Wordnet vs. Distributional determination of word similarity

Dongqiang Yang and 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. In this paper we propose a kind of distributional word similarity after extracting some syntactic relation in the sentence. As a supplement to the richness of WordNet, it acquires some lexical knowledge from large volume of corpus. It is also a powerful tool to judge word similarity.

Introduction and Aims

The problem of how similar words are or how closely words are associated has important applications in several areas including automatic thesaurus construction (Grefenstette 1992; Hearst 1992; Riloff and Shepherd 1997; Pichon and Sébillot 1998; Berland and Charniak 1999; Caraballo 1999), word sense disambiguation (Dagan, Marcus et al. 1993; Lin 1997; Dominic 2003) and information retrieval (Salton 1973; Grefenstette 1992; Leacock, Towell et al. 1996).

Word Association Norms (**refs**) is a knowledge base available for word association applications based on empirical results from Psychological word association experiments in which a subject is given a stimulus word and asked to respond with the first word that comes to mind. The data are provided in the form of lists of subject stimuli, responses and frequencies that can be interpreted as indicating the extent of the similarity between the paired words.

Word association data of a different kind can also be captured from corpora of naturally occurring speech, or text, and may be further categorized according to topic, register, ethnicity, etc.

WordNet and Roget's Thesaurus represent hand-crafted repositories of similar information, although in some respects richer information is provided by these – e.g. WordNet in some cases identifies the kind of relationship between words.

We are particularly interested in how to automatically collect, update and make effective use of WAN-like or WordNet-like data from large corpora.

Current approaches

The meaning of the word is to a large extent latent in the contexts where it occurs and most vocabulary is learned from context rather than a dictionary. Indeed lexicons, dictionaries and thesauri are just specialized contexts that are designed to facilitate catching the meaning of an unfamiliar word, but a mere gloss or definition is insufficient to fully convey the scope of such a word.

Each word may be regarded as interacting with and having a relationship with all its neighboring words (Firth 1957), but these relationships and interactions can vary quite considerably and are directly dependent on the both the grammatical and conceptual systems of both the speaker and listener as well as the context in which they are uttered. For example, "David got a cold and had to go … ", it is not difficult to infer that the omitted portion of the sentence should be related to going to the doctor or hospital or even home and bed. The missing content is likely to contain medical or pharmaceutical terms, but a raw unmotivated completion like "… to the shopping center" has too little an association with the cold to be likely (unless it was followed by something about doctors or medication). We can similarly guess aspects of the broad meaning of the word "eucalyptus", even without any pre-knowledge in it, through co-occurrence words that have some kind of syntactic relation to in contextual sentences – in different contexts we might learn it means a kind of gum tree, food for koalas, an oil for removing stickers, or a medication for relieving colds.

Roughly speaking, two types of relations between the stimuli and the responses in the WAN can be found: syntagmatic associations and paradigmatic associations. Syntagmatic association defines the relation between words in different slots in the sentence and implies the word co-occurrence in the syntactic structure of the sentence. Paradigmatic association represents the relationship between words that can occur in the slot in a sentence in both a syntactic and a semantic sense. To acquire these relationships in the corpus two kinds of techniques are relatively popular with respect to how to represent word meaning.

Manmade dictionary or thesaurus

Machine-readable dictionaries provide a well-designed pattern for inferring word senses. Lesk (Lesk 1986) utilizes overlapping between definitions for each sense of stimulus and responded words, i.e. how many common words they share, to stands for the choices in cross sense-comparison of the words. (Pedersen, Banerjee et al. 2003) propose to judge the concepts similarity through their gloss vectors from the matrix that are build on the 1.4 million glossary corpus of WordNet. The cell in the matrix is the co-occurrence frequency in each concept gloss in WordNet. This method is a partial improvement of the Lesk algorithm (Lesk 1986), in which the similarity is reflected by the common words of glosses for two different word senses. Owing to the concise gloss of each sense and the slow and rare updating of WordNet, this method has little use in practice.

Another class of knowledge-rich method makes use of semantic networks, or a semantically tagged corpus, to induce the relatedness between words by calculating semantic distance between two words in a the well-organized taxonomy such as WordNet or Roget's thesaurus (Pedersen, Banerjee et al. 2003; Yang and Powers

2005) . Note that different hierarchies may relate only the same part of speech (WordNet) or may handle mixed parts of speech (Roget). In WordNet's noun taxonomy it is primarily the hyper-hyponym and hol-meronym links that have been used to carefully categorize and distinguish the relationship between the different words and senses.

Knowledge-poor methods,

The frequency of co-occurrence in context can be represented using techniques such as word contingency matrices or n-gram statistics with or without making use of information about or deriving from syntactic structure. Given unsupervised learning is used to acquire the semantic information about a word, it should in principle cover broader domains and larger vocabulary sizes than the retrieval of word similarity from dictionary or semantic networks.

