Using Grammatical Relations to Automate Thesaurus Construction

Dongqiang Yang and David M.W. Powers

School of Computer Science, Engineering and Mathematics
Flinders University of South Australia
PO Box 2100, SA 5001, Adelaide
{Dongqiang.Yang, David.Powers}@flinders.edu.au

In this paper we introduce a novel method of automating thesauri using syntactically constrained distributional similarity. With respect to syntactically conditioned co-occurrences, most popular approaches to automatic thesaurus construction simply ignore the salience of grammatical relations and effectively merge them into one united 'context'. We distinguish semantic differences of each syntactic dependency and propose to generate thesauri through word overlapping across major types of grammatical relations. The encouraging results show that our proposal can build automatic thesauri with significantly higher precision than the traditional methods.

Keywords: syntactic dependency, distribution, similarity
ACM Classifications: I.2.7 (Natural Language Processing)

1. INTRODUCTION

The usual way of automatic thesaurus construction (ATC) is to extract the top *n* words in the similar word list of each seed word as its thesaurus entries, after calculating and ranking distributional similarity between the seed word and all of the other words occurring in the corpora. The attractive aspect of automatically constructing or extending lexical resources (Agirre, Ansa, Martinez and Hovy, 2001; Pantel, 2005; Pennacchiotti and Pantel, 2006) rests clearly on its time efficiency and effectiveness in contrast to the time-consuming and outdated publication of manually compiled lexicons. Its application mainly includes constructing domain-oriented thesauri for automatic keyword indexing and document classification in Information Retrieval (Grefenstette, 1992b; Sánchez and Moreno, 2005; Stamou and Christodoulakis, 2005), Question Answering (Leveling and Hartrumpf, 2005), Word Sense Disambiguation (Yarowsky, 1993; Lin, 1997; Resnik, 1997), and Word Sense Induction (Pantel and Lin, 2002).

As the ground of ATC, distributional similarity is often calculated in the high-dimensional vector space model (VSM). With respect to the *basic elements* in VSM (Lowe, 2001), the dimensionality of word space can be syntactically conditioned (i.e. grammatical relations) or unconditioned (i.e. '*a bag of words*'). Under these two context settings, different similarity methods have been widely surveyed, for example for '*a bag of words*' (Sahlgren, 2006) and for grammatical relations (Curran, 2003; Weeds, 2003). Moreover, the framework conducted by Padó and Lapata (2007) compared the difference between the two settings. They observed that the syntactically constrained VSM outperformed the unconditioned one that exclusively counts word co-occurrences

Manuscript received: 31 July 2008
Communicating Editor: Paul Watters

in a $\pm n$ window. Instead of comparing these two context representations in specific applications, we focus on how to effectively and efficiently produce similar words with syntactically conditioned co-occurrences.

2. PREVIOUS WORK

Without distinguishing the latent differences of grammatical relations in dominating word meanings in context, most approaches simply chained or clumped these syntactic dependencies into one unified context representation for computing distributional similarity such as in ATC (Hirschman, Grishman and Sager, 1975; Hindle, 1990; Grefenstette, 1992a; Lin, 1998; Curran, 2003), along with in Word Sense Disambiguation (Yarowsky, 1993; Lin, 1997; Resnik, 1997), word sense induction (Pantel and Lin, 2002), and finding the predominant sense (McCarthy, Koeling, Weeds and Carroll, 2004). These approaches improved the distributional representation of a word through a fine-grained context that can filter out the unrelated or unnecessary words produced in the traditional way of '*a bag of words*' or the unordered context, given that the parsing errors introduced are acceptable or negligible. It is clear that these approaches, based on observed events, often scaled each grammatical relation through its frequency statistics in computing distributional similarity, for example in the weighted (Grefenstette, 1992a) or mutual information based (Lin, 1998) Jaccard coefficient. Although they proposed to replace the unordered context with the syntactically conditioned one, they have partly overlooked the linguistic specificity of grammatical relations in word distribution. Accordingly, they in fact make no differentiation between them, which are analogy to computing distributional similarity with unordered context.

To explore major types of grammatical relations in deriving semantic similarity, Padó and Lapata (2007) experiment with a predefined (oblique) weighting scheme (Keenan and Comrie, 1977) in ranking dependency relationships, where the weight of subject-to-verb is 5, object-to-verb is 4, prepositional phrase-to-verb is 3, and so on. They assumed a direct dependency as an undirected path (with a length of 1) in the graph of syntactic dependencies. The optimal VSM they derived was equipped with inversely weighting dependencies within the path length less than 3, rather than this predefined scheme.

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

To derive German semantic verb classes through 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 step-by-step into a final verb representation (on the condition that the features have been thoroughly studied in verb semantics).

Instead of seeking for the prime word representation, chained through either weighting schema or the subtractive/additive fusion, we proposed to separately process each type of syntactically conditioned contexts in the course of ATC. Different from most popular ways of ATC, our proposal is to first categorize contexts in terms of grammatical relations, and then to overlap the top n similar words yielded in each type of grammatical relations to retrieve similar words. This is in contrast to averaging or weighting distributional similarity across grammatical relations, which is commonly adopted in the literature. In this way, we hypothesized that the advantage of using the syntactic constrained context could be fully exploited when deriving statistical semantics from word distributions.

3. CONTEXT INTERCHANGEABILITY OF SIMILAR WORDS

Word meaning can be regarded as a function of word distribution within different contexts in the form of co-occurrent frequencies, where similar words share similar contexts (Harris, 1985). Miller and Charles (1991) propose that word similarity depends on to what extent they are interchangeable across different context settings. The flexibility of one word or phrase substituting another indicates its extent to be synonymous providing that the alternation of meaning in discourse is acceptable. We calculated distributional similarity in different syntactic dependencies such as subject-predicate and predicate-object. Given the interchangeability of synonyms or near-synonyms in different contexts, we supposed that semantically similar words derived with distributional similarity should span at least two types of syntactically constrained contexts. In other words, once we can retrieve the thesaurus items from each dependency set, the final thesaurus comprises the intersection of the items across at least any two types of dependency sets.

3.1 Syntactic Dependency

The syntactically conditioned representation mainly rely on the following grounds: (1) the meaning of a noun depends 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 (Hirschman *et al*, 1975; Hindle, 1990); and (2) the meaning of a verb depends on its direct object, subject, or modifier such as the head of a prepositional phrase (Hirschman *et al*, 1975). These results are partly consistent with the findings in studying word association and the psychological reality of the paradigmatic relationships of WordNet (Fellbaum, 1998).

The syntactic dependencies can provide a clue for tracking down the meaning of a word in context. With the hypothesis of 'one sense per collocation' in WSD, Yarowsky (1993) observed that the direct object of a verb played a more dominant role than its subject, whereas a noun acquired more credits for disambiguation from its nominal or adjective modifiers. As an application of the distributional features of words, Resnik (1997) and Lin (1997) employed the selectional restraints in subject-verb, verb-object, head-modifier and the like to conduct sense disambiguation.

