LANGUAGE AND INFORMATION :

OLD IDEAS, NEW ACHIEVEMENTS

Karen Sparck Jones

Computer Laboratory
University of Cambridge

23.4.02

refs: www.cl.cam.ac.uk/~ksj/

The slogan :

in language and information processing -

don’t look for meanings, count mentions

Can it work ?

EXAMPLE ==>

TEXT :

Wombats are primarily active at night. They spend most of the day resting in their burrows, emerging only at night to forage for food - fruit, young plant shoots, etc. But wombats are sometimes seen during the day if the weather is grey and food is in short supply.

SUMMARISING :

Adequate ??   Wombats night food.

Desired :   Wombats are nocturnal vegetarians.

RETRIEVAL :

Query : wombat eating habits

Adequate match ? wombat*

Desired match : wombat* {eat/forage/food/fruit}

THEME :
distributional data about words
conveys enough about meaning for many purposes

[linguistic theory eg Harris]

computational practice
language description
* information processing tasks

fine for simple tasks eg retrieval
problem for complex tasks eg summarising

theoretical development :
probabilistic models

The talk :

1 How the work began

2 What's been achieved

3 Where we have to go

1 BEGINNINGS - late 50s

HP Luhn 1957

‘communication of ideas by way of words is carried out on the basis of statistical probability’

index documents via frequent terms

invoke thesaurus classes (KWIC, manual)

illustrated by example indexes

summarise documents using sentences where frequent words concentrated

demonstrated with ICSI papers 1958

IC51 INTERNATIONAL CONFERENCE ON SCIENTIFIC INFORMATION
AREA 5 PG 103

THE ANALOGY BETWEEN MECHANICAL TRANSLATION AND LIBRARY RETRIEVAL

MASTERMAN M CAMBRIDGE LANGUAGE RESEARCH UNIT CAMBRIDGE ENGLAND
NEEDHAM RM CAMBRIDGE LANGUAGE RESEARCH UNIT CAMBRIDGE ENGLAND
JONES ES CAMBRIDGE LANGUAGE RESEARCH UNIT CAMBRIDGE ENGLAND

AUTO ABSTRACT

6 NOW, IN THE PRESENT STATE OF RESEARCH THIS ANALOGY CAN ONLY BE DRAWN AT ALL PRECISELY BETWEEN ONE FORM OF LIBRARY RETRIEVAL PROCEDURE, AND ONE FORM OF MECHANICAL TRANSLATION PROCEDURE.. THESE TWO ANALOGOUS PROCEDURES ARE THOSE, IN EACH FIELD, WHICH MAKE USE OF A THESAURUS.

10 WE PROPOSE, THEN, THAT A CONCEPTUALLY BASED, THESAURUS TYPE OF LANGUAGE CLASSIFICATION SHOULD BE USED FOR A COMPLETELY GENERALISED RETRIEVAL PROCEDURE, THIS CLASSIFICATION PROCEDURE BEING, BY ITS NATURE, INTERLINGUAL.

14 TRANSLATION SPECIALISTS, AND, IN PARTICUALR, LINGUISTS DENY EVEN THE POSSIBILITY OF THE ANALOGY BY MAINTAINING THAT ANY CLASSIFICATION OF LANGUAGE BASED ON A THESAURUS CAN, AT BEST, ONLY HOPE TO TRANSLATE SEMANTIC MEANING, WHEREAS LANGUAGE IS PRIMARILY A SYSTEM OF GRAMMAR AND SYNTAX.. AND BOTH OF THESE ARE NOTORIOUSLY MONOLINGUAL.

18 THE OBJECT OF THIS PAPER IS TO REFUTE THIS CRITICISM BY SHOWING HOW A TYPE OF RETRIEVAL PROCEDURE, BASED ON A THESAURUS ALREADY BEING USED FOR THE EXPERIMENTAL TRANSLATION OF SEMANTIC MEANING, MIGHT ALSO BE EXTENDED SO AS TO TRANSLATE GRAMMAR AND SYNTAX.

Cambridge Language Research Unit 1956/8

thesaurus for lexical normalisation
AND concept determination

in document retrieval
machine translation

retrieval :
word substitution for matching
(eat <-> food)

translation :
sense disambiguation for transfer
(spend + day, night => pass time
not disburse

BUILDING THESAURI ie semantic classifications :

automatically from distributional data -
words with similar text behaviour a class

direct :
DOC1, DOC2 ... eat, food, forage, plants

==> < eat, food, forage >

indirect :
spend time, spend days
waste time, waste days

```
==> spend/waste [PASS]
    time/days   [PERIOD]
```

* GENERAL * MODEL FOR LIP :

1. exploit occurrence frequency as indicating
   discourse content salience

2. take cooccurrence frequency as showing
   conceptual relations

3. use old classes in interpreting
   new word conjunctions

4. formally model processes with probability :
   if x is frequent in text,
     text is probably about X
   if x and y are frequent together,
     text is probably about XY

ISSUES IN CLASSIFICATION :

data - need vast amounts (none in 1960)

approximate with simple keyword lists

finesse by bootstrapping from dictionaries

definitions -

context - document, sentence, constituent ?

similarity - form eg direct ?

coeffcent eg Jaccard ?

class - type eg overlapping, non-hierarchical ?

