"Real-Time" Identification of MWE Candidates in Databases from the BNC and the Web

"Identifying and Researching Multi-Word Units"
British Association for Applied Linguistics Corpus Linguistics SIG
Oxford Text Archive
Oxford 21 April 2005

William H. Fletcher
United States Naval Academy
(2004-05 Radboud University of Nijmegen)
http://pie.usna.edu
http://kwicfinder.com

Objectives of Presentation

- Describe background and biases
- Define key terms elastically
- Outline my software applications
- Sketch range of uses and target audiences envisioned
- Show and compare MRR and MI
- Encourage feedback and suggestions for further development

Background and Biases

- Multimedia in CALL – user (interface)
- KWiCFinder to…
  - Identify useful texts
  - Find examples of actual use for teaching and writing
  - Clarify linguistic questions
  - Explore emerging semantic fields
  - build web-based ad-hoc corpora
  - download free from KWiCFinder.com

Background and Biases

- kfNgram *n-grams, phrase-frames*

  free, flexible, GUI; fast even on large datasets (20MW)
- "Phrases in English" website
  - *n*-grams *(n = 1 – 8)*
  - phrase-frames: set of *n*-gram variants identical in all but one word
  - PoS-grams: set of n-gram variants with the same sequence of PoS tags
  - chargrams
  - now BNC; sub-corpora, MICASE and ANS to follow

Background and Biases

- Web as Corpus Search Engine Consortium
  - Initiative of Silvia Bernardini and Marco Baroni, University of Bologna, Forlì
  - Other WAC enthusiasts: Mr. Collocations Stefan Evert, Sebastian Hoffmann, Adam Kilgarriff and myself
  - Initial goal: gigaword Web corpus (800M English, 100M each German and Italian)

Background and Biases

- Emphasis on the **practical:** reasonable speed, acceptable precision and recall
- Motivations
  - on-the-fly subcorpora for PIE
  - kfNgramDB
    - overcome kfNgram limitations: static lists, straight frequency
    - managing KWiCFinder ad-hoc Web corpora
    - better integration all three tools

Objective

Evaluate and compare statistical techniques to identify MWE candidates for…

- corpus database queries for MWEs with specific lexical items
- subsequent screening, either manually or with processing-intensive metrics deemed more effective than those used here

Terms

- MWE cover-term for multi-word units, salient collocations, formulaic expressions
- Real-time / on-the-fly with “tolerable” delay
- Scalable from kilo- to mega- and giga-corpora

In practice “real time” for (sub)corpora ≤ 25 MW

Target Audience kfNgramDB

- (Corpus) linguists
  - compare subcorpora in large linguistic databases
  - identify content domain and text-type for Web as corpus
  - learn database principles by example on PC
- Language professionals
  - teachers, advanced language learners: readings, instructional materials, examples; identify (MW)Es
  - writers (L2 / L1): organize / maintain exemplars for *imitatio,* personal corpus and reference materials
  - translators: domain-specific parallel / comparable corpora, possibly compiled ad-hoc from Web sources

Relational databases (RDMS) – Why?

- organize linguistic data, "rapid" retrieval
- sophisticated queries *relating* the content of one field or table to others
- filter / focus results by relevant criteria
- dynamic interactive datasets, not static list
- standard query language SQL: skills transfer to other RDMS
- several powerful RDMS systems are
  - free
  - multi-platform (develop on PC, deploy on *nix)

Relational databases – Which?

- Microsoft Access
  - + "wizards" – easy to learn
  - + produces SQL queries adaptable to other RDMS
  - + excellent front-end to other RDMSs (e.g. MySQL)
  - – Windows only (MS Office Pro Suite)

Relational databases – Which?

MySQL

- +free, fast, scalable
- +tight integration with PHP for Web interface
- ±powerful non-standard SQL extensions
- +active development, large, helpful user base
- +user-defined C functions callable in queries (e.g. to calculate lexical association metrics)
- +embeddable in other applications
- +multiple platform

Which Lexical Association Metric?

"Gravity Counts for the boundaries of collocations"*

- Compares Mutual Information, T-score, Dice, Gravity Counts
- Gravity Counts take larger context into account
  - most useful for identifying collocation boundaries
  - *but data processing intensive*

* Daudaravičius, Vidas and Rūta Marcinkevičienė, *International Journal of Corpus Linguistics,* 9:2 (2004), 321-348.