Suppose that a tuple $<w_i, r, w_j>$ describes the words: w_i and w_j, and their bi-directional dependency relation r. For example, if w_i modifies w_j through r, all such w_j with r to w_i form a context profile for w_i, likewise w_i for w_j. In the hierarchy of syntactic dependencies (Carroll, Briscoe and Sanfilippo, 1998), the major types of grammatical relationships (r) can be generally clustered into:

- **RV**: verbs with all verb-modifying adverbs and the head nouns in the prepositional phrases;
- **AN**: nouns with noun-modifiers including adjective use and pre/post-modification;
- **SV**: grammatical subjects and their predicates;
- **VO**: predicates and their objects.

3.2 Context Interchangeability

The heuristic of deriving automatic thesauri with the interchangeability of synonyms or near-synonyms in grammatical contexts (dubbed as ***any two***) can be expressed:

- Nouns: $\bigcup_{i,j}(S_i \cap S_j)$ where i and j stand for any two types of dependency sets in terms of grammatical relations: **AN**, **SV**, and **VO**.
- Verbs: $\bigcup_{i,j}(S_i \cap S_j)$ where i and j stand for any two of **RV**, **SV**, and **VO**.

where for a given word, S is the thesaurus items produced through distributional similarity in a single dependency set. Note that we also used the heuristics of ***any three*** and ***any four*** to construct

automatic thesauri, but found most target words had no distributionally similar words under these stricter conditions than ***any two***. We did not attempt to demonstrate the conditions here.

We similarly hypothesized the union of all grammatical relations from the co-occurrence matrices as a baseline (dubbed as ***any one***), which computes distributional similarity with the union of all relations and can be indicated:

- Nouns: $S_{\bigcup_i}$ where *i* is one of **AN**, **SV**, and **VO**
- Verbs: $S_{\bigcup_i}$ where *i* is one of **RV**, **SV**, and **VO**

4. SYNTACTICALLY CONSTRAINED DISTRIBUTIONAL SIMILARITY

To automate thesauri, we first employed an English syntactic parser based on Link Grammar to construct a syntactically constrained VSM. The word space consists of four major syntactic dependency sets that are widely adopted in the current research on distributional similarity. Following the reduction of dimensionality on the dependency sets, we created the latent semantic representation of words through which distributional similarity can be measured so that thesaurus items can be retrieved.

4.1 Categorizing Syntactic Dependencies

To capture grammatical relations, 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*: proper nouns and their common nouns
  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 or indirect objects

2. *OD*: verbs and their distance-complement
3. *OT*: verbs and their time objects
4. *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, 1992a; Curran, 2003), or a full parser of MINIPAR (Lin, 1998) but we retrieve them instead through the Link Grammar parser.

Consider, for example, a short sentence from British National Corpus (BNC):

'Home care Coordinator, Margaret Gillies, currently has a team of 20 volunteers from a variety of churches providing practical help to a number of clients already referred.'

The parse of this sentence with the lowest cost in the link grammar parser is shown in Figure 1, where LEFT-WALL indicates the start of the sentence. We can classify four types of grammatical relations from this parse, namely:

- **RV**: <*currently, E, has*>, <*already, E, referred*>
- **AN**: <*home, AN, care*>, <*care, GN, coordinator*>, <*volunteer, Mp, team*>, <*church, Mp, variety*>, <*practical, A, help*>, <*client, Mp, number*>, <*referred, Mv, clients*>
- **SV**: <*coordinator, Ss, has*>
- **VO**: <*has, Os, team*>, <*providing, Os, help*>

After parsing the 100 million-word BNC and filtering out non-content words and morphology analysis, we separately extracted the relationships to construct four parallel matrixes or co-occurrence sets, denoted as R_X: $\mathbf{RV}_X$, $\mathbf{AN}_X$, $\mathbf{SV}_X$, and $\mathbf{VO}_X$ in terms of the four types of syntactic dependencies above. The row vectors of R_X denoted respectively $\mathbf{Rv}_X$, $\mathbf{An}_X$, $\mathbf{Sv}_X$, and $\mathbf{Vo}_X$ for the four dependencies. Similarly, the column vectors of R_X are denoted as $\mathbf{rV}_X$, $\mathbf{aN}_X$, $\mathbf{sV}_X$, and $\mathbf{vO}_X$ respectively.

Consider $\mathbf{SV}_X$ a *m* by *n* matrix representing subject-verb dependencies between *m* subjects and *n* verbs. We illustrate the **SV** relation using the rows ($\mathbf{Sv}_X$ or $\{X_{i,*}\}$) of $\mathbf{SV}_X$ corresponding to nouns

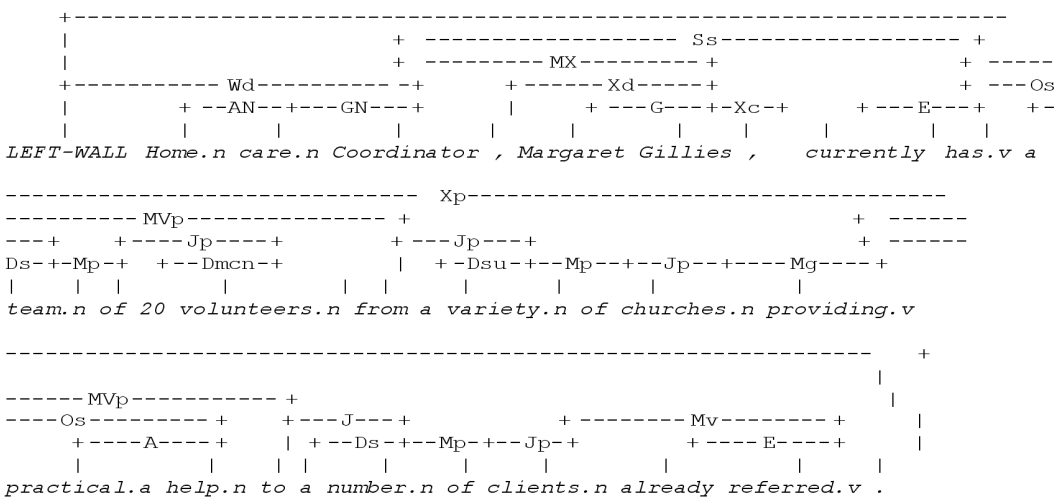


Figure 1: A complete linkage of parsing a sentence using Link Grammar

	Dim / *Freq*		1	2–10	11–20	21–30	> 31
$\mathbf{AN}_X$	48.5 by 37.6	Token	1,813.7	6,243.4	1,483.1	799.8	3,617.8
		Type	1,813.7	2,040.0	103.6	32.2	44.9
$\mathbf{RV}_X$	37.4 by 14.2	Token	863.1	2,276.4	481.4	234.9	692.2
		Type	863.1	751.9	33.8	9.5	10.9
$\mathbf{SV}_X$	32.7 by 11.3	Token	511.8	1,699.4	297.8	133.3	380.7
		Type	511.8	587.4	21.0	5.4	6.0
$\mathbf{VO}_X$	6.1 by 33.3	Token	488.5	1,811.5	475.4	266.2	1,286.9
		Type	488.5	575.1	33.1	10.7	15.6

Table 1: The statistics of the syntactically conditioned matrices (thousand)

conditioned as subjects of verbs in sentences, and 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 $\mathbf{SV}_X$ profiles the *j*th verb versus its subjects.

The parsing results are shown in Table 1, where *Dim* refer to the size of each matrix in the form of rows by columns, and *Freq* segmentations are the classification of frequency distribution, and Token/Type stands for the statistical frequencies of specific relationships with their corresponding dependency category R. The four syntactically conditioned matrices are extremely sparse with nulls in over 95% of the cells. Instead of eliminating the cells with lower frequencies, we kept all co-occurrences unchanged to avoid worsening data sparseness.