>>> criterion eg ‘clump’ ?

search procedures ?

mechanics - need a lot of power (little in 1960)

CLASSIFICATION EXPERIMENTS - 60s :

Retrieval -
- simpler case, urgent need
- pursued with enthusiasm
  - grouping methods eg Factor Analysis (Borko)
  - indexing uses eg term expansion (Stiles)
  - user aids eg semantic road maps (Doyle)
- large tests (Dennis)

Translation -
- lexicon-based grouping experiments (KSJ)

Foundation studies -
- parsed astrobotany text analysis (Harper)

ISSUES WITH TASKS :

Translation -
 no embedding systems for thesauri

Retrieval -
 no full texts for auto indexing

Summarising -
 no texts for auto abstracting

AND

in limited retrieval tests
 classifications did not work

FROM THE 60s TO THE 80s

Some studies eg sublanguage classes (Sager)
summarising cue words

Retrieval -
‘Keeping the unfashionable flag flying’

steady task advance :
- simple term indexing with frequency weights
- relevance feedback classification eg Salton
- probabilistic modelling eg Robertson :
  - get probability document relevant
    - via query term frequency

NEW FACTORS IN THE 80s

decent NLP tools eg grammars
but needing better detail

non-trivial application systems
but wanting better task functionality

rapidly-growing supply of m-r corpora
for improving resources
refining tasks

(much more powerful machines)

==> statistical revival

2 NOW, after 90s :

BUILDING SEMANTIC RESOURCES :
lexical associations, classifications
eg Church, Schuetze, Pereira

PROGRESS :
operations on vast scale -
lexicon eg 50 M words 1 M nv pairs
concept indexing eg LSI (Dumais)
surface grammars (‘language models’)

opportunity : the Web as Corpus

proofs of concept, process
EXAMPLE ==>

GROUPING EXAMPLE - Rooth et al :

v = mobilise CLASS 6 ‘GROUP ACTION’
most probable for object n
n = force, people, society, party ...
most <-> least frequent object

disambiguation for translation :

‘mobilisierung Gesellschaft’

=> mobilise society, not party

LACK OF PROGRESS ON RESOURCES :

large corpora but small vocabularies

association lists not classifications

crude classification models (eg K exclusive)

no system integration

not always task payoff (eg retrieval)

counter-opportunity : WordNet

IMPLEMENTING TASK SYSTEMS :

challenge : the Web as Information World

Translation - still traditional

Retrieval (growth, *many* players) -

statistical approaches
  endorsed in large tests
  featured in Web engines

EXAMPLE ==>

RETRIEVAL EXAMPLE - Robertson et al :

	Precision at rank 10
query :	
terms only	.11
weighted	.52
expanded	.61

Summarising (renaissance, *many* players) -

statistical methods with bells, whistles
eg cues, pruning, shallow parsing

Web applications

but problems - performance, multi-document

EXAMPLE ==>

SUMMARISING EXAMPLE - Boguraev et al :

‘One day, everything Bill Gates ...’
declares Gilbert Amelio, the boss at Apple
Computer ...

```
==>
   APPLE, MICROSOFT
    Apple lost $ 816 million
    Microsoft made $ 2 billion
    Apple is in a position
    Apple needs something dramatic
```

NEW IDEAS - A REVOLUTION ?

Language Modelling :

unified probabilistic paradigm
  for tasks
  also resources

derived from speech recognition
  applicable everywhere ?

KEY NOTIONS :

relation between two bags/strings ..

one generates the other, but with noise

language/information process, task is
   RECOVERING THE GENERATOR

natural formal account by exploiting Bayes

$$P(X|Y) = \frac{P(Y|X)\ P(X)}{P(Y)}$$

estimate P(Y|X), P(X) from frequency data
may ignore P(Y)

Speech recognition :

what is word string X, given sound string Y ?

ie best generator for noisy sounds Y ?

Translation :

what is the target string X given the source Y ?

ie best generator for the wrong words Y ?

Retrieval :

what is (relevant) document X, given query Y ?

ie best generator for the scanty terms Y ?

Summarising :

what is the summary text X for source text Y ?

ie best generator for the padded text Y ?

similarly for process and resource :
eg what descriptors X from word set Y ?

train from examples

allows complex units
reordered
probabilistic units
introduced units

EXAMPLE ==>

LM SUMMARISING EXAMPLE - Witbrock et al :

‘President Clinton met with his top Mideast advisors, including ...., in preparation for a session with ... Israel PM Netanyahu tomorrow. Paltestine leader Arafat is to meet with Clinton later ....’