- a bag of words

Here we assume that semantically related words that are likely to co-occur in a context (traditionally document) can be treated as an unstructured bag of words for the purposes of determining word associations/similarity. A matrix may be constructed in word-by-word or word-by-document order within fixed size window where any cell value could be the Term Frequency (TF) or TF*IDF (Inversed Document Frequency). And then any of a number of distance measures or clustering techniques may be used to characterize the similarity of words or associate related words (**ref**).

- syntactic relationships (shallow/deep parsing) – ??? ARE THESE KNOWLEDGE POOR???

Here we assume that specific semantic relationships relate to the grammatical structures in which they occur and thus judging word similarity requires comparing the syntactic components and their dependency relations in the surrounding parts of each word. This suggests that we first need to extract syntactic information using at least part-of-speech tagging or shallow parsing, and then need to judge word similarity in the context of the syntactic roles that relate the extracted chunks (Grefenstette 1993; Gasperin, Gamallo et al. 2001). After that, the grammatical context of the word can be translated into a set of attributes that includes might include part of speec (POS) tags such as ADJ, NN, NNPREP, SUBJ, DOBJ. Finally, again we will use an appropriate measure to determine the similarity of two words.

Lin (Lin 1998) employed a broad-coverage parser to find the dependency triples from the corpus, which is similar to Grefenstette's method. He extracted nearly 56 million dependency triples from the 64 million words (??? AFTER THROWING AWAY NON-CONTENT WORDS???). Then he computed the pair-wise similarity between the nouns, verbs and adjectives/adverbs using the metrics of Jacquard, Dice, Cosine, and Hindle. He compared the automatically constructed thesaurus through this method with the entries in WordNet and Roget's thesaurus.

A new approach

Yang and Powers (Yang and Powers 2005) propose a new model to inspect noun similarity based on the taxonomy of WordNet, and this sets the current gold standard for word similarity performance. Their algorithm performs far better than any current knowledge poor method relying on the statistical distribution of words – indeed their model has an almost 90 percent correlation with average human judgement for the 65 pairs of nouns (Rubenstein and Goodenough 1965), performs significantly better than most subjects, and agrees almost as well as different groups of humans of similar background. In this paper we put forward a totally new model of.distributional similarity and benchmark it against the data set used by Yang and Powers.

Basic procedures

Deese(Deese 1966) classifies part of speeches into syntagmatic and paradigmatic associations as discussed earlier:

1. Association responses for adverbs tend to reflect the syntagmatic relations.
2. Nouns are paradigmatic. The paradigmatic relations are not totally focused within a superordinate, subordinate, and co-ordinate hierarchy. The similarity of nouns is still accessed through intersection of their descriptions that specify the characteristics of entities. From a psychological perspective Deese concludes that

- Nouns can be related by being grouped together.
- Nouns can be related by conceptual or physical environment.
- Nouns can be related by sharing common attributes.

3. Verbs and adjectives fall in between syntagmatic and paradigmatic associations

Clark and Card (Clark 1969; Clark and Card 1969) argue that the conceptual relations are extracted from the syntactic structure in a sentence, which includes subject, direct object, and predicate of the sentence. Generally the syntagmatic relation provides us a clue for tracking down the meaning of the sentence after parsing the sentence and achieving the grammatical dependency.

Our task here is to find some mapping relations from the syntactic environment of stimuli and responses, which includes nouns, verbs, adjectives, and adverbs, to their conceptual contiguity in a real world.

Proposal

Parsing

“what is the context ?”
In the oxford dictionary, the context is carefully defined as:

1. *The weaving together of words and sentences; construction of speech, literary composition.*

2. *The connected structure of a writing or composition; a continuous text or composition with parts duly connected.*
3. *The connexion or coherence between the parts of a discourse.*
4. *The whole structure of a connected passage regarded in its bearing upon any of the parts which constitute it; the parts which immediately precede or follow any particular passage or 'text' and determine its meaning.*

The essence of its definitions is the connection or link of the context is able to cover some kind of meaning that can benefit understanding of whole discourse. A bag of words is a rough explanation of context, which is a random selection with too much noisy data. We assume shallow level context should be a fine-grained structure for which syntactic dependency is source. Otherwise no big difference exists between animal language and human language (Nowak, Plotkin et al. 2000).

To retireve these syntactic dependency as correct as possible we employ a wide-used free parser based on link grammar[1].