The matrices record the context with both syntactic dependencies and semantic content. These dual constraints yield rarer events than word co-occurrences in '*a bag of words*'. However, they impose more accurate or meaningful grammatical relationships between words providing the parser is reasonable accurate.

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

Given different methodologies to implementing parsing, it is hardly fair to appraise a syntactic parser. Molla and Hutchinson (2003) compared the Link Grammar 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 Link Grammar 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 explore the function of the syntactic dependencies: **RV**, **AN**, **SV**, and **VO** in deriving distributional similarity, which are acquired with the same Link Grammar parser, it is appropriate to use the Link Grammar parser to extract these dependencies.

4.2 Dimensionality Reduction in VSM

Singular Value Decomposition (SVD) often acts as an effective way of reducing the dimensionality of word space in natural language processing. A reduced SVD representation can diminish both 'noise' and redundancy whilst retaining the useful information that has the maximum variance. This approach has been dubbed Latent Semantic Analysis (LSA) (Deerwester, Dumais, Landauer, Furnas and Harshman, 1990; Landauer and Dumais, 1997) and maps the word-by-document space into word-by-concept and document-by-concept spaces. Note that the 'noisy' data in the raw co-occurrence matrices mainly comes from the results of wrong parsing and also redundancy exists as a common problem of expressing similar concepts in synonyms.

Typically at least 200 principal components are employed in Information Retrieval to describe the SVD compressed word space. Instead of optimising the semantic space versus other algorithms (through tuning the number of principal components in applications or evaluations), we specified a fixed dimension size for the compressed semantic space, which is thus not expected to be optimal for our experiment. We established 250 as a fixed size of the compressed semantic space. Among the singular values, the first 20 components account for around 50% of the variance, and the first 250 components for over 75%.

As is usual with the SVD/LSA application, we assume that the semantic representation of words is a linear combination of eigenvectors representing their distinct subcategorizations and senses, and that relating the uncorrelated eigenvector feature sets of different words can thus score their proximity in the semantic space.

4.3 Distributional Similarity

We consistently employed the cosine similarity of word vectors as used in LSA and commonly adopted in assessing distributional similarity (Salton and McGill, 1986; Schütze, 1992). The cosine of the angle θ, between vectors x and y in the *n*-dimensional space is defined as:

$$\cos\theta = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}_i}$$

where the length of x and y is $\|x\|$ and $\|y\|$.

Note that the accuracy and coverage of automatic term clustering inevitably depend on the size and domains of the corpora employed, as well as similarity measures. Consistently using one similarity method—the *cosine*, our main task in this paper is to explore the context interchangeability in automatic thesaurus construction, rather than to compare different similarity measures with one united syntactic structure that combines all the dependencies together. Although taking into account more similarity measures in the evaluations may solidify conclusions, this would take us beyond the scope of the work.

5. EVALUATION AND RESULTS

5.1 The 'Gold Standard' Thesaurus

It is not a trivial task to evaluate automatic thesauri in the absence of a benchmark set. Subjective assessment on distributionally similar words seems a plausible approach to assessing the quality of term clusters. It is practically unfeasible to implement it given the size of the term clusters. A low agreement on word relatedness also exists between human subjects.

The alternative way of measuring term clusters is to contrast them with existing lexical resources. For example, Grefenstette (1993) evaluated his automatic thesaurus with a 'gold standard' dataset consisting of Roget's Thesaurus ver. 1911, Macquarie Thesaurus, and Webster's 7th dictionary. 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, Curran (2003) incorporated more thesauri: Roget's Thesaurus (supplementing the free version of 1911 provided by Project Gutenberg with the modern version of Roget's Thesaurus II), Moby Thesaurus, The New Oxford Thesaurus of English, and The Macquarie Encyclopaedic Thesaurus.

The 'gold standard' datasets are not without problem due to their domain and coverage, because they are at best a snapshot of general or specific English vocabulary knowledge (Kilgarriff, 1997; Kilgarriff and Yallop, 2000). Moreover, the organization of thesauri forces different notions of being synonymous or similar, given the etymologic trend of words and different purposes of lexicographers. For example, as 1 of 1,000 topics in Roget's Thesaurus ver. 1911, there are two groups of synonyms {*teacher*, *trainer*, *instructor*, *institutor*, *master*, *tutor*, *director*, etc.} or {*professor*, *lecturer*, *reader*, etc.} under the topic of *teacher*. They express an academic concept of being in the position of supervision over somebody. In the noun taxonomy of WordNet, the synonym of teacher only consists of *instructor*, affiliated with the coordinate terms (sharing one common superordinate) such as *lecturer* and *reader*, or the hyponyms such as *coach* and *tutor*, or the hypernyms such *as educator* and *pedagogue*. As for *professor* and *master*, they both distance *teacher* by three links through their hypernym *educator*.

Subject to the availability of these thesauri or dictionaries, we incorporated both WordNet and Roget's Thesaurus, freely acquired, into the 'gold standard' thesaurus. WordNet only consists of paradigmatic relations and organizes a fine-grained semantic taxonomy mainly with the relationships of syn/antonym, IS-A, HAS-A, whereas Roget's Thesaurus covers both syntagmatic and paradigmatic relations and hierarchically clusters related words or phrases into each topic without explicitly annotating their relationships.

Kilgarriff and Yallop (2000) claimed that WordNet, along with the automatic thesauri generated under the hypothesis of similar words sharing similar syntactic structures, are *tighter* rather than *looser* in defining whether they are 'synonyms' or related words. This contrasts with Roget and the automatic thesauri derived through unordered word co-occurrences. Since we accounted for distributional similarity in the syntactically conditioned VSM, the reasonable way of evaluating it is to compare our automatic thesauri to WordNet. Apart from that, to perform a systematic evaluation on the relationships among distributionally similar words, we also included Roget as a supplement to the 'gold standard', as it covers words with both paradigmatic and syntagmatic relationships.

5.2 Similarity Comparison

We defined two distinctive measures to compare automatic thesauri with the 'gold standard', which are Sim_{WN} for WordNet and Sim_{RT} for Roget.

5.2.1 Similarity in WordNet

$Sim\mathrm{W}_N$ is based on the taxonomic similarity method we proposed (Yang and Powers, 2005, 2006). Since our method outperformed most popular similarity methods in terms of correlation with human similarity judgements, we employed them in the evaluation. Given two nominal or verbal concepts: $c1$ and $c2$, Sim_{WN} scores their similarity with:

$$Sim_{WN}(c1, c2) = \alpha_{str} \times \alpha_t \times \beta_t^{dist-1}, dist \leq \gamma$$

- α_{str}: 1 for nouns but for verbs successively falls back to α_{stm} the verb stem polysemy ignoring sense and form; or α_{der} the cognate noun hierarchy of the verb; or α_{gls} the definition of the verb.
- α_t: the path type factor to specify the weights of different link types, i.e. syn/antonym, hyper/hyponym and holo/meronym in WordNet.
- β: the probability associated with a direct link between concepts (type *t*).
- *dist*: the distance between two concept nodes
- γ: the path length *dist* is limited to depth factor γ, otherwise the similarity is 0

As for multiple senses of a word, word similarity maximizes its sense or concept similarity in WordNet.

