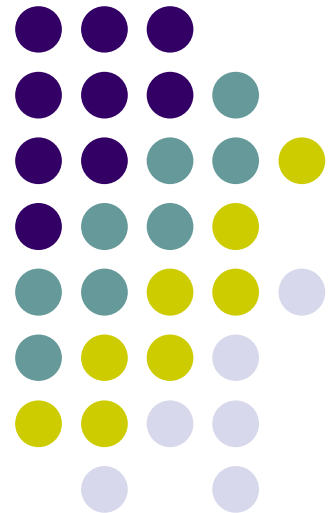
MTP First Stage Presentation

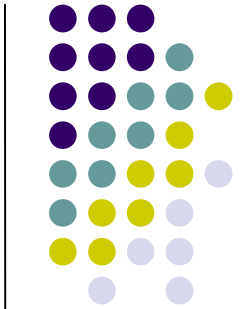
Multiword Expression Recognition

Anoop Kunchukuttan
Roll No: 06305407

Guide: Prof. Om Damani
Examiner: Prof. Pushpak Bhattacharyya



Outline



- What are Multi Word Expressions (MWE) ?
- Why care about MWEs ?
- MWE Characteristics & Classification
- MWE Extraction Methods
- MWE Extraction Evaluation
- Concluding remarks
- Problem Definition

What is a Multi Word Expression?

- A language word - lexical unit in the language that stands for a concept.

  e.g. *train, water, ability*
- However, that may not be true.

  e.g. *Prime Minister*
- Due to institutionalized usage, we tend to think of '*Prime Minister*' as a single concept.
- Here the concept crosses word boundaries.

Defining a Multi Word Expression

A Psycholinguistic Perspective

A sequence, continuous or discontinuous, of words or other elements, which is or appears to be prefabricated: that is stored and retrieved whole from memory at the time from use, rather than being subject to generation or analysis by language grammar.

Defining a Multi Word Expression

- Simply put, a multiword expression (MWE):
  a. crosses word boundaries
  b. is lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic

- E.g. *traffic signal, Real Madrid, green card, fall asleep, leave a mark, ate up, figured out, kick the bucket, spill the beans, ad hoc.*

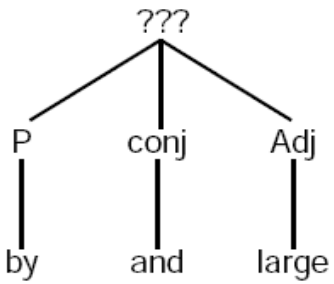
Idiosyncrasies elaborated

- **Statistical idiosyncracies**
  - Usage of the multiword has been conventionalized, though it is still semantically decomposable
  - E.g. *traffic signal, good morning*
- **Lexical idiosyncrasies**
  - Lexical items generally not seen in the language, probably borrowed from other languages
  - E.g. *ad hoc, ad hominem*

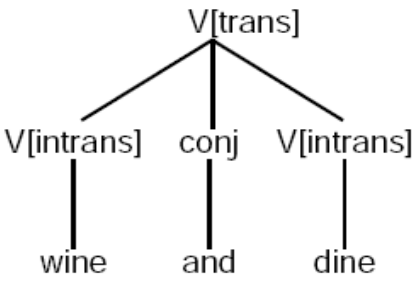
Idiosyncrasies elaborated (2)

- **Syntactic idiosyncrasy**

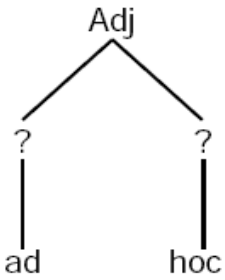
by and large



wine and dine



ad hoc



Conventional grammar rules don't hold, these multiwords exhibit peculiar syntactic behaviour

Idiosyncrasies elaborated (3)

Semantic Idiosyncrasy

- The meaning of the multi word is not completely composable from those of its constituents
- This arises from figurative or metaphorical usage
- The degree of compositionality varies
- E.g. *blow hot and cold* – keep changing opinions

  spill the beans – reveal secret

  run for office – contest for an official post.