Suppose a triple < *w1, r, w2*> to describe objects *w1, w2* and their dependency *r* where r has bi-directional actions on the pair of words *w1* and *w2*. For example if *w1* modifies *w2* in the kind of modification relation *r*, all such *w2* in the corpus form a context for *w1*, and all *w1* in the corpus will be context for *w2*. We cover 5 categories of relationships between words, as shown in table1:

1. action modifier (RV) : holds adverb and verb relations
   E*: Adverb.+ verb, adjective, other adverb etc..
   MV* : verb+ all kinds of modifing-verb (Adverb, participle etc.)
   MV* + I, J, Mg: verb+ ...+ preposition phrase, Participle (infinitival) modifiers etc.
2. object modifier (AN) : covers adjective and noun relation
   A, AN: adjective, noun + noun;
   M [a, g, v, p, r]: noun + all kinds of post-nominal modifier;
   Mp + * : noun + , + all kinds of post-nominal modifier;
3. agent to predict (SV): contain subjective noun and verb relations
   S*: subjects + verb
4. predict to argument (VO) : contain objective noun and verb relations
   O : verb + direct objects or indirect objects or infinitive complement
5. subject to object (SO) : keep track of noun to noun relations.
   SO: subject + … + object;

Table 1. an example sentence and different links in it

"The care of people in the community, with are ill with HIV infection and AIDS, together with the education of schoolchildren to help prevent the spread of this terrible disease is becoming more and more urgent." _exerted from BNC

[1] http://www.link.cs.cmu.edu/link/

Care people People community Community ill People ill HIV infection Schoolchildren Education Spread disease Terrible disease	Infection ill AIDS ill Education ill	Care urgent	Education help	Prevent spread
AN	RV	SO	SV	VO

After extracting the relationships and morphological analysis, we construct five raw matrixes, which can be taken as targets by contexts. For the 5 matrixes, we will handle them with same techniques in following sections. Without loss of generosity let's denote the subject –verb matrix as X_{sv}. X_{sv} is a $m \times n$ matrix in which rows correspond to subjects in each language segment, and columns correspond to verbs. x_{ij} is the total frequency of the ith subject with jth verb. The ith row of X_{sv} constructs the profile of ith subject in which it relates all different verbs in the corpus. And vice verse in the jth column of X_{sa} which reflect the attributes of the jth verb in the context of different subject.

Space transmission

The kernel problem is to find how translate the syntactic space or word occurrence space into semantic space, i.e. how to find a suitable function to represent the meaning of words. Some researchers directly shift the syntactic space X into semantic space with complex similarity metrics like mutual information of two objects (Hindle 1990), weighted Jaccard efficient (Grefenstette 1992). (Deerwester, Dumais et al. 1990) in the latent semantic analysis (LSI) dig out concept space with single value decomposition (SVD) to solve the problem of synonyms and polysonyms trouble problems in information retrieval. (Honkela and Hyvärinen. 2004) use independent component analysis (ICA) to extract linguistic features shared by words after constructing a context space of 2000 common words.

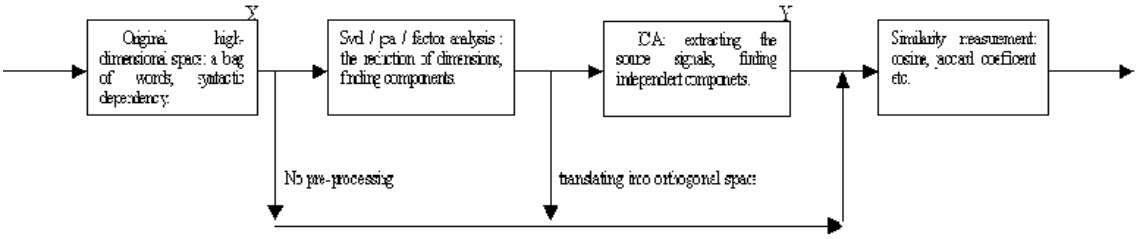

Fig. 1. the architecture of space transformation.

SVD

We assume SVD can detect a kind of pseudo-semantic pattern behind the noisy syntactic representation. The following is the decomposition of X_{SV}:

$$X_{SV} = USV^T \quad (1)$$

Where U is a $m \times r$ matrix of left single vectors from the standard eigenvectors of square matrix $X_{SV} X_{SV}^T$, r is the rank of X_{SV} ($r \leq \min(m, n)$), V^T is a $r \times n$ matrix or right single vectors from the eigenvectors of $X_{SV}^T X_{SV}$, S is a diagonal matrix of square root of single values from both $X_{SV} X_{SV}^T$ and $X_{SV}^T X_{SV}$. After SVD the columns of U renew the context of each subject as an eigen-verb, the rows of V^T renews the context of each verb as an eigen-subject. The whole point of SVD relies on the keeping of first l largest single values and corresponding l left column eigenvector and l right row eigenvectors to maximally approach X_{SV}. SVD filter out the redundant information and hold on the most useful information that has the maximum variance on each variable.