==> clinton to meet netanyahu arafat

LM ISSUES :

need training data (but can bootstrap)
powerful enough ?
convincing model ?

Speech - works well (eg everyone)
but inbuilt parallelism

Translation - experiment (eg Brown, Knight)
but complex units, dislocation

Retrieval - works well (eg Croft, Lafferty)
but coarse task

Summarising - experiment (eg Witbrock, Marcu)
but radical transformation

3. WHAT TO DO - SUMMARISING CHALLENGE :

Summary :
 condensing text transformation focused on
  important source content

role of source text structure -
 text has content structure
  organised for effective communication
  therefore emphasising important material

  BUT what structure ? how use ?

EXAMPLE ==>

SOURCE TEXT :

Wombats are not domestic animals and do not make good pets. They are very untidy. They spend most of the day asleep. They are liable to bite not only the hand that feeds them, but anything else they don't want to share their space with. So they are not pets for children. Wombats are also very picky about their food. They may not need caviar and champagne, but they certainly expect the best quality spinach leaves, avocados and iced water. So they are expensive as well as unfriendly pets for adults too.

ATTENTION STRUCTURE : CONTENT SALIENCE

basis for simple statistical summarising

Summarising rule :

Take most salient sentences using words

==>
(wombats, pets)

Wombats are not domestic animals and do not make good pets.

SENTENCES / PROPOSITIONS :

S / P1.1 wombats not domestic
P1.2 wombats not pets
P2 wombats untidy
P3 wombats nocturnal
P4 wombats bite X ...
P5 wombats not child pets
P6 wombats picky on food
P7 wombats like X foods ...
P8.1 wombats costly adult pets
P8.2 wombats unfriendly adult pets

LINGUISTIC (RHETORICAL) STRUCTURE :

```
S1.1 <domestic   > Description
S1.2 <pets       > Desc
S2   <untidy     >    Desc      elaboration of 1.1
S3   <nocturnal  >    Desc      elab           1.1
S4   <bites      >    Desc      elab           1.1
S5   <child pets >  Desc      refinement  of 1.2
S6   <picky      >    Desc      elab        of 1.1
S7   <like       >      Desc      elab        of 6
S8.1 <costly pets>  Desc      refine      of 1.2
S8.2 <unfriendly >  Desc      refine         1.2
```

```
Summarising rule : take top/first item
  ==>
  Wombats are not domestic.
```

WORLD-BASED STRUCTURE :

```
  creature
    [ wombat [ behaviour : P2
                           P3
                           P4
                           P6 [ P7 ] ]
             [ status : P1.1 ]
             [ roles : P1.2 [ P5 ] [ P8.1
                              P8.2 ] ] ]
```

Summarising rule : first/top item in fullest slot

==>

Wombats are untidy.

BUT

summarise for purposes

ie in context for uses, for users

proactive specification of requirements
not passive reflection of source

EXAMPLE ==>

SUMMARY PURPOSE 1 :
Summary for ‘Short Guide to Pets’,
wide readership, plain text

LINGUISTIC STRUCTURE :
Description (S1.1, S1.2) ... ...

Strategy : Take top-level descriptions
Express simply

==> Wombats are not domestic. They are not
child pets and are costly and unfriendly
adult pets.

SUMMARY PURPOSE 2 :

Summary for ‘Wombat Database’

limited readership, succinct text

LINGUISTIC STRUCTURE : ...

Strategy : Take top level descriptions ??

Take elaboration content
Express compactly

==> Wombats are untidy, nocturnal, bite,
and are picky eaters.

SUMMARY PURPOSE 1 : ‘Short Guide to Pets’

```
WORLD STRUCTURE :
 [ wombat [ behaviour : P2 ... ]
          [ status : ...
          [ roles : P1.2 ... ]

 Strategy : Take items in fullest ??

            Select for roles
            Express simply

 ==>  Wombats are not pets for children and
      are costly, infriendly pets for adults.
```

SUMMARY PURPOSE 2 :
Summary for ‘Wombat Database’
limited readership, succinct text

WORLD STRUCTURE : ...

Strategy : Take items in fullest
Express compactly

==> Wombats are untidy, nocturnal, biters
and choosy about food.

IMPLICATIONS FOR LM APPROACH :

In principle -
pertinent structure and mode of use
implicit in source-summary training data

In practice -
??

MOREOVER, LM approach
lacks flexibility for ‘ad hoc’ summary
lacks lever for purpose guidance

how useful could it be ?

so far,
statistical language and information processing

has done better than expected

maybe will need hybrid strategies

but right now,

statistical approach has a lot to offer

==> GO FOR IT !