Not a binary distinction

- MWEness is not a binary distinction
- Various levels of semantic compositionality
  - *let the cat out of the bag*
  - *lend a helping hand*
  - *fall asleep*
- Even human annotators may disagree

Why care about MWEs?

- A large fraction of words in English are MWEs (41% in Wordnet). Other languages too exhibit this behaviour.
- Conventional grammars and parsers fail.

  eg. *by and large* and compound nouns
- Semantic interpretation not possible through compositional methods
- Pains for machine translation – word by word translation will not work
- New terminology in various domains likely to be multi word. Implications for information extraction
- In IR, multiword queries mean multiword indexing

MWE processing tasks

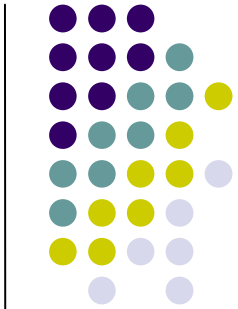
- **Extraction of MWE from corpus**
- Development of MWE lexicon and its representation
- Grammar formalisms for incorporating MWE required to provide robust grammars
- Semantic interpretation, role labelling of MWEs

Subject of this work: MWE extraction

- Will pave the way for lexicon representation and grammar incorporation
- An MWE lexicon will help research in the area

MWE Characteristics

Basis for MWE extraction

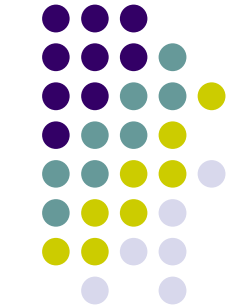


- Non-Compositionality
  - Non-decomposable – e.g. *blow hot and cold*
  - Partially decomposable – e.g. *spill the beans*

- Syntactic Flexibility
  - Can undergo inflections, insertions, passivizations
    e.g. *promise(d/s) him the moon*
  - The more non-compositional the phrase, the less syntactically flexible it is

MWE Characteristics (2)

Basis for MWE extraction



- Substitutability
  - MWEs resist substitution of their constituents by similar words

    E.g. '*many thanks*' cannot be expressed as '*several thanks*' or '*many gratitudes*'
- Institutionalization
  - Results in statistical significance of collocations
- Paraphrasability
  - Sometimes it is possible to replace the MWE by a single word

    E.g. *leave out* replaced by *omit*

Classifying Multi Word Expressions

Based on syntactic forms and compositionality

- Institutionalized Noun collocations

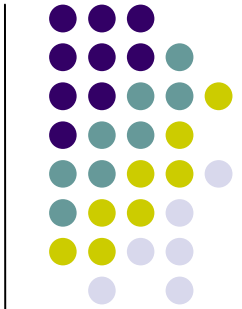
  E.g. *traffic signal, George Bush, green card*
- Phrasal Verbs (Verb-Particle constructions)

  E.g. *call up, eat up*
- Light verb constructions (V-N collocations)

  E.g. *fall asleep, give a demo*
- Verb Phrase Idioms

  E.g. *sweep under the rug*

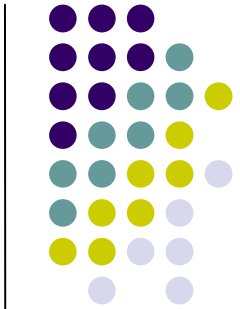
Extracting Multi Word Expressions



Basic Tasks

- Extract Collocations
  - Statistical evidence of institutionalization
  - Use of hypothesis testing
  - Maintain reasonably high recall
- Establish linguistic validity of collocation
  - Not all collocations make linguistic sense
  - Use filters to remove invalid collocations
- Measure semantic decompositionality of the MWE
  - Semantic idiosyncrasy an important characteristic of MWEness

