

# The Dekang Trilogy:

## Investigation of Three Closely Related Papers



Dekang Lin  
*University of Alberta*

Presented to CSC 2528 by Faye Baron and Saif Mohammad

# The Dekang Trilogy

---

- Extracting Collocations from Text Corpora (1998)
- Automatic Retrieval and Clustering of Similar Words (1998)
- Automatic Identification of Non-compositional Phrases (1999)

# Mutual Information

---

- $x$  and  $y$  with probabilities  $P(x)$  and  $P(y)$  have Mutual Information:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- $P(x, y)$  is the joint probability.
- If  $x$  and  $y$  are strongly associated,  $I(x, y) \gg 0$ .
- If  $x$  and  $y$  do not share an interesting relationship,

$$I(x, y) \approx 0.$$

Fano, R. 1961. [Transmission of Information: A Statistical Theory of Communications](#). MIT Press, Cambridge, MA.

# Word Association Ratio

---

- Similar to Mutual Information.
- $x$  and  $y$  are words,  $P(x)$  and  $P(y)$  are word probabilities.  
Estimated by corpus counts.
- $P(x, y)$ : probability that  $x$  appears before  $y$ .  
Estimated by corpus counts.
- $P(x, y)$  not symmetric.  
$$P(x, y) \neq P(y, x)$$

Church, K. and Hanks, P. 1989 [Word Association Norms, Mutual Information and Lexicography](#). In *Proceedings of the 23rd Annual Meeting of the ACL*.

# Dependency Triples

---

- *Dependency triples* are extracted from a corpus using a parser.
- Each **dependency triple** is made up of a **head**, a **dependency type**, and a **modifier**.
- The dependency type represents the relationship between the head and the modifier.

# Dependency Triples – Example

---

From the sentence *I have a brown dog* we get the following triples:

(have V:subj:N I)

(have V:compl:N dog)

(dog N:jnab:A brown)

(dog N:det:D a)

# Dependency Types

---

Label	Relationship between:
N:det:D	a noun and its determiner
N:jnab:A	a noun and its adjectival modifier
N:nn:N	a noun and its nominal modifier
V:compl:N	a verb and its noun object
V:subj:N	a verb and its subject
V:jvab:A	a verb and its adverbial modifier

# Collocations

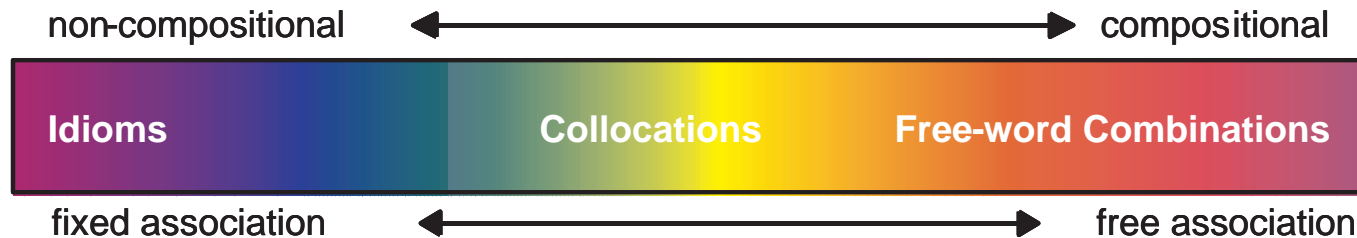
---

- There is no single clear accepted definition.
- Dekang refers to a collocation simply as *a habitual word combination*.
- McKeown and Radev describe a spectrum of association types between word combinations, from which three differentiable word combination categories emerge.

Kathleen R. McKeown and Dragomir R. Radev. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *A Handbook of Natural Language Processing*. Marcel Dekker, 2000.

# Spectrum of Word Combinations

---



Three categories emerge from the spectrum of association types between word combinations:

- free word combinations
- collocations
- idioms

# Free Word Combinations

---

- Combinations of words whose parts are **open-ended** and can be freely combined with other words.
- For example:
  - ▶ *to take the bus*
  - ▶ *the end of the road*
  - ▶ *to buy a house*

# Collocations

---

- *A group of words that occur together more often than by chance.*
- Parts are **fixed** — Substituting a synonym for one of the words may create an awkward expression.
- Expressions are **compositional** — The meaning of each individual word contributes to the expression meaning.
- For example:
  - ▶ *to trade actively*
  - ▶ *the Dow Jones Industrial Average fell (**number**) points*
  - ▶ *to file a lawsuit*
  - ▶ *orthogonal projection*

# Idioms

---

- **Strictly fixed** expressions — Substituting a synonym for one of the words would change the meaning.
- Expressions are **non-compositional** — The meaning of the expression cannot be derived from the meaning of the words from which the expression is composed.
- For example:
  - ▶ *to kick the bucket*
  - ▶ *par for the course*
  - ▶ *to catch up*
  - ▶ *dead end*

# Approaches to Collocation Extractions

---

- Choueka, Klein and Neuwitz. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *ALLC Journal*, volume 4.
- Church and Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* 16(1).
- Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1).

