# *Washing Machines, Heavy Water, and The Web*
# The Representation and Processing of Lexical Phrases

Robert Krovetz
Lexical Research

## Outline

- Background on Multi-Word Expressions
- Recognizing MWE's using different types of linguistic support
- Semantic classes for MWE's

- Background on Named Entity Recognition
- Experiments comparing 3 state-of-the-art classifiers
- Unit test for Named-Entities

## Aim: Recognize and Process Lexical Phrases

- Examples
  - *operating system, grass roots, singular value decomposition, Great Britain*.
- Dictionaries often contain such terms, but most of the terms we want are not found.
- Usually identified either by hand or by statistical methods.

## Facts about Lexical Phrases

- They are mostly two and three words expressions.
- The whole is more than the sum of the parts.
- The component words are often ambiguous in isolation.

## Facts about Lexical Phrases

- Frequent by token, but rare by type
- Component words co-occur more often than expected
- Resist substitution or paraphrase
- Total number of unique terms is unknown (even for practical purposes)

## Problems with Previous Approaches

- Statistical measures give mixed results. Sometimes ranking terms by raw frequency was the best.
- Most phrases occur too infrequently to provide reliable results.
- Low precision and recall.
- Unrealistic assumptions about statistical distribution and term independence.

## Different measures give different results

- **TMI**: *a little, did not, this is, united states, new york, know what, a good, a long, a moment, a small*
- **Log-likelihood**: *sherriff poulson, simon hugget, robin redbreast, eric torasian, colonel hillandale, colonel sapp, nurse leathereor, st. catherine*

## Motivation

- **Phrasebank** project at the NEC Research Institute.
- Goal – identify lexical phrases for support of a speech-to-speech machine translation system.

## Methodology

- Determine effectiveness and productivity of various types of support for recognizing lexical phrases.
- Compare linguistically-based methods with statistical methods.
- Integrate both approaches for the most effective results.

## Sources of Evidence

- Open and closed compounds – data base/ database
- Hyphenated forms – box-office/box office
- Genitive form – retrieval of information/ information retrieval

## Sources of Evidence

- Acronyms – AI (Artificial Intelligence
  Aortal Infarction
  Artificial Insemination)
- Inflectional Morphology – data base (s)
- Derivational morphology –
  New York/New Yorker
- Subject Codes –
  Protestant Minister

## More than one source of evidence can apply

- Hot dog (plural, compound)
- Information Retrieval (acronym, genitive)

# Initial Evaluation

- Prototype developed at NECI with 12,000 documents from Wall Street Journal.
- Candidate phrases identified using a part-of-speech tagger (Brill)
- Linguistic support assigned to candidate phrases
- Phrases were sorted by the degree of support
- Evaluated phrases with more than one source of support. Phrases were also evaluated by an independent judge

# Results of Evaluation

- There were 1650 phrases that had more than one type of linguistic support.
- About one-third of the evaluated phrases were considered “lexical”.
- There was a 75% agreement rate between the two judges about which phrases were lexical.
- The evaluation was difficult because of insufficient criteria for making a judgement.

## Assessment of Linguistic Support (Tipster corpus)

- Hyphenation is one of the largest sources (over 640K terms). Also a large number of inflectional variants, but less effective.
- Acronyms are the smallest source (6600 terms). Mostly technical terminology.
- Open/Closed compounds are in between. Many false positives due to "Space bar errors".

## What makes a Lexical Phrase *lexical*?

- Linguistic response – lexical phrases are non-compositional. But how can we know that from a computational perspective?
- Sometimes phrases *can* be composed and still be lexical!

## What makes a Lexical Phrase *lexical*?

- Conjecture: lexical phrases are lexical because of a lack of predictability for how to compose the meaning from the component words (interpreting).
- Conjecture: lexical phrases are lexical because of a lack of predictability for how to choose the component words (generation).

## Semantic Types for Lexical Phrases

- **Idiomatic:** "red herring", "hot dog", "end run"
- **Missing Object:** "washing machine", "operating system"
- **External value:** "room temperature", "small plane", "light truck"
- **Restricted word sense**: "stock market"
- **Unknown case relation:** "alligator shoes", "horse shoes", "dog sled"
- **Partially opaque**: "witching hour", "heavy water"

## *Hard water* vs. *Heavy water*

- **Hard water** has variation: harder water, hardest water.
- It also has an antonym-form: **water softener.**
- **Heavy water** does not have such variation or an antonym-form.
- Both are partially opaque; we know the expressions refer to water. We don't necessarily know how the adjectives relate.