Eigen-verbs mainly capture the feature pattern of subjects in pseudo-semantic space. Although we cannot fully understand the meaning of eigenvector in each single space SVD still can classify source signals with uncorrelated eigen-verbs or eigen-subjects, extract some underlying patterns from the syntactic space. If we assume the meaning of word is linear combination of eigenvectors, and matching feature sets of different words scores their proximity in the semantic space, semantic attributes, which are subject to these, are just uncorrelated but not independent owning to its ability to distinguish direction of semantic characters.

Suppose Multivariate data is distribution of gaussian, SVD/PCA (principal component analysis which has same effect when components are outputs of covariance matrix of X) is just trying to find as faithfully as possible to represent the syntactic space with the second order information from the covariance matrix. The pseudo-semantic space is located along with orthogonal direction of eigenvectors that maximize the variance of X_{SV}.

ICA

ICA is a kind of high-order methods that try to find information outside the covariance matrix.

$$Y = AS \quad (2)$$

Where Y is a $l \times m$ matrix of mixed signals which in our case is from the U^T or V^T of SVD of X_{SV}. S is a $l \times m$ matrix which is recovered as source signals or latent variable from ICA. A often refers to mixing matrix that ICA need to determine. Different from second-order methods like SVD and PCA, ICA manages to find the direction of original signals to make the semantic space meaningful (Hyvärinen 1999). All of the recovered signals are independent and nearly satisfy the non-gaussian distribution otherwise the uncorrelated variables in the covariance matrix is easy to find impendence in the gaussian distribution, which is enough in task of PCA. It is worth to mention that for the source signal extracted from ICA, we cannot decide the order of signals, which are arbitrary permutation of different output.

Experiment preparation

We parse British National Corpus (BNC), which contains nearly 100 million words (both text and spoken English materials) to get our basic syntactic space. After filtering out function words and common words, these syntactic relations are stored in the 5 matrix, see table 2. We then translated frequency of data item in each matrix into information content with its logarithm after frequency plus one.

We first decompose each matrix by SVD to get its pseudo-semantic spaces for rows and columns, and then we feed these spaces individually into FastICA[2] to get its real semantic space.

Table 2. all kinds of different frequency of relations after parsing the BNC, here type is how many kinds of relations happened n times, token corresponding to how many relations happened n times.

	Dimensions		1	2-10	11-20	21-30	>31
AV	37,456	Token	863,063	2,276,448	481,490	234,930	692,2

[2] http://www.cis.hut.fi/projects/ica/fastica/

	14,225	Type	863,063	751,866	33,792	9,508	10,85
AN	48,528	Token	1,813,673	6,243,391	1,483,063	799,765	3,617,8
	37,589	Type	1,813,673	2,039,980	103,586	32,283	44,89
SV	32,696	Token	511,845	1,699,445	297,791	133,309	380,7
	11,281	Type	511,845	587,350	20,953	5,392	5,95
VO	6,081	Token	488,481	1,811,506	475,408	266,182	1,286,9
	33,354	Type	488,481	575,108	33,172	10,731	15,61
SO	33,501	Token	926, 860	2,223,401	227,086	83,626	163,8
	34,128	Type	926,860	820,855	16,179	3,406	2,95

Evaluation

The significant single value.

The size of dimensionality of semantic space strongly depends on different sources of corpus and specific applications. The component of each entity denoted by vector in the semantic space corresponds to sematic features or attributes, which construct feature comparison model in human semantic memory. Some researchers employ whole components in context space to acquire word semantics (Grefenstette 1992; Lin 1997), others compress the context space to find latent semantic space, where semantic features are focused on the reduced number of components (Deerwester, Dumais et al. 1990). Even in methods facing reduced sematic space there are absent of consensus in defining how to describe feature sets. The impression of all the works grounded on SVD or PCA comes out to adjust the number of principle components to adapt distinctive applications. The best performance is subject to how many single values a system is outputting. Most language application projects employ at lest 200 principle components to describe the transmitted space and reduce the expensive computing. In the work of gene expression analysis by SVD, (Wall, Rechtsteiner et al. 2003)) only employ the first two principle components. Here our motive is not to make a big difference with other similar algorithms in some specific applications or evaluations, but to try to state that the term space of syntactic dependency can also anticipate semantic space of concepts. Henceforth we output a fixed number of the principal components as a fundamental dimensional size of pseudo semantic space before we commence each evaluation. It does not mean this number will be the optimal value for each standard evaluation, which makes the highest score in the trial. In what follows, we explain how to select the reasonable number.

In the hand –crafted semantic net, Roget's Thesaurus -1911, there are nearly 1,000 semantic categories, which organize over 40,000 words. With respect to expensive computation of SVD on our sparse matrixes and the huge word sense pairs (approximately 200,000) in WordNet we set up 1000 as default size of semantic feature set, viz. we export 1000 single values in each syntactic matrix and then to tailor an appropriate size of dimensions.