It is not realistic to set up an absolute threshold on similarity values in the evaluation of ATC, given that different human subjects and algorithms have different distributions and biases on similarity judgement. This is applicable to compare through similarity ranks instead of similarity values. Hirst and Budanitsky (2005) noted that for the 65 noun pairs (Rubenstein and Goodenough, 1965) with human similarity scores in a Likert scale from 0 to 4, no pairs are located between 2.36 to 1.83, which forms a significant gap on similarity scores between the top 28 pairs and the other 37. They proposed that some value in the gap can serve as a cut-off to divide the 65 pairs into two groups. After ranking 65 pairs by their similarity scores, we selected the cut-off point 2.36 to distinguish similar (≥ 2.36) and dissimilar pairs (< 2.36), which corresponds to the searching depth limit $\gamma = 4$ in Sim_{WN}. Likewise the cut-off of 2 on the 130 verb pairs (Yang and Powers, 2006) corresponds to $^3 = 2$. Thus for the noun candidates in ATC, we set up $^3 = 4$, to retrieve similar words. If two nodes are syn/antonyms or related to each other in the taxonomy within the shortest path length of four links, we counted them as a successful hit. Similarly for the verb case, the shortest path length is two links.

5.2.2 Similarity in Roget's Thesaurus

Roget's Thesaurus divides its hierarchy into seven levels from the top *class* to the bottom *topic*, and stores topic-related words under 1 of 1,000 topics. Sim_{RT} counted it a hit if two words are situated under the same *topic*.

Note that the relationships among the 'gold standard' words retrieved by Sim_{RT} are anonymous. Although WordNet only organizes paradigmatic relationships, Sim_{WN} does not distinguish in what way two words are similar, for example, IS-A, HAS-A, or a mixture of them, and only collects words within a distance from zero (syn/antonyms) to four links in WordNet.

5.3 Candidate Words in the 'Gold Standard'

We select 100 seed nouns and 100 seed verbs with term frequencies of around 10,000 times in BNC. The average frequency of these nouns is about 8,988.9, and 10,364.4 for these verbs. High frequency words are likely to be generic or general terms and the less frequent words may not happen in the semantic sets. In practice, the average frequency of the nouns in $\mathbf{An}_X$, $\mathbf{aN}_X$, $\mathbf{Sv}_X$, and $\mathbf{vO}_X$ is decreased to 3,361.1, 5,629.1, 1,156.7, and 1,692.1, and the verbs in $\mathbf{rV}_X$, $\mathbf{Vo}_X$, and $\mathbf{sV}_X$ are decreased to 3,014.3, 3,328.9, and 1,971.8, as we only extracted syntactic dependencies from BNC. Overall, the average frequency of the nouns is about 2,959.7 across $\mathbf{An}_X$, $\mathbf{aN}_X$, $\mathbf{Sv}_X$, and $\mathbf{vO}_X$, and 3,960.9 for the verbs across $\mathbf{rV}_X$, $\mathbf{Vo}_X$, and $\mathbf{sV}_X$.

		WordNet						Roget	Total
		SA	D1	D2	D3	D4	∑		
Noun	$\mathbf{aN}_X$	462	2,825	14,244	41,483	48,625	107,639	141,102	232,181
	$\mathbf{An}_X$	458	2,887	14,278	41,940	49,267	108,830	142,218	234,424
	$\mathbf{vO}_X$	439	2,619	13,027	37,433	43,620	97,138	133,733	214,727
	$\mathbf{Sv}_X$	434	2,607	12,938	37,355	43,274	96,608	131,527	212,156
∑X		469	2,979	14,967	44,185	52,054	114,779	146,435	244,245
Verb	$\mathbf{rV}_X$	1,282	24,702	58,617			84,601	81,713	144,545
	$\mathbf{Vo}_X$	1,260	24,265	57,225			82,750	79,771	141,039
	$\mathbf{sV}_X$	1,269	24,354	57,642			83,265	80,681	142,256
∑X		1,297	25,283	60,483			87,165	83,415	148,455

Table 2: The word relatedness distribution in the 'gold-standard' across each matrix

We first used Sim_{WN} and Sim_{RT} to compare each seed word to all other words from the dependency sets, namely $\mathbf{An}_X$, $\mathbf{aN}_X$, $\mathbf{Sv}_X$, and $\mathbf{vO}_X$ for nouns and $\mathbf{rV}_X$, $\mathbf{Vo}_X$, and $\mathbf{sV}_X$ for verbs, to retrieve its candidate words in the 'gold standard'. Instead of a normal thesaurus with a full coverage of PoS tags, we only compiled the synonyms of nouns and verbs that account for the major part of published thesauri and are more informative than other PoS tags. The word distribution within different distances to the 100 nouns and 100 verbs in the 'gold-standard' are listed in Table 2, where ∑X indicates the overall nouns from $\mathbf{An}_X$, $\mathbf{aN}_X$, $\mathbf{Sv}_X$, and $\mathbf{vO}_X$ and verbs from $\mathbf{rV}_X$, $\mathbf{Vo}_X$, and $\mathbf{sV}_X$ in the 'gold-standard'. For the 'gold-standard' words from WordNet, SA denotes syn/antonyms of the targets, and DI the words with exactly I link distance to targets (for nouns $I \leq \gamma = 4$; for verbs $I \leq \gamma = 2$); ∑ denotes the total number of 'gold-standard' words in each matrix; and Total means the overall number of 'gold-standard' words from both WordNet and Roget. In Table 2 the average number of 'gold-standard' words across each matrix is evenly distributed.

The agreement between the WordNet-style and Roget-style words in the 'gold-standard' across these matrices, that is, the ratio of the number of words retrieved by Sim_{WN} and Sim_{RT} in both WordNet and Roget against the total number of 'gold-standard' words, is on average 7.3% on nouns and less than 15.2% on verbs. We aggregated all the 'gold-standard' words across $\mathbf{An}X$, $\mathbf{aN}_X$, $\mathbf{Sv}_X$, and $\mathbf{vO}_X$ for nouns, as well as $\mathbf{rV}_X$, $\mathbf{Vo}_X$, and $\mathbf{sV}_X$ for verbs, which results in 244,245 nouns and 148,455 verbs overall in the 'gold standard'. The agreement between WordNet and Roget candidates on nouns and verbs is respectively about 6.9% and 14.9%. About 14.8% and 11.6% nouns in WordNet and Roget are of same, so are 25.4% and 26.5% for verbs. Each target noun on average owns about 1,148 WordNet, 1,464 Roget, and 2,442 Total words in the 'gold standard', and each target verb 872, 834, and 1485 words respectively.

5.4 A Walk-Through Example

For each seed word, after computing the *cosine* similarity of the seed with all other words in each dependency matrix, we produced and ranked the top *n* words as candidates. We then applied the two heuristics: ***any two*** and ***any one*** on these candidates to forming automatic thesauri.