## Lexical Phrases for Natural Language Processing

- Machine Translation:
  - Idiomatic phrases -- *hot dog* is not "warm canine"
  - External-value phrases – *room temperature* is translated differently in Japanese than "room" or "temperature" in isolation
  - Restricted word-sense – *stock market* should not be translated as a store that only sells soup, or a market in which we only trade gun barrels!
  - Missing-Object phrase – *washing machine* is "clothes wash machine" in Chinese. How do we supply the word "clothes"?

## Lexical Phrases for Natural Language Processing

- Information Retrieval
  - Idiomatic phrases --- need to be enclosed in quotes
  - All other classes --- need to give partial credit to component words. It is an open problem about how to assign credit.
- Cross-lingual IR and MT
  - Noun substitution and verb substitution ('dish' washer vs. 'flatware' washer (French); dish 'washer' vs. dish 'rinser' (German)

## Assessment of Recall

- Longman (4K phrases), Collins (11K), WordNet (27K)
- Terms in WordNet that are not in TREC corpus:
  - Monoamine Oxidase
  - Dirca Palustris
  - Yellow Mountain Saxifrage

## Assessment of Recall

- Corpus:  SourceFinder (425 Million words). Internally used at ETS.  Contents: magazine articles, journal articles, newspaper text).
- Gold Standard : WordNet (53,764 bigrams and 7,613 trigrams).
- Attested: 19,939 bigrams and 1700 trigrams.

## Current Work

- Project at the Educational Testing Service. The aim is to develop methods to assess breadth and depth of vocabulary.
- Vocabulary terms are grouped into topically organized clusters.
- Need to recognize terms such as “civil war” and “Abraham Lincoln” .

## Current Work

- Mutual Rank-Ratio (MRR) is an approach to ranking n-grams according to a new statistical measure. It is more effective than Mutual Information and log-likelihood at ranking MWE's from WordNet out of n-grams from a corpus. MRR was developed by Paul Deane.
- I am working with Paul on combining my approach (using supported phrases) with his.

## Current Work

- Advantages of MRR
  - Does not make independence assumption
  - Allows scores to be combined across n-grams of different lengths
  - Better suited to Zipfian distribution
  - Better precision than existing methods

## Different measures give different results

- **MRR**: *julius caesar, winston churchill, potato chips, peanut butter, fredrick douglas, ronald reagan, tia dolores, don quixote, cash register, santa claus*

## False Positives and MRR

- Morphological variants of established forms
- Partial n-grams (*york city)*
- Highly productive constructs (*January 2)*

## Mutual Rank Ratio

- For more information, see the paper:
  P. Deane, ``A Nonparametric Method for extraction of candidate phrasal terms'', in *ACL '05, Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* 2005.

## Named-Entity Recognition (NER)

- NER is a component part of Information Extraction. The aim is to recognize common types of entities like PERSON, LOCATION, and ORGANIZATION.
- Accuracy is usually determined in the context of competitions. Assessment has focused on extrinsic evaluation.

## State of the Art for NER

- Named-entity classes are based on guidelines, not standards.
- Criteria for membership in a class can change between competitions – creates difficulty comparing results.
- Proprietary software means that results might not be replicable by others in the community.

## Evaluation Methodology

- Compared three taggers: Stanford, LBJ, and BBN IdentiFinder.
- **Agreement on Classification –** What is the agreement rate on PERSON, ORGANIZATION, and LOCATION?
- **Ambiguity in Discourse –** How often does each tagger produce multiple classifications of the same token in a single document?