Extracting Multi Word Expressions



Basic Tasks

- Extract Collocations
- Establish linguistic validity of collocation
- Measure semantic decompositionality of the MWE

Pointwise Mutual Information (Church '90)

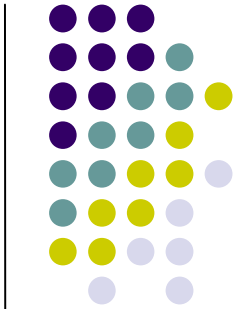
Pointwise Mutual information between words x and y

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

where, (x,y) is word pair being tested.
I(x,y) is the Pointwise Mutual Information between them

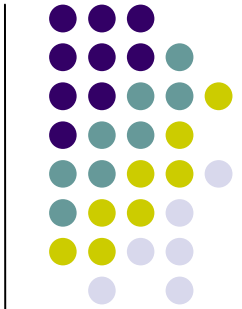
- The Pointwise Mutual Information between two words is a measure of the strength of their collocation.
- Window size determines flexibility/precision trade-off
- Overestimation of rare collocations, no notion of support
- Requires large corpus
- A good initial filter for selecting collocations

Pearson's chi-square test



- A statistical test of independence
- Based on assumption of normal distribution of word frequency, which could be a limitation
- Null hypothesis: the words are independent of each other.
- Higher the value of the chi-square statistic, the stronger the association between the words
- For small data collections, assumptions of normality and chi-square distribution do not hold. Hence, large corpus required

Pearson's chi-square test (2)
The Method



Make a contingency table of frequency counts

W_1,W_2	$W_1,\sim W_2$
$\sim W_1, W_2$	$\sim W_1, \sim W_2$

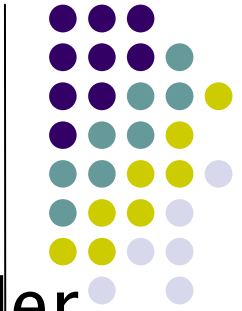
W_1,W_2 : number of times W1,W2 occurs together
$W_1,\sim W_2$: number of times W1 is not followed by W2
$\sim W_1, W_2$: number of times W1 does not precede W2
$\sim W_1, \sim W_2$: frequency of collocations containing none

Now,
O_{ij}=observed frequency in the table
E_{ij}= Expected frequency in each cell when W1 - W2 occur together by chance.
Expected frequency on each cell is equal to (row total * column total) / grand total

Now the chi-square statistic calculated below can be compared against the critical value

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Log Likelihood Ratio (Dunning '93)



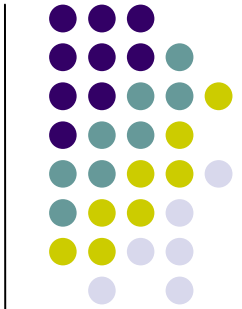
- Uses the log-likehood ratio hypothesis test, under the assumption of binary distribution of word frequency
- Null hypothesis (w2 independent of w1),

  H1: $P(w_2| w_1)=P(w_2|\sim w_1)$

  Alternate hypothesis (w2 depends on w1)

  H2: $P(w_2| w_1)\neq P(w_2|\sim w_1)$
- Can detect collocation in a small corpus too
- The quantity $-2*\log \lambda$ gives an indication of the collocation
  - asymptotically chi-square distributed.