In Table 3 we exemplify the top 20 similar words of *sentence* and *strike* yielded in each dependency set. Consider the distributionally similar words of *sentence* and *strike* in $\mathbf{aN}_X$ and $\mathbf{rV}_X$ for example. The words related to the linguistic sense of *sentence* consists of *syllable*, *words*,

	Similar words
$\mathbf{aN}_X$	*imprisonment term utterance penalty excommunication syllable words punishment prison prisoner phrase detention hospitalisation fisticuffs banishment verdict Minnesota meaning adjective warder*
$\mathbf{An}_X$	*words syllable utterance clause nictation word swarthiness paragraph text homograph discourse imprisonment nonce phrase hexagram adjective verb niacin savarin micheas*
$\mathbf{vO}_X$	*soubise cybele sextet cristal raper stint concatenation kohlrabi tostada apprenticeship ban contrivance Guadalcanal necropolis misanthropy roulade gasworks curacy jejunum punishment*
$\mathbf{Sv}_X$	*ratel occurrence cragsman jingoism shiism Oklahoma genuineness unimportance language gathering letting grimm chaucer accent taxation ultimatum arrogance test verticality habituation*
any two	*imprisonment words utterance word term punishment paragraph text phrase jail verb meaning noun poem language passage sequence syllable lexicon fine*
any one	*Imprisonment utterance penalty excommunication punishment prison prisoner detention hospitalisation banishment Minnesota meaning contrariety phoneme consonant counterintelligence starvation fine cathedra lifespan*

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

	Similar words
$\mathbf{rV}_X$	*hit punch dehumanize whack stab volley ball digitalise attack lunge effuse arm carbonate disaffect rest brake relish concert glug*
$\mathbf{Vo}_X$	*vault rubberstamp unteach riot slang backhand cauterize whiff alloy hammer canoodle burnish chime hit trouble filch parry occasion volley roll*
$\mathbf{sV}_X$	*overgeneralise chime abate hallucinate decouple embolden overplay unionise flurry crock clink outrun unionize rampage segue frequent blast concatenate sequestrate derecognise*
any two	*hit volley stab chime whack blast riot hammer blind kick parry vault smash trouble punch last bounce lash smack alloy*
any one	*carom hit stab chime whack volley blast abate hammer blind rubber-stamp sequestrate parry dehumanize slang last punch smash unteach still*

(b) The similar words to *strike* (as a verb)

Table 3: A sample of automatic thesaurus items

adjective, etc, in $\mathbf{aN}_X$, while the words with the judicial sense make up around half of the 20 words including *imprisonment*, *penalty*, and the like. The words such as *hit* and *punch* from $\mathbf{rV}_X$ are from the literal sense of *strike*, together with its metaphorical sense of the words such as *attack* and *arm*.

The heuristic of ***any two*** collected the intersection of thesaurus items across these dependency sets. For example, *punishment* and *words* are the similar words to *sentence*, which respectively occurred in $\mathbf{aN}_X$ and $\mathbf{vO}_X$ as well as in $\mathbf{aN}_X$ and $\mathbf{An}_X$; *hit* and *blast* are the similar words to *strike*, which respectively occurred in $\mathbf{Vo}_X$ and $\mathbf{rV}_X$ as well as in $\mathbf{Vo}_X$ and $\mathbf{sV}_X$.

5.5 Performance Evaluation

Instead of simply matching with the 'gold standard' thesauri, Lin (1998) proposed to compare his automatic thesaurus with WordNet and Roget on their structures, taking into account the similarity scores and orders of similar words respectively produced from distributional similarity and taxonomic similarity. This approach can account for thesaurus resemblance under the hierarchy of WordNet or Roget, which is an apparent advantage over straight word matching.

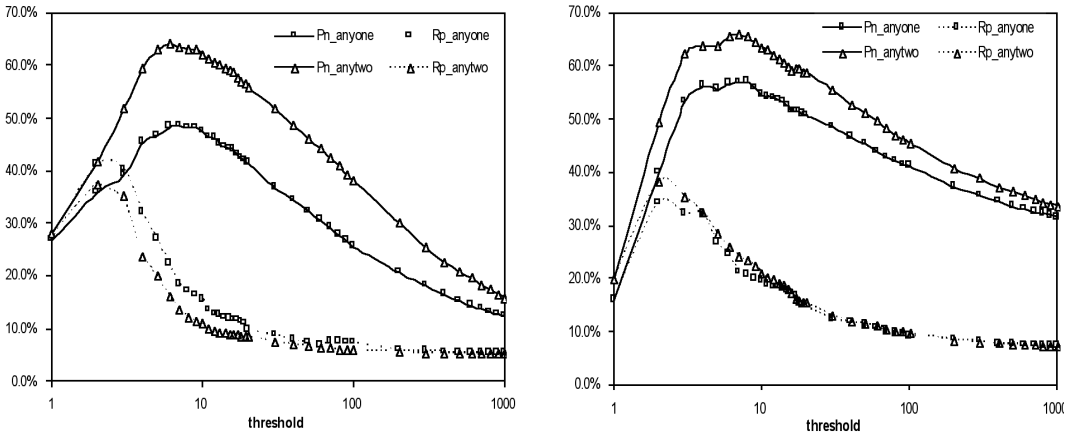
Instead of calculating the varied cosine similarity between each target vector yielded from automatic thesaurus and from WordNet or Roget (Lin, 1998), we adapted the concept of Precision (*Pn*) and Recall-precision (*Rp*) from information retrieval to demonstrate much sensible values of precision and recall for a ranked list. Given the top n similar words *S* for a target *T* in an automatic thesaurus *Pn* is defined as $|S|/n$, where $|S|$ refers to the number of *S* that can be retrieved in the top *n* similar words of *T* in WordNet or Roget. *Rp* is conditioned on precision and is correspondingly defined as $|S|/\sum d(S)$, where in terms of words $d(S)$ denotes minimum distance between *T* and *S* if *S* can be located within the top n similar words of *T* in WordNet or Roget. Analogously for the ranked word list from an automatic thesaurus, the top n similar words with respect to each sense of *T* in WordNet are produced in the order of hyper/hyponyms and holo/meronyms with exhausting initially synonyms and then antonyms, whereas the top *n* words in Roget can be subsequently acquired within +/-*n* (preceding/succeeding) words from *T* in each of its category. Through these redefined precision and recall *Pn* can stand for the coverage of the automatic thesaurus on potentially arbitrary senses or categories of *T* and *Rp* can describe relatedness of the thesaurus on the actual sense or category of *T*.

