$_X$ but not **Sv**$_X$. Both **Sv**$_X$ and **vO**$_X$ performed on a par without significant difference. The multiple linear regression shows that the combined model with **aN**$_X$, **An**$_X$, **vO**$_X$, and **Sv**$_X$ ($r = 0.74$) was significantly better than guessing the mean ($F = 15.394$, $p < 0.001$), where **An**$_X$, **vO**$_X$, and **Sv**$_X$ contributed little to the linear combination ($p > 0.05$) and **aN**$_X$ was the only significant contributor to the model ($p < 0.001$).

In contrast to the upper band of Sim_{WN}, an approach to taxonomic similarity, distributional similarity on **aN**$_X$, **An**$_X$, **vO**$_X$, and **Sv**$_X$ both significantly underperformed Sim_{WN} in correlating with human judgements.

Table 1-(a) also contains the correlation of the baseline **All**$_X$ with human ratings in this task ($r = 0.62$). Without any fusion, distributional similarity on **aN**$_X$ correlated better with human judgments than **All**$_X$, whereas **An**$_X$ performed nearly identically with **All**$_X$.

4.3 Discussion on the verb task

As shown in Table 1-(b), the *cosine* similarity in **rV**$_X$ with the correlation of about 60% predicted relatively more accurate semantic similarity than other two subsets: **sV**$_X$ and **Vo**$_X$, but the differences in their correlations were not significant. With the multiple linear regression on **Vo**$_X$, **rV**$_X$, and **sV**$_X$, we observed that 38.4% of variations in human judgement was accounted for in the combined model ($F = 26.151$, $p < 0.001$) that strongly correlated with the observed values ($r = 0.62$). Both **rV**$_X$ and **Vo**$_X$ made a significant contribution in the model with the exception of **sV**$_X$.

As for the taxonomic similarity in Table 1-(b), distributional similarities on **Vo**$_X$, **rV**$_X$, and **sV**$_X$ were significantly inferior to Sim_{WN} in terms of correlations with human judgements on the 130 pairs.

With respect to the united dependency set, consisting of **Vo**$_X$, **rV**$_X$, and **sV**$_X$, only **rV**$_X$ performed competitively against the baseline **All**$_X$.

4.4 Frequency bias

Due to the hypothesis of distributional representations, distributional similarity of words should correlate with the common features they share (Harris, 1985). We defined and collected the Intersection Attribute Frequency (**IAF**), which indicates on average how many common attributes any two words share in each dependency set R_X. For the 65 noun pairs, **IAF** on **aN**$_X$ (65.2) was larger than it on **An**$_X$ (49.2), **vO**$_X$ (26.6)**,** and **Sv**$_X$ (20.9), which corresponded well to their orders of the correlations in Table 1-(a). For the 130 verb pairs, **IAF** on **rV**$_X$ (168.9) was greater than it on **Vo**$_X$ (139.1) and **sV**$_X$ (105.1), which tallied with the relatively higher correlation on **rV**$_X$ ($r = 0.59$) than on **Vo**$_X$ ($r = 0.49$) and **sV**$_X$ ($r = 0.41$) in Table 1-(b). This is in accordance with the intuition that the more features words share, the more similar they are, which could account for the difference between the dependencies in predicting semantic features.

5 Thesaurus Construction

Instead of comparing distributional similarity with the ‘gold-standard’ of human similarity judgement, one of the application-style evaluations on distributional similarity is to automatically produce a thesaurus entry for each target word, through which the accuracy of synonyms or near-synonyms captured can indirectly measure the capabilities of the syntactic dependencies in predicting lexical semantics.

The usual way of creating an automatic thesaurus is to extract the top *n* words in the similar word list of each target as the entries of its thesaurus, after calculating and ranking the distributional similarity between the target and all of the other words. The accuracy and coverage of thesauri inevitably depend on the size and domains of the corpora used, as well as the measures of computing distributional similarity.

Given the same distributional similarity (*cosine*) across the dependency sets, the results of thesaurus construction can test semantic constraints of grammatical relations. Instead of a normal thesaurus with a full coverage of PoS tags, we only compile the thesaurus entries of nouns and verbs that account for the major part of published thesauri.

