DEFT Workshop @ TALN2011

Simple formula for losing DEFT with more than 90% of correct guesses

written by
Paris 8 PhD. Student & member of

Lutin Userlab
Cité des sciences et de l'industrie

Daniel Devatman Hromada

The Problem

To associate to N (scientific) articles one among N abstracts summarizing the respective article.

The Hypothesis

Simple unweighted addition of relative probabilities of all the words present in the abstract can be exploited as sufficiently adequate approximation of

fulltext → abstract

summarization process.

The Intuition behind Hypothesis

If the term T present in abstract A_i occurs solely in fulltext F_X and nowhere else, T can be taken into account as a **strong marker** of association (A, F_X) (hapaxes, names etc.)

If T occurs twice in F_2 and once in F_3, the contribution of T to overall scoring of coupled associations shall be $(A_i, F_2) = 2 (A_i, F_3)$ etc.

The relative frequency

i.e. relative probability of term t occuring in article a
when compared with the rest of the corpus

$$P_{t,a} = F_{t,a} / F_{t,total}$$

$F_{t,a}$ = number of occurrences of t in a
$F_{t,total}$ = number of occurrences of t in all articles

...note that $F_{t,a}$ for all articles as well as and $F_{t,total}$ can be obtained in one sole pass through the array of articles...

The scoring formula

For every candidate [abstract a, fulltext f] couple we calculate the score by summing up the relative probabilities of all terms present in abstract A

$$score_{a,f} = \sum_{t=1}^{t=T} \sum_{f=1}^{f=N} P_{t,a}$$

Where *t* is the term present in abstract A and $P_{t,f}$ is a relative frequency (pre-calculated in the first pass) of term t in relation to fulltext candidate f (chosen from the set of N fulltexts)

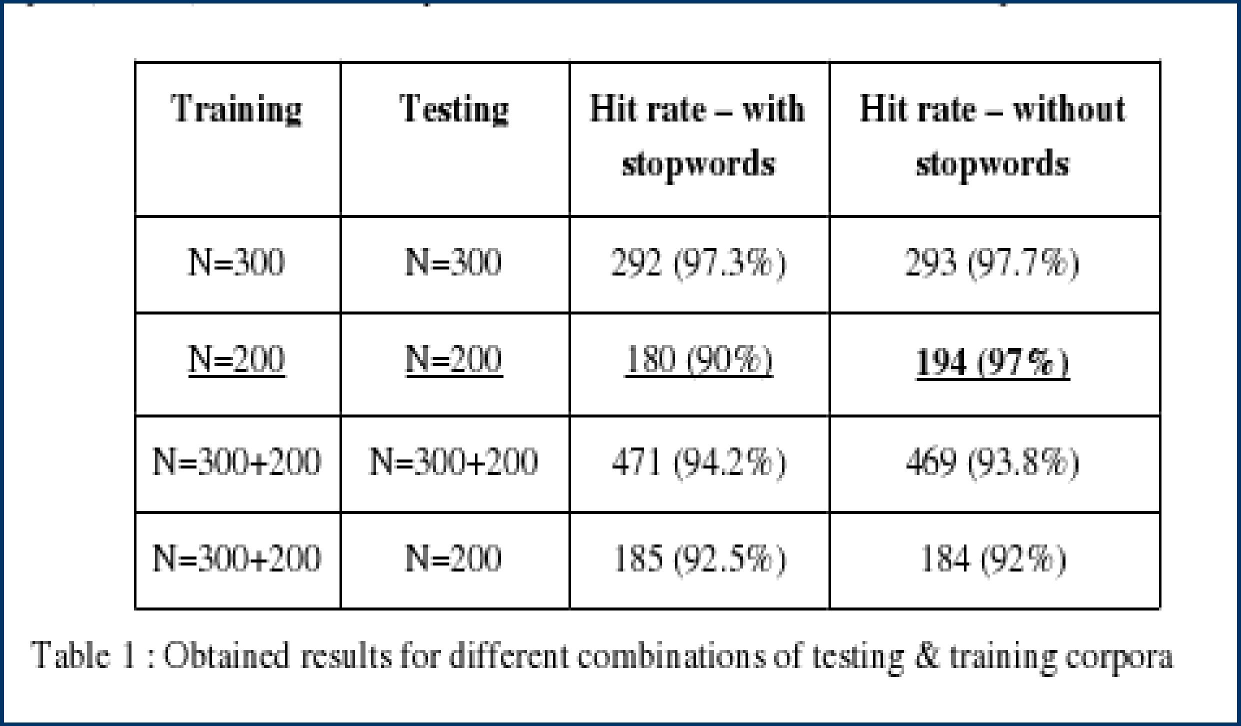
...note that score for all candidate [a,f] couples can be calculated in just one pass through array of abstracts...