5.6 Results

We took the top *n* similar words derived from each co-occurrence matrix for ***any two*** or ***any one***, with *n* varying from 1 to 1000 in ten steps, roughly doubling each time. The results are shown in Table 4. We individually listed *Pn* and *Rp* values with respect to WordNet, Roget, and Total (the union of WordNet and Roget).

		any one						***any two***					
		WordNet		Roget		Total		WordNet		Roget		Total	
N		*Pn*	*Rp*	*Pn*	*Rp*	*Pn*	*Rp*	*Pn*	*Rp*	*Pn*	*Rp*	*Pn*	*Rp*
1	noun	22.0	22.0	15.0	15.0	27.0	27.0	24.0	24.0	12.0	12.0	**28.0**	**28.0**
	verb	13.0	13.0	7.0	7.0	16.0	16.0	15.0	15.0	8.0	8.0	**20.0**	**20.0**
2	noun	31.0	35.2	19.0	23.7	36.0	**41.2**	34.0	34.0	20.0	20.0	**42.0**	37.5
	verb	39.0	31.7	9.5	12.0	40.0	34.2	48.5	34.4	11.0	13.3	**49.5**	**38.2**
5	noun	42.4	21.1	22.2	29.5	46.8	**27.1**	56.6	17.1	28.4	24.0	**63.2**	20.0
	verb	54.2	25.6	20.2	17.1	55.8	26.9	62.6	27.4	23.8	15.0	**64.0**	**28.7**
10	noun	43.4	11.8	19.4	18.5	47.5	**15.5**	56.6	10.4	26.9	17.1	**62.3**	11.0
	verb	53.3	19.5	18.0	17.5	54.7	19.6	62.3	21.7	20.9	15.9	**63.7**	**21.2**
20	noun	37.7	9.5	16.1	13.8	41.6	**9.8**	50.2	8.7	22.7	16.5	**56.0**	8.4
	verb	49.3	15.0	13.9	15.0	50.9	14.7	57.5	15.6	16.1	13.8	**59.0**	**15.4**
50	noun	29.0	8.0	11.2	11.2	32.3	**7.4**	41.4	7.2	16.7	9.5	**46.4**	6.8
	verb	43.8	11.9	10.0	10.9	45.4	11.3	49.5	12.2	11.4	9.9	**51.3**	**11.5**
100	noun	22.9	8.4	8.2	9.5	25.7	**7.4**	33.8	6.6	12.8	6.6	**38.4**	5.9
	verb	39.7	10.0	7.7	8.4	41.2	9.2	44.1	10.4	8.4	7.5	**45.6**	**9.8**
200	noun	18.6	6.9	5.9	7.8	20.9	**5.9**	26.6	6.2	8.9	6.2	**30.2**	5.5
	verb	36.0	9.3	5.9	6.5	37.4	**8.6**	39.6	9.3	6.4	6.2	**41.0**	8.5
500	noun	13.6	6.4	3.9	6.1	15.4	**5.5**	18.6	6.0	5.4	5.8	**21.0**	5.3
	verb	32.6	8.5	4.2	5.7	33.8	**7.7**	35.1	8.5	4.6	5.3	**36.4**	**7.7**
1000	noun	11.0	6.3	2.8	5.5	12.4	**5.4**	14.1	6.1	3.6	5.5	**16.0**	5.2
	verb	30.5	8.2	3.4	4.9	31.6	**7.3**	32.7	8.2	3.6	4.9	**33.8**	**7.3**

Table 4: The adapted precision and recall in ATC (percentage)



(a) Nouns **(b) Verbs**

Figure 2: The overall performance comparison of any two vs any one

6. DISCUSSION

6.1 'any two' vs 'any one'

In Figure 2, it is clear that in terms of *Pn* measurement ***any two*** consistently outperformed ***any one*** for both nouns and verbs in thesaurus construction. The improvement in the precision of the ***any two*** clusters over the ***any one*** heuristic was significant ($p < 0.05$, paired t test). This is achieved under the condition of comparable *Rp*. Before reaching the threshold 200, the overall *Rp* for verbs for ***any two*** almost stay higher than for any one***, which is contrary in the case of nouns. Since then*** no noticeable difference can be observed. The reason behind this could be that some 'gold-standard' words derived from a matrix may never occur in the thesaurus entries from another matrix, which are consequently neglected in ***any two***.

We also extend this work to the words with intermediate (around 4,000) and low (around 1,000) term frequencies in BNC. For the 100 nouns and 100 verbs with the intermediate frequencies, 3,753.9 and 3,675.2 respectively, the average frequency of the nouns across $\mathbf{An}_X$, $\mathbf{aN}_X$, $\mathbf{Sv}_X$, and $\mathbf{vO}_X$ is 1,274.7, and the verbs across $\mathbf{rV}_X$, $\mathbf{Vo}_X$, and $\mathbf{sV}_X$ is 1,422.0. For the 100 nouns and 100 verbs with low frequencies: 824.1 and 864.6, the average frequency of the nouns across $\mathbf{An}_X$, $\mathbf{aN}_X$, $\mathbf{Sv}_X$, and $\mathbf{vO}_X$ is 297.0, and the verbs 342.2 across $\mathbf{rV}_X$, $\mathbf{Vo}_X$, and $\mathbf{sV}_X$. For the intermediate and low frequency words, the heuristic of any two still significantly outperformed the ***any one*** in yielding automatic thesauri ($p < 0.05$) with higher precision.

As the threshold increasing from 1 to 1000 in Table 4, both the nominal and verbal parts of thesaurus using the heuristics of ***any two*** and ***any one*** could corroborate a preference for relationships from WordNet rather than from Roget, since both *Pn* in WordNet contributed majority of the overall *Pn* in contrast to it in Roget. Note that from the figures shown in Table 2, we can observe that the overlap between WordNet and Roget is rather small, where only 14.8% of WordNet or 11.6% of Roget for nouns co-occur, so does 25.4% of WordNet or 26.5% of Roget for verbs. This could be caused by filtering out more Roget words present in the ***any one*** or ***any two*** thesaurus. This trend keeps unchanged even when more unrelated words could be introduced as the threshold approached 1000.

The lexical entries of *sentence* and *strike* in the ***any two*** and ***any one*** thesauri are listed in Table 5, where the threshold is varied from 1 to 20 with roughly doubling each time. In contrast to the

words produced from each dependency set in Table 3, both ***any two*** and ***any one*** can filter out apparently unrelated words such as *soubise* and *cybele* for *sentence* and *digitalise* and *carbonate* for *strike*. However, some truly similar words were also missed out in the ***any two*** thesauri, for example, *verdict* and *clause* in Table 3 (a), as well as *attack* and *filch* in Table 3 (b). Although the *any one* thesaurus can produce semantically similar words as its counterpart does for the low threshold, for the high threshold augmented to 20 the overall entries of *sentence* and *strike* in the ***any two*** thesaurus in Table 5 contain more closely related words than in the ***any one*** thesaurus. For example, both *Minnesota* and *counterintelligence* for *sentence* in Table 3 (a), along with *slang* and *unteach* for *strike* in Table 3 (b) were ruled out through the ***any two*** heuristic, which were regarded as valid words in the ***any one*** thesaurus. This can be partly complemented through increasing the threshold. Even with the threshold 50, the overall thesaurus entries were still acceptable with approximately 50% of total precision.

6.2 The Predominant Sense

Word senses in WordNet are ranked by their frequencies, where the first sense often serves as the predominant sense of a word. The predominant sense often serves as a back-off in sense

	Similar words	
	any one	***any two***
1	*imprisonment*	*imprisonment*
2	*imprisonment utterance*	*imprisonment words*
5	*imprisonment utterance penalty excommunication punishment*	*imprisonment words utterance word term*
10	*imprisonment utterance penalty excommunication punishment prison prisoner detention hospitalization banishment*	*imprisonment words utterance word term punishment paragraph text phrase jail*
20	*imprisonment utterance penalty excommunication punishment prison prisoner detention hospitalization banishment Minnesota meaning contrariety phoneme consonant counterintelligence starvation fine cathedra lifespan*	*imprisonment words utterance word term punishment paragraph text phrase jail verb meaning noun poem language passage sequence syllable lexicon apprenticeship*

(a) The similar words of *sentence* (as a noun) with both linguistic and judicial senses

	Similar words	
	any one	***any two***
1	*carom*	*hit*
2	*carom hit*	*hit volley*
5	*carom hit stab chime whack*	*hit volley stab chime whack*
10	*carom hit stab chime whack volley blast abate hammer blind*	*hit volley stab chime whack blast riot hammer blind kick*
20	*carom hit stab chime whack volley blast abate hammer blind rubber-stamp sequestrate parry dehumanize slang last punch smash unteach still*	*hit volley stab chime whack blast riot hammer blind kick parry vault smash trouble punch last bounce lash smack alloy*

(b) The similar words of *strike* (as a verb) with both literal and metaphorical senses

Table 5: A sample of ATC produced through the heuristic *any two* under varied thresholds

disambiguation. To study the sense distribution of the words in automatic thesaurus, we also calculated *Pn* on the condition of extracting the 'gold-standard' words exclusively related to the first sense of a target (*First*), in contrast to all the senses.