5.1 Candidate words

	Similar words
aN$_X$	*Imprisonment term utterance penalty excommunication syllable words punishment prison prisoner phrase detention hospitalisation fisticuffs banishment verdict Minnesota meaning adjective warder*
An$_X$	*words syllable utterance clause nictation word swarthiness paragraph text homograph dis-course imprisonment nonce phrase hexagram adjective verb niacin savarin micheas*
vO$_X$	*soubise cybele sextet cristal raper stint concatenation kohlrabi tostada apprenticeship ban contrivance Guadalcanal necropolis misanthropy roulade gasworks curacy jejunum punishment*
Sv$_X$	*ratel occurrence cragsman jingoism shiism Oklahoma genuineness unimportance language gathering letting grimm chaucer accent taxation ultimatum arrogance test verticality habituation*
All$_X$	*Imprisonment utterance penalty excommu-nication punishment prison prisoner detention hospitalisation banishment Minnesota meaning contrariety phoneme consonant counter-intelligence starvation fine cathedra lifespan*

(a) The similar words to *sentence* (as a noun)

	Similar words
rV$_X$	*assault rape criticize arm slaughter abduct mortar accuse defend fire avow lash badmouth blaspheme slit singe flame kidnap persecute*
Vo$_X$	*raid criticise bomb realign outwit beleaguer guard raze bombard criticize resemble spy pulse misspend reformulate alkalinise meta-stasise placard ruck glory*
sV$_X$	*Ambush invade fraternize palpitate patrol wound pillage bomb billet shell fire liberate kidnap raid garrison accuse assault arrest slaughter outnumber*
All$_X$	*raid bomb assault criticize ambush accuse fire guard bombard patrol rape storm infiltrate wound kidnap criticise garrison alkalinise torture spy*

(b) The similar words of *attack* (as a verb)

Table 2: A sample of the distributional 'thesauri'

We select 100 nouns and 100 verbs with term frequencies of around 10,000 times in BNC. Highly frequent words are likely to be functional words and the less frequent words may not happen in the semantic sets. In fact, the average frequency of the nouns in **An**$_X$, **aN**$_X$, **Sv**$_X$, and **vO**$_X$ are respectively about 3400, 5600, 1200, and 1700, and the verbs in **rV**$_X$, **Vo**$_X$, and **sV**$_X$ 3000, 3300, and 2000, as we only extracted syntactic dependencies from BNC.

For a target word in each sub-syntactic set, we produced and ranked the top 20 words as candidates for the automatic thesaurus after computing distributional similarity of the target with all other words in each sub-syntactic set. The population of the nouns or the verbs consists of 2000 words. In Table 2, we exemplify the top 20 similar words of *sentence* (as a noun) and *attack* (as a verb).

5.2 Evaluation

It is not a trivial work to evaluate distributional thesauri in the absence of a benchmark set. After constructing a 'gold standard' dataset consisting of Roget's Thesaurus (1911), Macquarie's Thesaurus, and Webster's 7th dictionary, Grefenstette (1993) evaluated his automatic thesaurus extracted from Grolier's Encyclopaedia using distributional similarity on syntactic dependencies. If two words were located under the same topic in Roget or Macquarie, or shared two or more terms in their definitions in the dictionary, they were counted as a successful hit for synonyms or semantic-relatedness.

To improve the coverage of the 'gold standard' dataset in the experiment, Curran (2003) incorporated more thesauri: Roget Thesaurus (both the free version provided by Project Gutenberg (1911) and the modern version of Roget's II), Moby Thesaurus, The New Oxford Thesaurus of English, and The Macquarie Encyclopedic Thesaurus.

Instead of simply matching with the 'gold standard' thesauri, Lin (1998) proposed to compare the structures of his automatic thesaurus to WordNet and Roget through his taxonomic similarity approach, i.e. taking into account the order of the similar words produced from distributional similarity. Inspired by Lin's work (1998), we also defined two different similarity measures to compare the automatic thesaurus with the 'gold standard', i.e. Sim_{WN} for WordNet and Sim_{RT} for Roget. Instead of recording the similarity scores produced in Sim_{WN} and Sim_{RT} we counted the number of similar words within similarity thresholds.

- Sim_{WN}: There are numerous noun similarity methods in the WordNet similarity package of Pedersen et al. (2004). However, since the similarity method proposed by Yang and Powers (2005; 2006) was competitive and also worked on the 130 verb pairs unlike other algorithms, we employed their algorithm in the evaluation. Note that their methods were in fact based on edge-counting in the taxonomy of WordNet. In the task, we set up a shorter searching depth limit $\gamma = 4$ for nouns to identify words that are more similar, and $\gamma = 2$ for verbs. If two distributionally similar words are syn/antonym or connected with each other in the taxonomy with the shortest path length less than the depth limit, we counted them as a successful hit, i.e. semantic relatedness.