Generally we don't clearly distinguish PCA and SVD on their functions of compressing vector space. But if we are reluctant to demean X before we perform SVD we cannot state that the variances of the compressed semantic space (X'* V or equally U*S) are exactly captured by the square of single value (S). In our case of extra sparse matrix over 95 percent of entries in the matrixes is zero we can safely claim that demeaning operation has little difference on the variance of reconstructed semantic space, and the square of ordered single vales indicate the significance of single vectors, which capture the direction of maximum variance in the syntactic matrix. We decide *l* largest single vales by their contributions in the 1, 000 components, shown as following,

$p_i = s_i^2 / \subset \mathrm{diag}(S^2)$ (wrong, after demeaning it goes right)

where, s_i is the ith single vale in S after SVD, $\mathrm{diag}(S^2)$ is the diagonal vector of square of S.

Everitt and Dunn (2001), Wall, et al. (2003) set up a threshold, 0.7/*l*, where *l* is equal to 1, 000 in our case, to measure the significant level of each single vale to the variance. We established 250 as a fixed size of each semantic space after checking the threshold. With the visualization of this (Figure 2) we can read that the first 2 component groups account for nearly 50 percent of variance, and after that the elbow curve turns to be smoother. We distinguish the first 250 single values that satisfy our need of the threshold, and hold nearly 76 percent of the variance in the syntactic space, as general size of semantic dimension. The other 750 single vales are cancelled, as their corresponding eigenvectors are insignificant features, as far as their rates of variances are concerned.

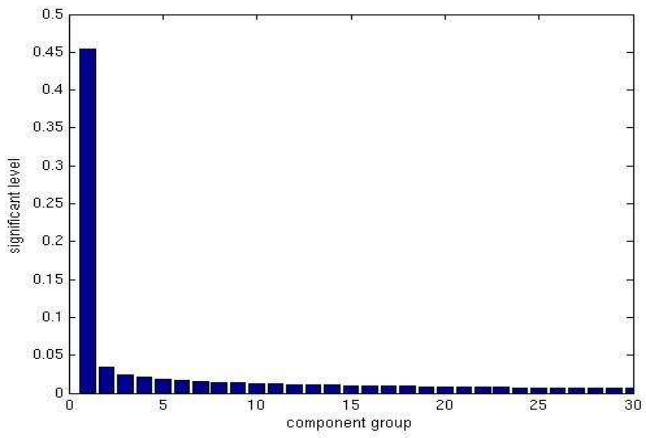

Fig. 2. visualization of significant single values (every 10 single values as a group)

Some similar words

After reducing the dimensions of syntactic space, we should give some results on first 10 similar words we find after comparing all of the words in the pseudo-semantic space. Another thing is about how to evaluate the synonyms list of a target word using distributional similarity when we can't find a common benchmark set and there are absence of standard valuation to practise (some hiring subjective judgement, some using WordNet or Roget or dictionary to compute the size of intersection between the list and the synonyms or definitions of these knowledge bases).

We want to set up a fixed distance (one link or 2, 3 inks between the words in the taxonomy of WordNet), which will limit the maximum steps between the target word and the word in the list. If we find there exists some kind of links between target word and candidate word (one of similar words to the target word) in a definite distance, we take it as a hit. Therefore, we can derive a hit ratio on a small data set, i.e. 12 words, which is employed by Yarowsky (1995)in his word sense disambiguation work.

In the same time the problem is the objectivity of the WordNet, since some hierarchies are not well organized. For example our distributional strategy can find a similar word, vaccine, or the target, drug. But the WordNet can't find any relation between them in the distance of less than four. In other words it doesn’t mean our distributional similarity can't give the right synonyms or near synonyms of the target word. it is subjected to the correctness of WorldNet.

Table 3. the first 10 similar words of each target in the syntactic space (RAW), psudo semantic space (SVD), and semantic space (ICA)

	RAW	SVD	ICA
tank	Gallon, litre, bucket, vapour, demitasse, tubful, acre-foot, saucepan, vole, bowl	Tanker, boiler, aquarium, container, battery, compartment, reactor, rig, vessel	Pump, aquarium, tanker, boiler,valve, cylinder, pipe, container, catheter, pumps
space	Tastefulness, grate, doorway, desktop, spawner, tuning, manhole, sewer, software, enrolment.	Universe, demister, handwheel, bathroom, room, wardroom, rooms. headroom, sanctum, privacy	Keratitis, courtroom, geum, sumac, gallery, pomposity, chamber, de-stanlinisation, astrodome, cubbyhole.
sake	Haji, visitor, service, professional, care, visiting, hazard, americanization, enrollee,spa	In, nurture, epitome, love, thanks, tote, damn, spirit, now, semblance.	Semblance, detriment, haji, rest, baptism, forty, prospect, secret, subsidisation, smidgeon.
poach	Incubate, live, kill, employ, recruit, know, fertilise, cheer, persuade, help	Sieve, pour, butter, sprinkle, sponge, saute, fry, sizzle, oil, tenderise	Fry, saute, sprinkle, sponge, cook, simmer, pan, macerate, reheat, sieve
plant	Weapon, reactor, warhead, non-proliferation, fission, stations, deterrent, disamament, submarine, holocaust	Solidago, kingcup, species, tolmiea, polyploid, silphium, cattail, phanerogam, maguey, lycopene	Solidago, kingcup, radiolaria, helleborus, dichloride, shrub, marchantia, lycopodiate, nepeta, perennial
palm	Palm, bookbinding, eyeshade, grenade, forefinger, nuthatch,	Fingertip, knee, chin, shoulder, hand, cheek,	Fingertip, forefinger, knee, cheek, wrist, elbow,