Overall the precision of *First* sense is not less than 50% of the precision of all sense for both nouns and verbs in the ***any two*** heuristic. This implies that distributionally similar words derived using the ***any two*** heuristic are more semantically related to the first sense of a target, around 50% or more, than other senses. Even for the ***any one*** heuristic, around 50% of the words that match a 'gold-standard' for any sense, hold semantic relatedness with the first senses of targets.

The unbalanced sense distribution among the thesaurus items shows the uneven usages of words with respect to the Zipf's Law (1965). Kilgarriff (2004) also noted Zipfian distribution of both word sense and words when analysing the Brown corpus and BNC. The predominant sense of a word can be formed through their distributionally similar words instead of laborious sense annotation work, which serves as an important resource in sense disambiguation.

6.3 Distributional Similarity and Semantic Relatedness

Semantic similarity is often regarded as a special case of semantic relatedness, while the latter also contains word association. Distributional similarity consists of both semantic similarity and word association between a seed word and candidate words in its thesaurus items, except for the 'noisy' words (due to the parsing or statistical errors) that hold no plausible relationships with the seed. Consider the distributionally similar words of *sentence* produced in $\mathbf{aN}_X$ in Table 3 (a) for example. Only three words, namely *term*, **phrase**, and **verdict**, were connected with *sentence* through the similarity measurement of Sim_{WN} in WordNet, whereas 14 words such as *phrase* and *penalty* shared the same topics with *sentence* in Roget. The noun *sentence* consists of three senses in WordNet,

- *sentence#n#*1: a string of words satisfying the grammatical rules of a language
- *sentence#n#*2: (criminal law) a final judgment of guilty in a criminal case and the punishment that is imposed
- *sentence#*n#3: the period of time a prisoner is imprisoned

The word *sentence* is also located in Section 480 (*Judgement*), 496 (*Maxim*), 535 (*Affirmation*), 566 (*Phrase*), and 971 (*Condemnation*) in Roget. For example, the nominal part of Section 480 is,

> *480. Judgment. [Conclusion.]*
>
> *N. result, conclusion, upshot; deduction, inference, ergotism[Med]; illation; corollary, porism[obs3]; moral. estimation, valuation, appreciation, judication[obs3]; dijudication[obs3], adjudication; arbitrament, arbitrement[obs3], arbitration; assessment, ponderation[obs3]; valorization. award, estimate; review, criticism, critique, notice, report. decision, determination, judgment, finding, verdict,* ***sentence****, decree; findings of fact; findings of law; res judicata[Lat]. plebiscite, voice, casting vote; vote &c. (choice) 609; opinion &c. (belief) 484; good judgment &c. (wisdom) 498. judge, umpire; arbiter, arbitrator; assessor, referee. censor, reviewer, critic; connoisseur; commentator &c. 524; inspector, inspecting officer. twenty-twenty hindsight [judgment after the fact]; armchair general, Monday morning quarterback.*

Generally *sentence#n#*1 in WordNet can be projected into Section 496 and 566, and *sentence#n#*2 into Section 480 and 971, and *sentence#n#*3 into Section 535. With respect to the evaluation of Sim_{WN} in WordNet, *term* in Table 3 (a) is the hypernym of *sentence#n#*3; and *phrase* and sentence#n#1 distance themselves in three links, say, sentence#n#1 has a meronym of clause

that is a coordinate of *phrase*; and *sentence#n#2* bears the same hypernym with *verdict* within four links. Apart from the paradigmatic relationships in WordNet, the three words also connect with *sentence* through Sim_{RT} in Roget, where words such as *verdict* and *sentences* are located under the same section—*Judgement* (480). However, *sentence* holds more relations of being in the same domain with its similar words in the thesaurus from $\mathbf{aN}_X$. For example, *penalty* and *sentence* come from/exist in Section 971, which expresses the notion of criminality deserving a penalty in a way of judicial sentence, and *prisoner* and *sentence* are situated in Section 971, which illustrates being in prison resulting from judgements in a court in the context of criminal law.

As we compute distributional similarity on the assumption of similar words sharing similar contexts conditioned by grammatical relations, in general more paradigmatic relations can be found than syntagmatic ones. In Table 4, the higher precision for WordNet than for Roget's Thesaurus show that distributionally similar words are more semantically similar rather than associated words. This is consistent with the conclusion of Kilgarriff and Yallop (2000) on computing distributional similarity that the hypothesis of similar words sharing similar contexts constrained by grammatical relations can yield *tighter* or WordNet-style thesauri, whereas the hypothesis of similar words sharing unconditioned co-occurrences can yield *looser* or Roget-style thesauri. Note that distributionally similar words could be semantically opposite to each other, given the common grammatical relations they often share. For example, in the automatic thesaurus produced with ***any two***, the nouns *failure* and *success*, or *strength* and *weakness*, are antonymous, as well the verbs *cry* and *laugh*, *deny* and *admit*.

It is clear that the 'gold standard' is subject to the vocabulary size of WordNet and Roget's Thesaurus. The worse case is from the 1911 version of Roget's Thesaurus we adopted, where words generated in modern times are not contained. For example words such as *software* and its distributionally similar words, including *emulator*, *unix*, *NT*, *Cobol*, *Oracle* (as the database system), *processor*, and *PC*, are not included in the 1911 version of Roget. We selected the target word with relatively higher frequencies in BNC and did a simple morphology analysis in the construction of the matrices using word-mapping table in WordNet, so that all nouns and verbs from automatic term clustering can be covered (at least in WordNet). However, not all word relationships in automatic thesauri could be contained in WordNet, even though we have included Roget to supply richer relationships. For example, take the words *sentence* and *detention*. In Table 3 (a) *detention* is listed in the top 20 similar words to *sentence* on $\mathbf{aN}_X$, but they have no direct or indirect links in WordNet, nor are they situated under any *topic* or *section* in Roget, but their intense association has become commonly used. Likewise, *kidnap* as one of the top 20 similar words to *attack* on $\mathbf{rV}_X$ in Table 3 (b), which is distributionally similar to *attack*, but there are no existing connections between them in WordNet and Roget.

7. CONCLUSION

Despite the introduction of grammatical relations in ATC, most methods still combined these relations into one united representation for distributional similarity computation, which worked analogously to these based on the premise of '*a bag of words*'. Instead of the united representation, we first categorized grammatical relations into four types of syntactically conditioned contexts, and then retrieved similar words under the assumption of their context interchangeability across any two types of contexts. Our method can improve ATC with significantly higher precision than the traditional methods. Future research will focus on how to cluster and extract word senses from the thesaurus entries, as well as on how to harvest semantic relations among the thesaurus entries.