# Approaches to Collocation Extractions

---

- Pearce. 2002. A comparative evaluation of collocation extraction techniques. *International Conference on Language Resources and Evaluation*
- Richardson. 1997. *Determining Similarity and Inferring Relations in a Lexical Knowledge Base*. Ph.D. Thesis, The City University of New York.

# Extracting Collocations

---

- The experiment was conducted on the 55-million-word Wall Street Journal and the 45-million-word San Jose Mercury.
- The corpora were parsed and dependency triples generated.
- A correction algorithm was applied to improve the reliability of the parsed triples.
- Mutual Information was applied to the dependency triples to *weed out coincidences* and retain only collocations.
- Used a term bank to evaluate coverage of the collocations.

# Algorithm to Correct Parsed Triples

---

- Assumes triples are generally *right* more than *wrong*.
- Applies correction rules to switch relationship classification of potentially *wrong* triples.
- A correction rule consists of a threshold  $\theta$ , and a pair of dependency types  $(rel, rel')$  that can be confused. (*e.g.*, verb-object, noun-noun.)
- Thirty correction rules were used.

# Algorithm to Correct Parsed Triples

---

- If both  $(w_1, rel, w_2)$  and  $(w_1, rel', w_2)$  are found in the corpus and the ratio of their frequency counts is greater than  $\theta$ , then the frequency count of the lower is added to the higher, and then set to zero. It is assumed that the lower frequency triple is an error.
- A manual inspection of 200 randomly selected corrections showed this correction algorithm to be 95% accurate.
- 699,219 pairs of words had multiple dependency relationships.

# Using Mutual Information to Select Collocations

---

A dependency triple  $(w_1, rel, w_2)$  can be regarded as three events:

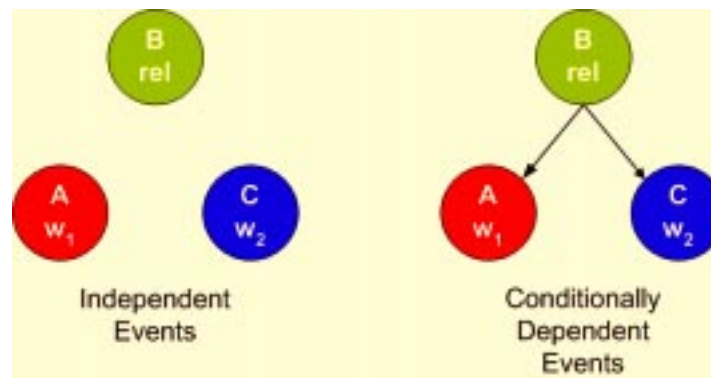
**A:** a randomly selected word is  $w_1$ ,

**B:** a randomly selected dependency type is  $rel$ ,

**C:** a randomly selected word is  $w_2$ .

# Using Mutual Information to Select Collocations

---



Combining the Mutual Information of a triple as defined by Alshawi and Carter (1994) with an assumption of conditional independence and Bayesian Networks (Pearl, 1988) we get:

$$\log \frac{P(A, B, C)}{P(B) \times P(A | B) \times P(C | B)}$$

# Calculating Mutual Information — the Pieces

---

Where:

- $w_1$  represents the first word in the triple
- $rel$  represents the relationship
- $w_2$  represents the second word in the triple
- $*$  is a wild card for all words belonging to triples that match the remainder of the expression
- $|| ||$  is used to denote the frequency count of the contained expression

We get the expressions ...