Log Likelihood Ratio (2)
The Method



The log-likelihood ratio calculated as

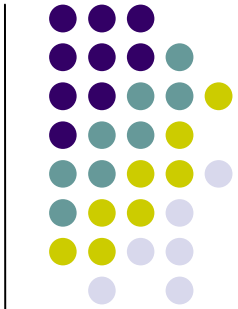
$$\lambda = \frac{max_{\omega \in \Omega_0} H(w;k)}{max_{\omega \in \Omega} H(w;k)}$$

The likelihood of the observed frequency of w2

$$H(p_1, p_2; k_1, n_1, k_2, n_2) = p_1^{k_1}(1-p_1)^{n_1-k_1} p_2^{k_2}(1-p_2)^{n_2-k_2}.$$

The following are the quantities involved
$p_1 = P(w_2|w_1)$, $p_2 = P(w_2|\sim w_1)$, $n_1 = c_1$, $k_1 = c_{12}$
$n_2 = n - c_1$, $k_2 = c_2 - c_{12}$
c_1, c_2, c_{12} =corpus frequencies of w_1,w_2,w_1w_2
n=total number of words in the corpus
For the alternate hypothesis, the MLE estimates of p1, p2 are,
$p_1 = k_1/n_1$ and $p_2 = k_2/n_2$
For the null hypothesis, we have $p_1 = p_2 = p$.
$p = (k_1 + k_2)/(n_1 + n_s)$

Expectation/Variance based measure (Smadja '93)



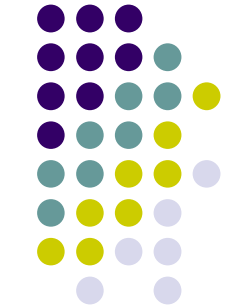
- Consider a fixed size window around every word
- For every word *w*, count frequency f_i of all words w_i in a neighbourhood window.(w,w_i) are candidate collocation pairs.
- For every pair (w,w_i), count the number of occurences p_{ij} at any position *j* in window of *w*.
- Now apply the following tests

- Strength: Check if the collocation has high association

$$k_i = \frac{f_i - \bar{f}}{\sigma}$$

$$k_i > k_0$$

Expectation/Variance based measures (2)



- Spread: Select spiky distributions, exhibiting skewed distribution of collocate

$$\bar{p}_i = \frac{\sum_j p_{ij}}{\text{wsize}}$$

$$U_i = \frac{(\sum_j p_{ij} - \bar{p}_i)^2}{\text{wsize}}$$

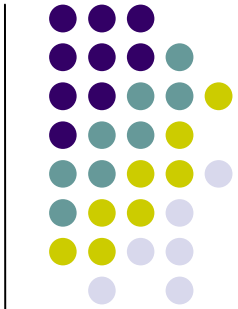
$$U_i > U_0$$

- Peakiness: identify interesting peaks, having minimum frequency support

$$p_{ij} \geq \bar{p}_i + (k_\beta \times \sqrt{U_i})$$

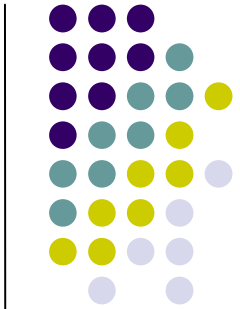
Candidate collocation pairs satisfying these criteria are MWE

Critique



- Large corpus is needed
- Data sparsity
  - N-gram collocations
- Alternative modeling of text
  - Poisson distributions

Extracting Multi Word Expressions



Basic Tasks

- Extract Collocations
- Establish linguistic validity of collocation
- Measure semantic decompositionality of the MWE

Linguistic filters

- Not all kinds of collocations are valid.
  - eg. *the ... of* may pass as a significant collocation, but is linguistically invalid.
- Don't work for syntactically idiosyncratic collocations

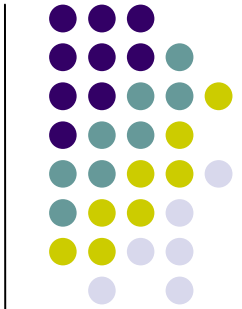
Use of POS tags

- Use POS tags to retain only certain syntactic collocations:
  - Noun-Noun Noun compounds
  - Adjective-Noun Noun compounds
  - Verb-Noun Idioms
  - Verb-Preposition Phrasal verbs
- Burden of handling syntactic variability

Dependency Relations

- Use a parser to identify syntactic dependencies
- The relationship triples from the parse supply potential collocations
  - E.g. *(make,direct_object,light)* is generated for *'make light'*
- Linguistically valid collocations generated
- Structured, principled method.
- Error in the parsing reflects in collocation extraction.