- *Sim$_{RT}$*: Roget's Thesaurus divides its hierarchy top class to the bottom topic, and stores topic-related words under one of 1000 topics. We counted it a hit if two words are situated under the same topic or the higher level of the same section, i.e. the distance between two words was no more than 2 levels.

5.3 Results

		WordNet						Roget	Total
		SA	*D1*	*D2*	*D3*	*D4*	∑		
Noun	**aN**$_X$	2.8	7.5	10.0	8.2	5.3	33.7	27.5	46.7
	An$_X$	1.5	5.5	9.6	8.6	5.3	30.6	22.3	43.4
	vO$_X$	1.6	4.5	5.9	5.1	4.1	21.2	17.9	33.0
	Sv$_X$	1.1	2.9	4.8	5.0	3.7	17.4	14.1	29.2
	All$_X$	3.0	7.3	11.2	8.7	5.6	36.1	30.1	46.9
Verb	**rV**$_X$	5.3	16.5	13.7			35.5	31.1	46.2
	Vo$_X$	4.1	13.8	13.3			31.1	26.9	43.4
	sV$_X$	2.7	9.6	12.0			24.3	24.1	37.7
	All$_X$	4.0	20.0	12.8			36.7	30.2	47.9

Table 3: The evaluation results of noun and verb thesauri.

The results of our automatic thesauri for the nouns and verbs in the sub-syntactic sets are listed in Table 3. For *Sim$_{WN}$*, *SA* denotes the accuracy on the syn/antonyms of the targets, and *DI* the accuracy on the words with exactly *I* link distance to targets (for nouns $I \leq \gamma = 4$; for verbs $I \leq \gamma = 2$); ∑ denotes the overall accuracy. For *Sim$_{RT}$*, Roget indicates the overall accuracy in Roget, and Total the overall accuracy in both WordNet and Roget.

5.4 Discussion

In Table 3 both the noun thesaurus from **aN**$_X$ and the verb thesaurus from **rV**$_X$ achieved the highest overall accuracy in WordNet, Roget, and Total. The paired-sample T-test on the accuracy of each target in each sub-syntactic set showed that (1) distributional similarity extracted significantly more similar nouns from **aN**$_X$ than other three dependency sets: **An**$_X$, **vO**$_X$, and **Sv**$_X$, and from **An**$_X$ than the other two sets: **vO**$_X$ and **Sv**$_X$; (2) there were not significant difference between **rV**$_X$ and **Vo**$_X$ in retrieving real similar verbs through distributional similarity, but both of them were significantly better than **Sv**$_X$.

The baseline **All**$_X$, incorporating more grammatical relations into one representation, i.e. **aN**$_X$, **An**$_X$, **vO**$_X$, and **Sv**$_X$ for nouns and **Vo**$_X$, **rV**$_X$, and **sV**$_X$ for verbs, retrieved more synonyms or near-synonyms in its automatic thesauri than other single dependency set. The advantage of **All**$_X$ against others is not a surprise given the syntactic dependencies it combined. However **All**$_X$ vs. **aN**$_X$ and **rV**$_X$ shows no significant discrepancy on accuracy, which also implied the strength of the head-modifier relations on dominating lexical semantics.

We further varied the threshold from 20 to 50 words increasingly with 10 words to study the effect of the size of term clusters on accuracy. We found that the results were similar, and the drop of the overall accuracy of nouns and verbs was on average 4% and not significant ($p < 0.05$).

These homogeneous results in retrieving semantically similar or related words, together with those in judging semantic similarity, indicated that the head-modifier relations strongly correlates with semantic properties for nouns and verbs.

5.5 Frequency bias

As indicated in the previous evaluation, we also collected the **IAF** statistics of 2,000 noun and 2,000 verb pairs in these dependency sets, which can signify to what extent two words share common distributional structures in each dependency set. The highest **IAF** 135.4 in **aN**$_X$ (respectively 92.4 in **An**$_X$, 35.9 in **vO**$_X$, and 28.1 in **Sv**$_X$) and 87.6 in **rV**$_X$ (53.1 and 45.8 in **Vo**$_X$ and **sV**$_X$) corresponds to the highest accuracy of each dependency set in yielding automatic thesaurus construction. These results were consistent with them in the relatively small data sets of 65 noun-pairs and 130 verb-pairs from the previous section, where **IAF** is proportional to the correlations of distributional similarity on each type of grammatical relations with human similarity judgements.