# Calculating Mutual Information — the Expressions

---

$$P(A, B, C) = \frac{\|w_1, rel, w_2\| - c}{\|*, *, *\|}$$

$$P(B) = \frac{\|*, rel, *\|}{\|*, *, *\|}$$

$$P(A | B) = \frac{\|w_1, rel, *\|}{\|*, rel, *\|}$$

$$P(C | B) = \frac{\|*, rel, w_2\|}{\|*, rel, *\|}$$

get plugged into:  $\log \frac{P(A, B, C)}{P(B) \times P(A | B) \times P(C | B)}$

and reduced to:  $\log \frac{\|w_1, rel, w_2\| - c \times \|*, rel, *\|}{\|w_1, rel, *\| \times \|*, rel, w_2\|}$

# Evaluation

---

- Church and Hanks (1990), and Choueka (1988) did not measure coverage.
- Alshawi and Carter (1994) scored their collocations by their use in parse tree selection.
- Dekang mimicked this technique using the Susanne corpus containing parse trees of 64 texts in the Brown Corpus of American English.

# Evaluation

---

- First he converted constituency parse trees into dependency trees.
- Then he extracted dependency triples for collocations containing the relations: subject-verb, verb-object, adjective-noun, noun-noun.
- Only triples with duplicate occurrences were kept as collocations.

# Evaluation

---

- For each of the *Susanne-triples*, if the triple was extracted with the identical two words and relationship, it was considered *correct*. If the words matched but the relationship was wrong, it was considered *incorrect*. Otherwise, it was considered *additional*

# Evaluation

---

$$\textit{coverage} = \frac{\textit{correct}}{\textit{recurring}}$$

$$\textit{precision} = \frac{\textit{correct}}{\textit{correct} + \textit{incorrect} + \textit{additional}}$$

Brown Corpus categories include:

**A** press reportage

**G** belles lettres, biography, memoirs, etc.

**J** *learned* (technical and scholarly prose)

**N** adventure and Western fiction

# Evaluation Results

---

	A	G	J	N
recurring	548	268	592	256
correct	358	147	164	139
incorrect	5	2	4	5
additional	0	1	4	0
coverage	65.3%	54.9%	27.7%	54.2%
precision	98.6%	98.6%	97.6%	96.4%

# Preliminary Conclusions

---

- I question how Dekang can call this triple extraction *collocation extraction* when at most, two-word, non-contiguous expressions are extracted.
- Perhaps this triple extraction would, if combined with the theory of Smadja and his predecessors, yield more lengthy and meaningful collocations.

# Semantic Relatedness

*Antarctica - penguin*

*Antarctica - table*

Which pair is *more* related?

Lin [Aug, 1998] provides a strictly corpus based measure of similarity which utilizes information triples and word association ratio.

Lin, D. 1998. *Automatic Retrieval and Clustering of Similar Words*. COLING-ACL98, Montreal, Canada, August, 1998.

# Why Semantic Relatedness?

---

- Information retrieval
- Word sense disambiguation
- Spelling correction
- Identifying discourse structure
- Text summarization
- Word prediction
- Automatic thesaurus creation

# Source of Information and Measures

---

- Ontologies and semantic networks

Roget's Thesaurus, WordNet

Rada [1989], Morris and Hirst [1991], Hirst and St-Onge [1998]

- Text corpus

SemCor, Brown, BNC

Hindle [1990], Shütze [1997], [Lin \[Aug, 1998\]](#), Yoshida [2003]

- Combination

Based on information content.

Resnik [1995], Jiang and Conrath [1997], Lin [1997]

# Definition of Similarity

---

- Definition not coupled with any semantic network.
- The similarity measure is provable.

$$sim(A, B) = \frac{I(common(A, B))}{I(description(A, B))}$$

- $common(A, B)$ : proposition that states commonalities between them.
- $description(A, B)$ : proposition that describes  $A$  and  $B$ .
- $I(s)$  is the information contained in proposition  $s$ .

Lin, D. 1998. [An Information-Theoretic Definition of Similarity](#). *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin, July, 1998.

# Intuition

---

- $sim(A, B)$  directly proportional to the commonality between  $A$  and  $B$ .

The more the commonality, more is the similarity.

- $sim(A, B)$  inversely related to difference between  $A$  and  $B$ .

The more the difference, the less  $A$  and  $B$  are similar.

- $A$  and  $B$  are maximally similar when  $A$  and  $B$  are identical, irrespective of the amount of commonality.