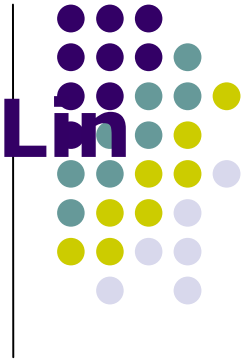
Extracting Multi Word Expressions



Basic Tasks

- Extract Collocations
- Establish linguistic validity of collocation
- Measure semantic decompositionality of the MWE

Substitution by similar words(Lin '99)



- **Key Idea**: If a MWE is semantically non-decomposable, substituting a constituent word with a similar word produces an expression which has different distributional characteristics

  E.g. '*fall asleep*' could be substituted by '*stumble asleep*'
- Measure of non-compositionality,

  Δ = *PMI of the MWE – PMI of substitute collocation*
- Greater the difference between the PMI of the MWE and that of the substitute collocation, the more non-decomposable the MWE is
- Substitute with (a) the most similar word (b) mean PMI of top-k similar words
- It might as well indicate institutionalization

Using Selectional Preferences (Moiron '07)

- **Key Idea**: Verbs have preference for certain nouns as their arguments.
- Analogous to the notion of selectional preference of a verb for a noun class
- The stronger the preference compared to similar nouns, the more likely it an MWE
- Resnik's selectional preference measures adapted
- Data sparsity could be a problem

Using Selectional Preferences (2)

- Resnik's selectional preference measures
  - Strength of association

$$S_v = \sum_n p(n|v) \log \frac{p(n|v)}{p(n)}$$

  - Selectional preference of a verb for a noun

$$A_{v \to n} = \frac{p(n|v) \log \frac{p(n|v)}{p(n)}}{S_v}$$

  - Preference within a certain word cluster

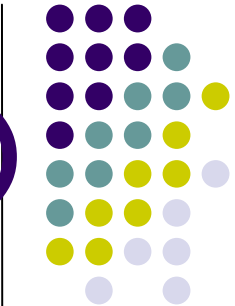
$$R_{v \to n} = \frac{A_{v \to n}}{\sum_{n' \in C} A_{v \to n'}}$$

Measuring Syntactic Fixedness (Fazly '06)

- **Key Idea:** Exploit the fact that idiomatic phrases are less syntactically flexible than compositional phrases. In this work, V-N collocations are considered
- V-N collocations are subject to variations in the form of passivization, determiner type and pluralization.
- Various patterns of variations identified:

Patterns					
v	det:NULL	n_{sg}	v	det:NULL	n_{pl}
v	det:*a/an*	n_{sg}			
v	det:*the*	n_{sg}	v	det:*the*	n_{pl}
v	det:DEM	n_{sg}	v	det:DEM	n_{pl}
v	det:POSS	n_{sg}	v	det:POSS	n_{pl}
v	det:OTHER	[$n_{sg,pl}$]	det:ANY [$n_{sg,pl}$] be $v_{passive}$		

Measuring Syntactic Fixedness (2)



- Estimate the prior probabilty of a pattern over the entire corpus

$$P(pt) = \frac{\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} f(v_i, n_j, pt)}{\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} \sum_{pt_k \in \mathcal{PS}} f(v_i, n_j, pt_k)}$$