	luggage, finishing, gesell, riveter, tiu	breast, pillow, forefinger, forearm	breast, tendon, fist, thigh
motion	Tempo, suffusion, pirouette, fuze, hemiplegic, foxtrot, vote, acclimatization, trickster, accretion	Movement, spin, deuteron, elongation, vibration, stabiliser, current, sensor, imaging, rudder	Spin, velocity, deuteron, linac, amendment, resolution, hysteresis, exosphere, dichroism, spectrogram
duty	Obligation, bioethics, requirement, responsibility, nuisance, enrollee, erythropoietin, provisions, undertaker, provision	Responsibility, authority, jurisdiction, sick, ordinance, elint, subdeacon, prerogative, infraction, custody	Responsibility, obligation, liability, sick, disqualification, excise, post, discretion, assagai, underdog
drug	Medication, elytron, technique, pesticide, gelatin, isoflurane, butanol, peekaboo, preordination, umbo	Medication, antibiotic, therapy, diuretic, contraceptive, chemotherapy, steroid, poison, vaccine, contraception	Ampicillin, antibiotic, diuretic, medication, aminophylline, penicillin, morphine, naloxone regimen, therapy
ax	Axis, dining-hall, xylophone, mullion, bitter, goniometer, lintel, curlew, troll, slab	route, axis, dining-hall, plank, diagonal, groove, airlock, intersection, locus, channel	Axis, blade, groove, runway, octagon, facelift, aisle, vein, shaft, diagonal
bass	Harpo, jacquard, whammy, helix, decker, bogey, fingermark, bind, liposome, bookkeeping	Guitar, harpo, nystagmus, violin, amplifier, whammy, liposome, dabbler, melody, dishwasher	Guitar, harpo, nystagmus, violin, amplifier, liposome, whammy, melody, dabbler, dishwasher
crane	Zeta, why, threshing, dredger, locomotive, naphtha, engine, tugboat, projector, oh	Falcon, mainsail, tanker, truck, hatchback, tractor, diesel, speedboat, dredger, hammer	Kettle, falcon, tanker, bream, wheels, waterwheel, gull, denominator, octopus, slug

Table 4. the related word found. Syn: synoynm; 1: only one link between the target word and all its 10 similar words. (distance is 1); 2: 2 links >3: more than 3 links. All: total links between the target word and all its 10 similar words.

	RAW					SVD					ICA				
	syn	1	2	>3	all	syn	1	2	>3	all	syn	1	2	>3	all
tank			3	6	9		2	2	5	9		1	2	1	4
space				7	7				7	7				5	5
sake				8	8				7	7				2	2
poach															
plant												1		7	8
palm				8	8										
motion				7	7										
duty	2	1	1		4										
drug		1		2	3										
ax	1			4	5										
bass				5	5										
crane				3	3										
total															

65 pairs of good enough word pairs.

To evaluate finding word association is not a relaxing task. There is no consensus on the standard for evaluation of lexical similarity. Plenty of researchers validate their word similarity metrics with human judgements on the same dataset (cf. (Yang and Powers 2005)). This dataset is from psycholinguistic test performed by Rubenstein and Goodenough (Rubenstein and Goodenough 1965), on judging synonymy of word pairs, where they hired 51 subjects, two groups of college undergraduates, to evaluate 65 pairs of nouns by assigning a similarity from 0 to 4. we still employ the dataset to access extent to what we can find a pair of words are similar.

In this task, we test 3 kinds of space: syntactic space where data entry is simply coroccurence frequency without any other preprocessing, pseudo semantic subspace after SVD, and semantic subspace through ICA. We also list edge-counting method proposed by (Yang and Powers 2005) in their measuring word similarity using taxonomy of wordnet, which perform best in the current literature, to investigate what and how they are different. The result is shown in Table 3, which indicates that SVD (pseudo semantic space) and ICA (semantic space) has nearly same correlation with the human judgement (i.e. r = 0.695), which are both higher than syntactic dependency space.

Table 5. on the 65 pairs of words the result of different techniques, including edge-counting, syntactic dependency, SVD and ICA. Also lists the correlation with human judgement in the last row.