REFERENCES

AGIRRE, E., ANSA, O., MARTINEZ, D. and HOVY, E. (2001): Enriching WordNet concepts with topic signatures. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburg, USA.

CARROLL, J., BRISCOE, T. and SANFILIPPO, A. (1998): Parser evaluation: A survey and a new proposal. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 447–454. Granada, Spain.

CURRAN, J.R. (2003): From distributional to semantic similarity. Ph.D thesis. University of Edinburgh.

DEERWESTER, S.C., DUMAIS, S.T., LANDAUER, T.K., FURNAS, G.W. and HARSHMAN, R.A. (1990): Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6): 391–407.

FELLBAUM, C. (1998): *WordNet: An Electronic Lexical Database*. Cambridge, MA, The MIT Press.

GREFENSTETTE, G. (1992a): Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 324–326. Newark, Delaware.

GREFENSTETTE, G. (1992b): Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 89–97. Copenhagen, Denmark.

GREFENSTETTE, G. (1993): Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches. In *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*, 143–153.

HARRIS, Z. (1985): Distributional structure. In *The Philosophy of Linguistics,* KATZ, J.J. (ed). New York, Oxford University Press. 26–47.

HINDLE, D. (1990): Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 268–275. Pittsburgh, Pennsylvania.

HIRSCHMAN, L., GRISHMAN, R. and SAGER, N. (1975): Grammatically-based automatic word class formation. *Information Processing and Management*, 11: 39–57.

HIRST, G. and BUDANITSKY, A. (2005): Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering,* 11(1): 87–111.

KEENAN, E. and COMRIE, B. (1977): Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8: 62–100.

KILGARRIFF, A. (1997): I don't believe in word senses. *Computers and the Humanities*, 31(2): 91–113.

KILGARRIFF, A. (2004): How dominant is the commonest sense of a word? In *Proceedings of the 7th International Conference (TSD 2004, Text, Speech and Dialogue)*, 103–112. Brno, Czech Republic.

KILGARRIFF, A. and YALLOP, C. (2000): What's in a thesaurus? In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC-2000*, 1371–1379. Athens, Greece.

LANDAUER, T.K. and DUMAIS, S.T. (1997): A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104: 211–240.

LEVELING, J. and HARTRUMPF, S. (2005): University of Hagen at CLEF 2004: Indexing and translating concepts for the GIRT task. In *Proceedings of Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, 271–282. Bath, UK.

LIN, D. (1997): Using syntactic dependency as a local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 64–71. Madrid, Spain.

LIN, D. (1998): Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics*, 768–774. Montreal, Quebec, Canada.

LOWE, W. (2001): Towards a theory of semantic space. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 576–581. Edinburgh, UK.

McCARTHY, D., KOELING, R., WEEDS, J. and CARROLL, J. (2004): Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 267–287. Barcelona, Spain.

MILLER, G.A. and CHARLES, W.G. (1991): Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1): 1–28.

MOLLA, D. and HUTCHINSON, B. (2003): Intrinsic versus extrinsic evaluations of parsing systems. In *Proceedings of European Association for Computational Linguistics(EACL), workshop on Evaluation Initiatives in Natural Language Processing*, 43–50. Budapest, Hungary.

PADÓ, S. and LAPATA, M. (2007): Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2): 161-199.

PANTEL, P. (2005): Inducing ontological co-occurence vectors. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 125–132. Ann Arbor, Michigan.

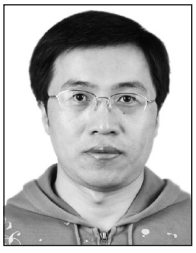
PANTEL, P. and LIN, D. (2002): Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 613–619. New York, NY, USA.

PENNACCHIOTTI, M. and PANTEL, P. (2006): Ontologizing semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, 793–800. Sydney, Australia.

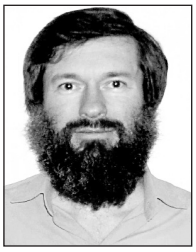
PLAS, L.V.D. and BOUMA, G. (2005): Contexts for finding semantically similar words. In *Proceedings of the 20th International Conference on Computational Linguistics*, 173-186. Geneva, Switzerland.

RESNIK, P. (1997): Selectional preference and sense disambiguation. In *Proceedings of ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How?* 52–57. Washington, USA.
RUBENSTEIN, H. and GOODENOUGH, J.B. (1965): Contextual correlates of synonymy. *Communications of the ACM*, 8(10): 627–633.
SAHLGREN, M. (2006): The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D thesis. Stockholm University.
SALTON, G. and McGILL, M. J. (1986): *Introduction to Modern Iformation Retrieval*. New York, NY, USA, McGraw-Hill.
SÁNCHEZ, D. and MORENO, A. (2005): Web-scale taxonomy learning. In *Proceedings of the Workshop on Learning and Extending Lexical Ontologies by using Machine Learning*, Bonn, Germany.
SCHULTE IM WALDE, S. (2006): Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2): 159–194.
SCHÜTZE, H. (1992): Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, 787–796. Minneapolis, Minnesota, USA.
SLEATOR, D. and TEMPERLEY, D. (1991). *Parsing English with a Link Grammar*. Computer Science Technical Report. CMU-CS-91-196. Carnegie Mellon University.
STAMOU, S. and CHRISTODOULAKIS, D. (2005): Retrieval efficiency of normalized query expansion. In *Proceedings of the 6th International Conference (CICLing 2005): Computational Linguistics and Intelligent Text Processing*, 593–596. Mexico City, Mexico.
WEEDS, J. E. (2003): Measures and applications of lexical distributional similarity. Ph.D thesis. University of Sussex.
YANG, D. and POWERS, D.M.W. (2005): Measuring semantic similarity in the taxonomy of WordNet. In *Proceedings of the Twenty-Eighth Australasian Computer Science Conference (ACSC2005)*, 315–322. Newcastle, Australia.
YANG, D. and POWERS, D.M.W. (2006): Verb similarity on the taxonomy of WordNet. In *Proceedings of the 3rd International WordNet Conference (GWC-06)*, 121–128. Jeju Island, Korea.
YAROWSKY, D. (1993): One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop*, 266–271. Princeton, New Jersey.
ZIPF, G.K. (1965): *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. N.Y., Hafner Pub. Co.

BIOGRAPHICAL NOTES



Dongqiang Yang



David Powers

Dr Dongqiang Yang is a research fellow in the School of Computer Science, Engineering and Mathematics at Flinders University of South Australia (Flinders). His research interests mainly include natural language processing, information retrieval and data mining. He received his PhD in Computer Science from Flinders, where he is also the member of Artificial Intelligence and Language Technology Laboratories. As a primary investigator, he is working on the project that aims to automate collection and classification of online educational resources.

Dr David Powers is Professor of Computer Science and Director of the Artificial Intelligence and Language Technology Laboratories in the School of Computer Science, Engineering and Mathematics at Flinders University of South Australia. His research interests focus around learning of language and ontology, with a particular emphasis on cognitive approaches and bio-plausible models. Prof. Powers also has a variety of applications in various stages of commercialization in the areas of web search, speech control interface, brain computer interface and computer assisted education.