- For a given V-N collocation, calculate posterior probability of every pattern

$$P(pt|v, n) = \frac{f(v, n, pt)}{\sum_{pt_k \in \mathcal{PS}} f(v, n, pt_k)}$$

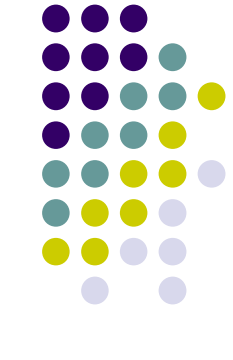
- Calculate the KL divergence between the two distributions, which gives a measure of the syntactic fixedness of the V-N collocation. Greater the KL divergence, lesser is the compositionality of the collocation

$$\text{Fixedness}_{\text{syn}}(v, n) = \sum_{pt_k \in \mathcal{PS}} P(pt_k|v, n) \log \frac{P(pt_k|v, n)}{P(pt_k)}$$

Latent Semantic Indexing (Baldwin '03, Katz '06)

- **Key Idea:** The degree of compositionality is indicated by the similarity of the MWE vector with that of the composition of the constituent vectors in concept space.
- Represent the MWE and its constituents in concept space
- Get a lower dimensional representation by performing a SVD
- Compose constituent words by a vector sum of their LSI representations.
- Cosine similarity between the MWE vector and the composed vector gives a measure of the decomposability. Greater the similarity, greater is the decomposability

Using multi-lingual word alignment (Tiedemann '06)

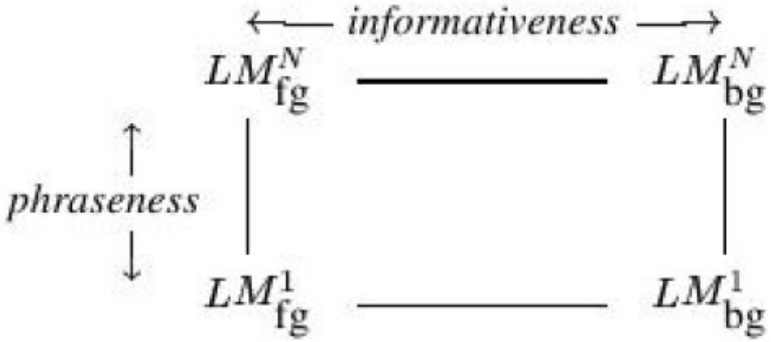


- **Key Idea:** It is difficult to translate idiomatic expressions from one language to another, while literal expressions can be translated word by word.
- Methodology:
  - Align the parallel corpora and create translation links for every word i.e. List of possible translations of the word.
  - Words of idiomatic MWE are likely to have more translations than that of composable expressions. This uncertainty is expressed as an entropy measure. More idiomatic the expression, the higher the entropy.

$$H\ T_s|s \quad -\sum_{t \in T_s} P\ t|s\ log P\ t|s$$

Language Modelling (Tomokiyo 2003)

- Use a foreground and background corpus for domain specific term extraction
- Build multiple models



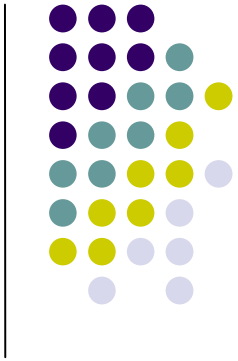
- Difference between:
  - foreground unigram and n-gram model distributions indicator of collocation significance (phraseness)
  - foreground and backgram n-gram model distributions indicator of term novelty (informativeness)
- Data sparsity an issue

To wrap up

- Use a combination of all relevant measures discussed, with due weight given to each

- No standard data sets, evaluation practices
  - In case of binary classification of MWE, measure precision and recall
  - In case of ordinal ranking of MWE, calculate Kendall's Tau coefficient or Spearman Rank correlation method
  - Gold standards for MWE evaluation
    - Human annotation
    - WordNet, idiom dictionaries (SAID, etc.).