# Applied to a Corpus

---

- Counts of triples which have the word (say,  $w_1$ ) constitute the description of  $w_1$ .

$$\| \textit{football}, \textit{obj-of}, \textit{play} \| = 6$$

$$\| \textit{football}, \textit{nmod-of}, \textit{player} \| = 4$$

$$\| \textit{football}, \textit{subj-of}, \textit{bounce} \| = 3$$

- Words  $w_1$  and  $w_2$  may share triples.

$$\| \textit{football}, \textit{obj-of}, \textit{play} \| = 6$$

$$\| \textit{cricket}, \textit{obj-of}, \textit{play} \| = 8$$

Both triples match the pattern  $(*, \textit{obj-of}, \textit{play})$ .

# Commonality

---

- Let  $I(w, r, w')$  denote information in:

$$\|w, r, w'\| = c$$

- Lin defines  $common(w_1, w_2)$  as:

$$\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))$$

- $T(w)$  is the set of all pairs  $(r, w')$  such that  $I(w, r, w')$  is positive.

# Description and Similarity

---

- Description of  $w$ :

$$\sum_{(r,w') \in T(w)} I(w, r, w')$$

- Similarity between  $w_1$  and  $w_2$ :

$$\begin{aligned} \text{sim}(w_1, w_2) &= \frac{I(\text{common}(w_1, w_2))}{I(\text{description}(w_1, w_2))} \\ &= \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w') \in T(w_1)} I(w_1, r, w') + \sum_{(r,w') \in T(w_2)} I(w_2, r, w')} \end{aligned}$$

# Distributional Similarity

---

- The notion that words which usually have similar contexts are *similar* is known as **Distributional Similarity**.

- Context is represented as counts of triples.

Hindle [1990], Lin [Aug, 1998]

- Context represented by co-occurrences.

Shütze [1997], Dagan [1997], Yoshida [2003]

# Other Corpus-Based Measures

---

$$\begin{aligned} H & : \sum_{(r,w) \in T(w_1) \cap T(w_2) \wedge r \in \{\text{subj-of}, \text{obj-of}\}} \min(I(w_1, r, w), I(w_2, r, w)) \\ H_r & : \sum_{(r,w) \in T(w_1) \cap T(w_2)} \min(I(w_1, r, w), I(w_2, r, w)) \\ \text{cos} & : \frac{|T(w_1) \cap T(w_2)|}{(|T(w_1)| \times |T(w_2)|)^{1/2}} \\ \text{Lin} & : \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w') \in T(w)} I(w, r, w') + \sum_{(r,w') \in T(w)} I(w, r, w')} \end{aligned}$$

# How about ...

---

$$H_r : \sum_{(r,w) \in T(w_1) \cap T(w_2)} \min(I(w_1, r, w), I(w_2, r, w))$$

$$sim_{Lin} : \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w') \in T(w)} I(w, r, w') + \sum_{(r,w') \in T(w)} I(w, r, w')}$$

$$sim_{New} : \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} 2 \times \min(I(w_1, r, w), I(w_2, r, w))}{\sum_{(r,w') \in T(w)} I(w, r, w') + \sum_{(r,w') \in T(w)} I(w, r, w')}$$

# Example

---

- Let  $w_1$  and  $w_2$  share a triple of form:  $(*, r, w_3)$
- Let  $I(w_1, r, w_3) \gg 1$  (say,  $x$ ) and  $I(w_2, r, w_3) \approx 1$ .

- Addition to denominator  $MI_1 + MI_2$ :

$$sim_{Lin}: \approx x + 1 \quad sim_{New}: \approx x + 1$$

- Addition to numerator:

$$sim_{Lin}: MI_1 + MI_2 \approx x + 1$$

$$sim_{New}: 2 \times \min(MI_1, MI_2) \approx 2$$

# Thesaurus Creation

---

- Applied to 64 million corpus from of the WSJ, San Jose Mercury and AP Newswire.

5469 nouns, 2173 verbs and 2632 adjectives/adverbs

occurred at least a 100 times in parsed corpus.

- Pairwise similarity between all nouns, verbs and adjectives computed.
- Thesaurus entry for each word ( $w$ ) created.

$w(pos) : w_1 s_1, w_2 s_2, \dots, w_n s_n$

$w_i$  is a word,  $s_i = sim(w, w_i)$

# Sample Entries

---

**brief (noun):** affidavit 0.13, petition 0.05, memorandum 0.05, motion 0.05, lawsuit 0.05, deposition 0.05, slight 0.05, prospectus 0.04, document 0.04, paper 0.04

**brief (verb):** tell 0.09, urge 0.07, ask 0.07, meet 0.06, appoint 0.06, elect 0.05, name 0.05, name 0.05, empower 0.05, summon 0.05, overrule 0.04

- Set of thesaurus entries looks good.
- Notice the very small similarity values.