6 Conclusion

Through human similarity judgements and automatic thesaurus construction, we study the major types of syntactic dependencies in expressing semantic salience. The consistent results show that semantic features of nouns and verbs are most strongly characterised by the head-modifier relations. The distinctive linguistic features of these syntactic dependencies provide an empirical basis for how to better model word meanings. Our future work would be to fuse these features in the distributional representation of words, and tailor them for specific applications.

References

Budantisky, Alexander and Graeme Hirst (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics 32(1): 13-47.

Curran, James R. (2003). From Distributional to Semantic Similarity. Ph.D thesis

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman and Eytan Ruppin (2002). Placing Search in Context: The Concept Revisited. ACM Transactions on Information Systems 20: 116-131.

Grefenstette, Gregory (1992). Sextant: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis. In the 30th Annual Meeting of the Association for Computational Linguistics, 324-326. Newark, Delaware.

Grefenstette, Gregory (1993). Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches. In the Workshop on Acquisition of Lexical Knowledge from Text, 143-153.

Harris, Zellig (1985). Distributional Structure. In The Philosophy of Linguistics J. J. Katz, (ed). New York, Oxford University Press: 26-47.

Keenan, Edward and Bernard Comrie (1977). Noun phrase accessibility and universal grammar. Linguistic Inquiry 8: 62-100.

Landauer, Thomas K. and Susan T. Dumais (1997). A Solution to Plato's Problem: the Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review 104: 211-240.

Lin, Dekang (1997). Using Syntactic Dependency as a Local Context to Resolve Word Sense Ambiguity. In the 35th Annual Meeting of the Association for Computational Linguistics, 64-71. Madrid, Spain.

Lin, Dekang (1998). Automatic Retrieval and Clustering of Similar Words. In the 17th International Conference on Computational Linguistics, 768-774. Montreal, Quebec, Canada.

McCarthy, Diana, Rob Koeling, Julie Weeds and John Carroll (2004). Finding Predominant Senses in Untagged Text. In the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 267-287. Barcelona, Spain.

Molla, Diego and Ben Hutchinson (2003). Intrinsic versus Extrinsic Evaluations of Parsing Systems. In European Association for Computational Linguistics(EACL), workshop on Evaluation Initiatives in Natural Language Processing, 43-50. Budapest, Hungary.

Padó, Sebastian and Mirella Lapata (2007). Dependency-based construction of semantic space models. To appear in Computational Linguistics 33(2).

Pantel, Patrick and Dekang Lin (2002). Discovering Word Senses from Text. In the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 613-619. New York, NY, USA.

Pedersen, Ted, Siddharth Patwardhan and Jason Michelizzi (2004). WordNet::Similarity - Measuring the Relatedness of Concepts. In the Nineteenth National Conference on Artificial Intelligence (AAAI-04), 1024-1025. San Jose, CA.

Plas, Lonneke van der and Gosse Bouma (2005). Contexts for Finding Semantically Similar Words. In the 20th International Conference on Computational Linguistics, 173-186. Geneva, Switzerland.

Resnik, Philip (1997). Selectional Preference and Sense Disambiguation. In ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How?, 52-57. Washington, USA.

Sahlgren, Magnus (2006). The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces. Ph.D thesis

Schulte im Walde, Sabine (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. Computational Linguistics 32(2): 159-194.

Sleator, Daniel and Davy Temperley (1991). Parsing English with a Link Grammar, Carnegie Mellon University.

Weeds, Julie Elizabeth (2003). Measures and Applications of Lexical Distributional Similarity. Ph.D thesis

Yang, Dongqiang and David M.W. Powers (2005). Measuring Semantic Similarity in the Taxonomy of WordNet. In the Twenty-Eighth Australasian Computer Science Conference (ACSC2005), 315-322. Newcastle, Australia, ACS.

Yang, Dongqiang and David M.W. Powers (2006). Verb Similarity on the Taxonomy of WordNet. In the 3rd International WordNet Conference (GWC-06), 121-128. Jeju Island, Korea.

Yarowsky, David (1993). One Sense per Collocation. In ARPA Human Language Technology Workshop, 266-271. Princeton, New Jersey.

Zipf, George Kingsley (1965). Human Behavior and the Principle of Least Effort: an Introduction to Human Ecology. N.Y., Hafner Pub. Co.