Summary



- MWE is an umbrella term for very varied syntactic categories
- Need to understand the language features for each MWE type and translate them into extraction policies.
- Primary Methods: Hypothesis testing, substitutionality, selectional preferences, syntactic fixedness and contextual features.
- Development of standard evaluation measures and datasets required

Further work

- Develop efficient methods for extraction of MWE for smaller corpus
- Extraction of multiword terms in a domain-restricted corpus
- Extraction of MWEs for Hindi/Marathi
  - Lack of NLP resources for Indian languages
  - Free word order

References

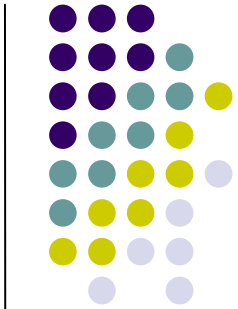
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multi-word expressions: A Pain in the neck for NLP. In *Proceed-ings of CICLing* , 2002.
- Sriram Venkatapathy and Aravind K. Joshi. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of HLT/EMNLP,* 2005.
- Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 1993
- KW Church, P Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 1990
- F Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 1993

References (2)

- D. Lin. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, University of Maryland, 1999.
- T. Baldwin, C. Bannard, T. Tanaka, and D.Widdows. An Empirical Model of Multiword Expressions Decomposability. *In Proc. of the ACL-2003 Workshop on Multiword Expressions,* 2003.
- Fazly and S. Stevenson. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the EACL*, Trento, Italy, 2006.
- Tim de Cruys and Begona Villada Moiron. Semantics-based multiword expression extraction. *ACL-2007 Workshop on Multiword Expressions., 2007*
- Takashi Tomokiyo, Matthew Hurst, A Language Model Approach to Keyphrase Extraction. *ACL Workshop on MWE, 2003*

References (3)

- D. McCarthy, B. Keller, and J. Carroll.Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, Japan*., 2003
- Philip Resnik. Selection and Information: A Class-Based Approach to Lexical Relationships. PhD thesis, University of Pennsylvania, 1993.
- Irina Dahlmann and Svenja Adolphs. Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)? *ACL-2007 Workshop on Multiword Expressions, 2007.*
- B.Villada Moiron and J. Tiedemann. Identifying idiomatic expressions using automatic word alignment. *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a multilingual context*, 2006.



Thank You

Substitution by similar words(Lin '99)

- Lin uses an automatically generated thesaurus for finding similar words and defines a PMI measure taking into account the dependency relations in which the words take part, thus capturing syntactic relations too.
- PMI formula

$$P(w, w') = \frac{\| w, r, w' \| \times \| *, r, * \|}{\| w, r, * \| \times \| *, r, w' \|}$$

||x, y, z || is the cardinality of the triple x, y, z
r is the dependency relation through which w and w_0 are related.
* means any word relation

Distributed Frequency of Object (Tapanainen '98)

- This measure is applicable for Verb-Noun collocations
- **Key idea**: If an object appears only with one verb (or few verbs) in a large corpus, the collocation is expected to have idiomatic nature

  e.g. 'sure' has 'make' as its verb in 'make sure'. It is unlikely that 'sure' will be associated with other verbs.
- To capture this phenomenon, DFO is defined as:

$$d(o) = \frac{\sum_i^n f(v_i, o)}{n}$$

where,
$f(v_i,o)$ is the frequency of verb v_i and noun-object o occuring together
n is the number of verbs in the corpus

Particle Overlap for Phrasal Verbs (McCarthy '03)

- This method is applicable for phrasal verbs
- The particle in literal verb-particle construction contributes to the semantics of the phrase. e.g. *climb up*

  However, in phrasal verbs, it is more for the effect than for the literal meaning e.g. *speak up*
- **Test:** Replace the verb with related verbs and see if it forms a likely verb-particle construction
  - replacing '*climb*' with related verbs – *walk up*, *run up*, *limp up*, *crawl up*, which are plausible
  - replacing '*speak*' with related verbs - *talk up*, *chatter up*, which don't make sense and hence is not likely to be found in corpus
- This test measures the number of related verb-particle constructions that can be listed for the given V-P from an automatically generated thesaurus. More number of phrasal verbs with same particle indicates higher compositionality