# RNN

---

- Two words are a pair of **Respective Nearest Neighbors** if each is the other's most similar word.

## **Reciprocally Nearest Neighbors - Hindle**

543 noun, 212 verb and 382 adjective RNN pairs found.

N: earnings - profit, plan - proposal, share - stock

V: fall - rise, injure - kill, concern - worry

A: high - low, bad - good, extremely - very

# Evaluation

---

- Two thesauri created automatically from WordNet (Lin, 1997) and Roget's thesaurus (a simple cosine method).  
Entries just as the corpus-based one.
- A measure defined to calculate similarity of two entries for the same word in different thesauri.
- Average similarity of corpus based methods, with that of Roget and WordNet calculated.

# Similarity: Thesauri Entries

---

- A pair of thesauri entries:

$$w(pos) : w_1 s_1, w_2 s_2, \dots, w_n s_N$$

$$w'(pos) : w'_1 s_1, w'_2 s_2, \dots, w'_n s_N$$

- Their Similarity:

$$\frac{\sum_{w_i=w'_j} s_i s'_j}{\left( \left( \sum_{i=1}^N s_i^2 \right) \left( \sum_{j=1}^N s'_j{}^2 \right) \right)^{1/2}}$$

Yoshida et. al. [2003] suggest a method that involves root mean squares of the differences in  $s_i$  and  $s_j$ .

# Comparison with WordNet-Based Thesaurus

---

	<b>Average</b>	<b>Std. devn.</b>
<i>Roget</i>	0.178	0.0016
<i>sim<sub>Lin</sub></i>	0.212	0.0015
<i>Hindle</i>	0.204	0.0014
<i>Hindle<sub>r</sub></i>	0.165	0.0012
<i>cosine</i>	0.199	0.0014

# Comparison with Roget-Based Thesaurus

---

	<b>Average</b>	<b>Std. devn.</b>
<i>WordNet</i>	0.178	0.0016
<i>sim<sub>Lin</sub></i>	0.149	0.0014
<i>Hindle</i>	0.147	0.0014
<i>Hindle<sub>r</sub></i>	0.115	0.0011
<i>cosine</i>	0.136	0.0013

# Assumption and Inferences

---

- Assumption: The more a thesaurus is similar to WordNet and Roget, better is the thesaurus.
- $sim_{Lin}$ ,  $Hindle_r$  and  $cosine$  thesauri more similar to WordNet than Roget.
- Use of all syntactic relations better than just verb-obj and subj-obj.  
*Hindle<sub>r</sub>* better than *Hindle*.
- $sim_{Lin}$  is better than  $Hindle_r$  which is better than  $cosine$ .

# Interesting Questions!

---

- The correlation coefficient of the similarity values with Rubenstein-Goodenough (65 pairs) and Miller-Charles (30 pairs).
- Applications of the measure to other problems such as spelling correction and word sense disambiguation.
- Are all syntactic relations helpful?

# Susceptibility to Strong Collocations

---

- Comparison of word pairs such that one or both have strong collocations will result in very low scores.
- Consider closely related words *merry* and *playful*.
  - merry - christmas*, is expected to have a very high MI.
  - playful - christmas*, not likely to appear in corpus.
- Score added to numerator of sim: 0
- Score added to denominator: a large MI value.
- Result: Low similarity value.

# Concluding Remarks on Lin's Similarity Paper

---

- Provides a strictly corpus based measure of similarity with similarities ranging from 0 to 1.

Even the best RNN's have similarities around 0.5.

- Applied it to create a thesaurus and proposed a method to evaluate automatically created thesauri.
- Correlation with human judgments worth determining.
- Efforts on what kinds of relations to use and making the measure resilient to strong collocational effects, worthwhile.

# Automatic Identification of Non-compositional Phrases

---

- Dekang leverages his work on extracting *collocations* and creating a *thesaurus of similar terms* to extract *non-compositional phrases* or *idioms*.
- Builds on the work of Tapanainen et al. (1998)

Pasi Tapanainen, Jussi Piitulainen, and Timo Jävinen. 1998. Idiomatic object usage and support verbs. In *Proceedings of COLING/ACL-98*

# Related Work — Tapanainen, Piitulainen, & Jävinen

---

- The basic idea is that *if an object appears only with one verb (of few verbs) in a larger corpus we expect that it has an idiomatic nature.*
- Looks at the  $V : \text{comp1} : N$  relationship only.