Scores / Word pairs		Human	Y&P	SYN	SVD	ICA
gem	jewel	3.94	0.9	0.1023	0.2576	0.2578
midday	noon	3.94	0.9	0.1304	0.4517	0.4522
automobile	car	3.92	0.9	0.0799	0.1332	0.1334
cemetery	graveyard	3.88	0.9	0.2146	0.2623	0.2656
cushion	pillow	3.84	0.85	0.1451	0.2654	0.2675
boy	lad	3.82	0.85	0.6337	0.5543	0.5547
cock	rooster	3.68	0.9	0.0666	0.244	0.2479
implement	tool	3.66	0.85	0.1501	0.3318	0.333
forest	woodland	3.65	0.9	0.2270	0.5633	0.5644

coast	shore	3.6	0.85	0.4739	0.6567	0.6569
autograph	signature	3.59	0.85	0.0047	0.0349	0.0347
journey	voyage	3.58	0.85	0.4685	0.4669	0.4675
serf	slave	3.46	0.417	0.0952	0.1529	0.1526
grin	smile	3.46	0.9	0.5755	0.7754	0.7753
glass	tumbler	3.45	0.85	0.4082	0.6533	0.6537
cord	string	3.41	0.85	0.0397	0.326	0.327
hill	mound	3.29	0.9	0.1260	0.2677	0.2727
magician	wizard	3.21	0.9	0.0901	0.2651	0.2682
furnace	stove	3.11	0.595	0.0682	0.4182	0.4204
asylum	madhouse	3.04	0.85	0.0367	0.2192	0.2202
brother	monk	2.74	0.85	0.0523	0.0737	0.0761
food	fruit	2.69	0.417	0.2737	0.417	0.4178
bird	cock	2.63	0.417	0.1419	0.0849	0.0894
bird	crane	2.63	0.85	0.1221	0.3033	0.3051
oracle	sage	2.61	0.204	0.0219	0.0556	0.0561
sage	wizard	2.46	0.204	0.0439	0.1083	0.1103
brother	lad	2.41	0.292	0.1237	0.1298	0.1311
crane	implement	2.37	0.292	0.0214	0.1094	0.1119
magician	oracle	1.82	0.143	0.0057	-0.018	-0.017
glass	jewel	1.78	0.143	0.0252	0.057	0.0578
cemetery	mound	1.69	0.034	0.1100	0.0972	0.1024
car	journey	1.55	0	0.0770	0.0211	0.0214
hill	woodland	1.48	0.204	0.0701	0.088	0.0915
crane	rooster	1.41	0.1	0.0057	0.0866	0.0891
furnace	impleme	1.3	0.20	0.020	0.241	0.243

	nt	7	4	7	7	9
coast	hill	1.26	0.292	0.1766	0.1146	0.1157
bird	woodland	1.24	0.07	0.1183	0.0266	0.0287
shore	voyage	1.22	0	0.1829	0.3679	0.3686
cemetery	woodland	1.18	0.049	0.0896	0.1494	0.1521
food	rooster	1.09	0	0.0179	0.0236	0.025
forest	graveyard	1	0.049	0.0637	0.1322	0.1352
lad	wizard	0.99	0.292	0.0504	0.0999	0.1023
mound	shore	0.97	0.292	0.0541	0.0042	0.0069
automobile	cushion	0.97	0.417	0.0625	0.1656	0.168
boy	sage	0.96	0.204	0.0285	0.0242	0.0253
monk	oracle	0.91	0.1	0.0328	0.0874	0.0882
shore	woodland	0.9	0.204	0.0776	0.1072	0.1085
grin	lad	0.88	0	0.0475	0.0674	0.068
coast	forest	0.85	0.143	0.0737	0.1389	0.1396
asylum	cemetery	0.79	0.024	0.0395	0.1388	0.1412
monk	slave	0.57	0.292	0.0642	0.1525	0.1556
cushion	jewel	0.45	0.143	0.0420	0.175	0.176
boy	rooster	0.44	0	0.1987	0.0072	0.0087
glass	magician	0.44	0.1	0.0231	0.0212	0.0242
graveyard	madhouse	0.42	0	0.0347	0.1317	0.1334
asylum	monk	0.39	0.034	0.0194	-0.021	-0.017
asylum	fruit	0.19	0.143	0.0084	0.0037	0.0061
grin	implement	0.18	0	0.0090	-0.03	-0.029
mound	stove	0.14	0.143	0.1236	0.1058	0.1113
automobile	wizard	0.11	0	0.0145	-0.019	-0.017

autograph	shore	0.06	0	0.0010	-0.085	-0.085
fruit	furnace	0.05	0.143	0.0086	-0.057	-0.054
noon	string	0.04	0	0.0229	-0.009	-0.007
rooster	voyage	0.04	0	0	-0.052	-0.05
cord	smile	0.02	0	0.0143	-0.029	-0.027
correlation			0.897	0.5096	0.6955	0.695