Choosing the candidate

Hypothesis : Highest score signifies the presence of the biggest amount of coupling markers with big relative contributions.

So we just sort the $\text{score}_{a,f}$ couples in descending order and couple every a with f from the highest position in such ordered list.

Results



Training	Testing	Hit rate – with stopwords	Hit rate – without stopwords
N=300	N=300	292 (97.3%)	293 (97.7%)
N=200	N=200	180 (90%)	**194 (97%)**
N=300+200	N=300+200	471 (94.2%)	469 (93.8%)
N=300+200	N=200	185 (92.5%)	184 (92%)

Table 1 : Obtained results for different combinations of testing & training corpora

* stopword (CPAN Lingua::StopWords) related experiments were conducted only after reception of results from DEFT organising committee

Conclusion

Hypothesis *« Simple unweighted addition of relative probabilities of all the words present in the abstract can be exploited as « ???sufficiently ??? » adequate approximation of fulltext → abstract summarization process »*

… was not falsified (we had >90% hit rate without recourse to any « heavy » machine learning or semantic space construction techniques)

… offers a swift (1 formula, 2 array passes, 77 lines of code and less than 100 seconds of calculation) answer to the problem of [abstract, fulltext] coupling

… can yield some simple but interesting insights about the (cognitive?) nature of summarization process

… indicates that in case of isolating (chinese) or rather isolating (english, french...) languages, the surface « frequency-based » features of the text can be quite useful

```
#articles are in « art » directory, abstracts are in « res »
directory
print '<?xml version="1.0" encoding="utf-8" ?>'."\n<corpus>\n";
#1st pass - creating total & article-relative word frequency
histograms for all articles
my %word_freq_in_article;
my %word_freq_in_all_articles;
@artz=glob("art/*.pur");
for $art (@artz) {
  $art=~/^art\/(\d\d\d)/;
  $file=$1;
  open(A,$art);
  while (<A>) {
    @wordz=split(/[^\w]/);
    for $word (@wordz) {
      if (!$word_freq_in_all_articles{$word}) {
        $word_freq_in_all_articles{$word}=1;
        $word_freq_in_article{$word}{$file}=1;
      } elsif (!$word_freq_in_article{$word}{$file}) {
        $word_freq_in_all_articles{$word}++;
        $word_freq_in_article{$word}{$file}=1;
      } else {
        $word_freq_in_all_articles{$word}++;
        $word_freq_in_article{$word}{$file}++;
      }
    }
  }
}
#2nd pass – we take every word W from every abstract and then look
at the frequencies of W in all articles
my @keylist;
my %abstract_article;
foreach $f (<res/*.res>) {
  $i{$f} = -s $f;
}
;
@re_filez = (sort{ $i{$b} <=> $i{$a} } keys %i);
for $resfile (@re_filez) {
  $resfile=~/^res\/(\d\d\d)/;
  $abstract=$1; push @keylist,
    $abstract;
  open(F,$resfile);
  while (<F>) {
    if (/<p>(.*?)<\/p>/) {
      $content=$1;
      @wordz=split(/[^\w]/,$content);
      for $word (@wordz) {
```

```
        for $article (keys%{$word_freq_in_article{$word}}) {
          $abstract_article{$abstract}{$article}=0 if
(!$abstract_article{$abstract}{$article});
          #formula which attributes the score to every (abstract,
article) couple
          $abstract_article{$abstract}{$article}+=
($word_freq_in_article{$word}{$article} /

($word_freq_in_all_articles{$word})) if
$word_freq_in_article{$word}{$article};
        }
      }
    }
  }
}
our @used;
our @keyz;
sub r {
  $depth=$_[0]; if (grep($_ eq $keyz[$depth], @used)) {
    r($depth+1);
  } else {
    return $keyz[$depth];
  }
}
for $abstract (@keylist) {
  %abhash=%{$abstract_article{$abstract}};
  #descendant ordering of (abstract, article) couples gives us the
best candidates
  @keyz = sort {$abhash{$b} <=> $abhash{$a}} (keys(%abhash));
  $key=r(0);
  if ($abhash{$keyz[0]}>($abhash{$keyz[1]}+0.23)) {
    push @used,$key;
  }
  print "<doc><resume fichier=\"$abstract.res\" /><article
fichier=\"$key.art\" /></doc>\n";
  $hit++ if ($resultz{$abstract}==$key);
}
print "</corpus>\n";
```

DEFT-related conclusion

Our hypothesis
was not falsified but was definitely

not sufficient to win DEFT2011

(our results were undoubtably worst, so sorry guys for lowering the overall average :)

Congratulations to the winners !

What is the probability of occurrence of words

« Thank You for Your attention»

on the last slide like this one ???

And I thank also Mr. Adil ElGhali for having presented these slides