Mutual Rank Ratio

- Paul Deane, Educational Testing Service, "A Nonparametric Method for Extraction of Candidate Phrasal terms", Association for Computational Linguistics 2005.
- "lexical association metric for knowledge-free extraction of phrasal terms", identification of MWUs in untagged text
- Based on ratio of "global" to "local" *shared ranks*
- Performance similar or superior to other metrics identifying 2- and 3-grams in WordNet…

  …*when n-grams including the top 160 ranked types are excluded*

Mutual Rank Ratio

shared rank: "tied" items assigned same rank e.g.

- items 10-15 all have frequency 512
- shared rank is (10 + 15) / 2 = 12.5
- next item ranked 16 *(higher if shared)*

global and *local rank*

United Kingdom

- *local rank* of a specific n-gram *(LR)*
  - united kingdom
- *global rank* of phrase-frames of which *n*-gram is a variant *(GR)*
  - * kingdom *(the k., animal k., his k. … united k.)*
  - united * *(u. kingdom, u. states, u. nations, u. distillers…)*

Mutual Rank Ratio

Formula

$$MRR = \frac{\left(GR_{united\ *} \cdot GR_{*\ kingdom}\right)^{1/2}}{LR_{united\ kingdom}}$$

*n*th root of product of all Global *(phrase-frame)* Ranks divided by Local *(*n*-gram)* Rank

Mutual Rank Ratio Pros & Cons

- \+ Easy to calculate, especially if *n*-grams and phrase-frames are already known *(PIE, kfNgram)*
- \+ Finds MWUs in untagged text >= others*
- \+ Weighting reflects Zipfian distribution

- \- Excludes MWUs…
  - \- with top types *(state of the art, matter of principle)*
  - \- not in phrase-frames (“singletons”)

* if most frequent types excluded

Mutual Information

- Popular metric for finding rare word pairs
- Formula *(after D & M)*

  $\mathrm{MI}(x,y) = \log_2 (N \cdot f(x,y) / f(x) \cdot f(y))$

  N corpus size

  f(x,y) frequency of co-occurrence

  f(x), f(y) total frequency in corpus
- Calculated for pairs of words with frequency rank > 150, span 2-4 words;

  n-grams with these pairs retrieved (could include *state of the art, matter of principle)*

Mutual Information Pros & Cons

+ Straightforward calculation with parameters needed for some other metrics
+ Finds "elusive" items, including singletons
+ Complements MRR

- Strong bias toward the infrequent:

  Two co-occurring rare words will show a high score, but two co-occurring frequent words will show a low score. *(D & M 325)*

MRR and MI Compared

- Minimal overlap in MWEs (top 500 items < 20% shared; ranking very different)
- Complementary: both identify different sets of "interesting" MWE candidates
- Both
  - calculation on-the-fly in series of SQL queries alone impractical / intractable on PC for corpora > 5MW
  - hybrid approach with programmatic math faster, more scalable

MRR and MI Compared

Top-ranked 500 *n*-grams

- by MRR but not by MI
- by MI but not by MRR
- by both (<20% of total)

in

- MICASE *(MIchigan Corpus of Academic Spoken English, 1.8 MW)*
- EuroParl *(European Parliament transcripts, 500 KW)*

Click for word lists

MRR and "Singletons"

- In a large tagged corpus (BNC), Mutual Rank Ratio strands many MWE "singletons", *n*-grams lacking a phrase-frame for at least one wildword position
- Frequent singletons should be reviewed for potential MWEs
- Singletons less problematic for smaller untagged corpora

Click for word lists

Toward Gigacorpora

- Today's RDMSs excel at locating and relating millions of records, but do not scale well into the billions
- Search engine technology points the way
- Doug Cutting's *Lucene* open source text indexer (Java) handles large plain-text collections
- Hybrid approach
  - Lucene to locate documents / passages
  - RDMS to manage text metadata, markup

“Real-Time” Identification of MWE Candidates

Reactions and suggestions encouraged.

http://www.kwicfinder.com/

http://pie.usna.edu

fletcher@usna.edu