## Initial Evaluation

- Corpus – SourceFinder (425 million words; more than 270,000 documents).
- Stanford and LBJ taggers used for initial evaluation.
- Result – tagging mistakes were apparent by inspection. "Berners-Lee" was tagged as a **PERSON,** as well as **ORGANIZATION.**

## Sample of Classification Mistakes - PERSON

| Stanford | LBJ |
|---|---|
| Shiloh | A.sub.1 |
| Yale | What |
| Motown | Jurassic Park |
| Le Monde | Auschwitz |
| Drosophila | T.Rex |

## Sample of Classification Mistakes - ORGANIZATION

| Stanford | LBJ |
|---|---|
| RNA | Santa Barbara |
| Arnold | FIGURE |
| NaCl | Number: |
| AARGH | OMITTED |
| Drosophila | Middle Ages |

## Sample of Classification Mistakes - LOCATION

| Stanford | LBJ |
|---|---|
| Hebrew | The New Republic |
| ASCII | DNA |
| Tina | Mom |
| Jr. | Ph.D. |
| Drosophila | Drosophila |

## Agreement Rate by Class

| | Common Entities | Percentage |
|---|---|---|
| **PERSON** | 548,864 | 58% |
| **ORG** | 249,888 | 34% |
| **LOC** | 102,332 | 37% |

## Ambiguity within a Discourse

***Stanford***

| | Overlap | Co-occurrences |
|---|---|---|
| PERSON-ORG | 98,776 | 40% |
| PERSON-LOC | 72,296 | 62% |
| ORG-LOC | 80,337 | 45% |

# Ambiguity within a Discourse

***LBJ***

| | Overlap | Co-occurrences |
|---|---|---|
| PERSON-ORG | 58,574 | 68% |
| PERSON-LOC | 55,376 | 69% |
| ORG-LOC | 64,399 | 63% |

# Comparing all 3 Taggers

- Corpus – American National Corpus.
- Many of the same problems were encountered.
- Taggers performed very well for entities that were common in each class. But there were errors even with phrases that were frequent.

## Unit Test for NER

- Tests for the following:
  - Orthography
  - Terms that are in upper case, but not named-entities (RNA, AAARGH)
  - Last names in close proximity to full name
  - Terms that contain punctuation marks that at are not named-entities (A.sub.1)

## Unit Test for NER

- Variation in form (MIT, M.I.T, Massachusetts Institute of Technology)
- Acronyms (ETS, UN), with and without expansion.
- Terms that contain a preposition – this helps test for correct identification of extent.
- Knowledge-based classification (Amherst, MA and Amherst College).

**We Didn't Start the Fire**
**(**with apologies to billy joel)

We didn't start the fire
It was Bar-Hillel's churnin'
for some better learnin'

LBJ and JFK
OMG it's MLK
DNA and RNA
are not ORGANIZATIONS
that's not OK

- Amherst, MA and Amherst College
  Separating them requires knowledge
- The Web is not a PERSON
  as we can plainly see
  Despite the upcoming
  Singularity
- Berners-Lee *is* a PERSON semantically
  We need such recognition
  If his dream will come to be
- We didn't start the fire
  it was BAR-HILLEL's churnin'
  for some better learnin'

# This is not a Unit Test

(a tribute to Rene Magritte and RMS)

- Although we created this test with humor, we intend it as a serious test of the phenomena we encountered. These problems include ambiguity between entities (such as Bill Clinton and Clinton, Michigan), uneven treatment of variant forms (MIT, M.I.T., and Massachusetts Institute of Technology - these should all be labeled the same in this text - are they?), and frequent false positives such as RNA and T. Rex. .....

# Proposal to help improve NER

- We propose that the community focus on four classes: **PERSON, LOCATION, ORGANIZATION,** and **MISC.**
- Rationale:

1. These classes are more difficult than dates, times and currencies.
2. Widespread disagreement between taggers on these classes.
3. Need a class for handling terms that do not fit the first three classes.

## Proposal to help improve NER

- Create test sets across a variety of domains. It is not enough to work with newswire and biomedical text.
- Use standardized test sets that are designed to test for different types of linguistic phenomena.
- Report accuracy rates separately for different classes.
- Establish a way for tagging systems to express uncertainly about a classification.

## More on the unit test

- For more information, see the paper:

  Robert Krovetz, Paul Deane, and Nitin Madnani, "The Web is not a PERSON, Berners-Lee is not an ORGANIZATION, and African-Americans are not LOCATIONS: an Analysis of the Performance of Named-Entity Recognition," in *Proceedings of the ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World, 2011*

# Conclusion

- Multi-word Expressions represent significant challenges for natural language processing.
- It is difficult to determine what makes a lexical phrase *lexical* without semantic criteria.
- We need more integration between a "systems" approach and an "AI" approach .