Where  $DF(o)$  is the distributional frequency of each object noun  $o$  we get:

$$DF(o) = \sum_{i=1}^n \frac{|v_i, V : \text{comp1} : N, o|^a}{n^b}$$

# Observe Idiom Characteristics

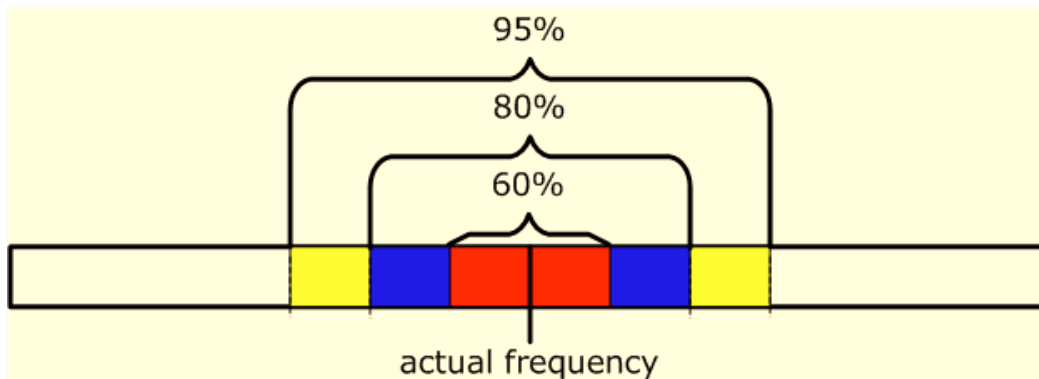
---

Looking at the idiom *red tape*

adjective	noun	freq	mutual info
red	tape	259	5.87
yellow	tape	12	3.75
orange	tape	2	2.64
black	tape	9	1.07

# Confidence Intervals

- The maximum likelihood estimate of the probability of a triple is  $MLE(A, B, C) = \frac{\|w_{1,rel}, w_2\|}{\|*,*,*\|}$
- To increase certainty of the prediction, expand the range of expected frequency. The larger the interval, the more likely that the frequency will fall within that range.



# Confidence Intervals — Mutual Information Ranges

- The range expansion is estimated as  $\frac{\|w_1, rel, w_2\| \pm z_N \sqrt{\|w_1, rel, w_2\|}}{\|*, *, *\|}$
- Dekang looks at the 95% confidence interval

Sample  $z_N$  values

N%	50%	80%	90%	95%	98%	99%
$z_N$	0.67	1.28	1.64	1.96	2.33	2.58

Incorporate in Mutual Information Calculation to get:

$$\log \frac{(\|w_1, rel, w_2\| \pm z_N \sqrt{\|w_1, rel, w_2\|}) \times \|*, rel, *\|}{\|w_1, rel, *\| \times \|*, rel, w_2\|}$$

# Rule for Identifying Non-compositional Phrases

---

A collocation  $\alpha$  is non-compositional if there **does not** exist another collocation  $\beta$  such that:

- (a)  $\beta$  is obtained by substituting the head or the modifier in  $\alpha$  with a similar word, and
- (b) There is an overlap between the 95% confidence interval of the mutual information values of  $\alpha$  and  $\beta$ .

# Example of Rule Application

---

	freq	mutual	lower	upper
verb-object	count	info	bound	bound
make difference	1489	2.928	2.876	2.978
make change	1779	2.194	2.146	2.239

# Evaluation

---

- There is no established measure of success for idiom retrieval.
- Dekang uses NTC's English Idioms Dictionary (NTC-EID) as the gold standard.
- Retrieves all idioms whose head begins with one of {*have, company, make, do, take, path, lock, resort, column, gulf*}.
- Also compare Longman Dictionary of English Idioms (LDOEI) with the gold standard to illustrate the agreement between manually compiled dictionaries.

# Evaluation — Results

---

Evaluated against NTC-EID

	Precision	Recall	Parse Errors
Dekang	15.7%	13.7%	9.7%
LDOEI	39.4%	20.9%	N.A

# The Dekang Trilogy

---

- Dekang has combined the well-known *mutual information theory* with *Dependency triples* to create new extraction techniques.
- Though they seem amazingly simple, they are in fact amazingly powerful.
- It would be useful to investigate how they can be expanded to become more useful tools.