Discussion

For the 65 pairs words, ICA failed to improve scoring the similarity of words after the projecting of original syntactic space with SVD. Yang and Powers (Yang and Powers 2005) employed Wilcoxon signed-rank test as an alternative to the t-test to inspect the difference of correlation values of similarity metrics, since they score the dataset with different scale or equal-interval. Here we still repeat such significance test to analyse whether their differences in each space are significant. The one-tailed Wilcoxon sign-ranked test (at 95% of confident level) shows us that Y&D is significantly better than SYN ($p = 0.002$), SVD ($p = 0.014$) and ICA ($p = 0.021$). SVD and ICA are at near significant level to SYN (respectively $p = 0.095$ and $p = 0.076$).

Berland, M. and E. Charniak (1999). Finding parts in very large corpora. Proceedings of the 37th conference on Association for Computational Linguistics, College Park, Maryland.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. Proceedings of the 37th conference on Association for Computational Linguistics.

Clark, H. H. (1969). "Linguistic processes in deductive reasoning." Psychological Review **76**: 387-404.

Clark, H. H. and S. K. Card (1969). "Role of semantics in remembering comparative sentences." Journal of experiment psychology **82**: 545-53.

Dagan, I., S. Marcus, et al. (1993). Contextual word similarity and estimation from sparse data. Proceedings of the 31st conference on Association for Computational Linguistics, Columbus, Ohio.

Deerwester, S. C., S. T. Dumais, et al. (1990). "Indexing by Latent Semantic Analysis." Journal of the American Society of Information Science **41**(6): 391-407.

Deese, J. (1966). The structure of associations in language and thought. Baltimore,, Johns Hopkins Press.

Dominic, W. (2003). A mathematical model for context and word-meaning. the 4th international and interdisciplinary conference on modeling and using context, Stanford, California.

Everitt, B. S. and G. Dunn (2001). Applied Multivariate Data Analysis. London, Armold.

Firth, J. R., Ed. (1957). A synopsis of linguistic theory. Selected Papers of J.R. Firth 1952-1959, Longman.

Gasperin, C., P. Gamallo, et al. (2001). Using Syntactic Contexts for Measuring Word Similarity. Workshop on Semantic Knowledge Acquisition & Categorisation (ESSLLI 2001). Helsinki.

Grefenstette, G. (1992). Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis. 30th annual meeting of the association for computational linguistics.

Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval.

Grefenstette, G. (1993). Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches. Workshop on acquisition of lexical knowledge from text columbus.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics.

Hindle, D. (1990). Noun classification from predicate-argument structures. 28th Annual Meeting of the Association for Computational Linguistics.

Honkela, T. and A. Hyvärinen. (2004). Linguistic Feature Extraction using Independent Component Analysis. Int. Joint Conf. on Neural Networks (IJCNN2004), Budapest, Hungary.

Hyvärinen, A. (1999). "Survey on Independent Component Analysis." Neural Computing Surveys **2**: 94-128.

Leacock, C., G. Towell, et al. (1996). Towards building contextual representations of word senses using statistical models. Corpus processing for lexical acquisition, MIT Press**:** 97-113.

Lesk, M. (1986). Automatic sense diaambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone. the 5th annual international conference on systems documentation, ACM Press.

Lin, D. (1997). Using syntactic dependency as a local context to resolve word sense ambiguity. Proceedings of the 35th annual meeting of the association for computational linguistics, Madrid.

Lin, D. (1998). Automatic retrieval and clustering of similar words. proceedings of the COLING-ACL'98.

Nowak, M. A., J. B. Plotkin, et al. (2000). "The evolution of syntactic communication." Nature **404**.

Pedersen, T., S. Banerjee, et al. (2003). Maximizing Semantic Relatedness to Perform Word Sense Disambiguation.

Pichon, R. and P. Sébillot (1998). "Automatic acquisition of meaning elements for the creation of semantic lexicons." Information Society: 69-72.

Riloff, E. and J. Shepherd (1997). A Corpus-Based Approach for Building Semantic Lexicons. Proceedings of EMNLP-2.

Rubenstein, H. and J. B. Goodenough (1965). "Contextual correlates of synonymy." communications of the ACM **8**(10): 627-633.

Salton, G. (1973). "Comment on "an evaluation of query expansion by the addition of clustered terms for a document retrieval system"." Computing Reviews **14**(232).

Wall, M. E., A. Rechtsteiner, et al. (2003). Singular Value Decomposition and Principal Component Analysis. A Practical Approach to Microarray Data Analysis. D. P. Berrar, W. Dubitzky and M. Granzow, Kluwer:Norwell, MA**:** 91-109.

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

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA.