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

Daniel Devatman Hromada

(1) PhD. student Université Paris 8 – St. Denis
hromi@kyberia.sk

Abstract

A simple scoring method based upon unweighted summing of relative word occurrence probabilities was implemented in order to find the solution to the second problem of data-mining competition “Défi Fouille de Textes”. The objective - to couple, one by one, the limited set of scientific articles with the limited set of abstracts summarizing these articles – was attained in two computational passes with >97% hit rate for the training corpus only, >90% for the testing corpus only, >94% for the testing+training corpora combined and 92.9% for the testing corpus exploiting the knowledge acquired from training corpus. These results indicate that relative frequencies of individual words yield useful “features” for coupling the full text with its summarized counterpart.

Résumé

Une méthode simple, basée sur l'addition des probabilités d'occurence des mots, fut mise en place afin de proposer la solution à la deuxième tâche de Défi Fouille de Textes 2011. Cette tâche consistait à identifier à quel article scientifique – parmi un ensemble fermé d'articles - correspondait un résumé appartenant à un ensemble fermé de résumés. La formule que nous proposons atteignit >97% de taux de réussite quand évaluée sur le corpus d'apprentissage seulement, 90% quand évaluée sur le corpus de test seulement, >94% quand évaluée sur les deux corpora combinés, et 92.9% pour l'évaluation du corpus de test prenant en compte les distributions de probabilité obtenus à partir du corpus d'apprentissage. Ces résultats indiquent que les fréquences relatives des mots individuels peuvent être considerées comme les “traits” pertinents pour l'appariement des articles avec leurs résumés.

Mots-clés: appariement résumé / article, fréquence relative, mot vide, hapax
Keywords: full-text / abstract coupling, relative frequency, stopword, hapax

1 Introduction

The goal of the second task of challenge Défi Fouille de Textes (DEFT) 2011 was to couple N scientific articles with N abstracts summarizing the contents of the respective articles, where N=300 for the training corpus and N=200 for the testing corpus. Due to the lack of resources, our team had decided to try to ignore more state of the art approaches based on Artificial Neural Networks, Support Vector Machines, Nearest Neighbors or boosting and decided to confront the DEFT corpora with a very simple, yet intuitively[1] appealing idea.

The basic idea is simple: since abstract is nothing else than the condensation of the full-text, it can be stated that if the word W occurs in the abstract, it will tend to occur in the full-text as well; and more it is present in the abstract, more would one expect to find it in the associated full-text. Since opposite is also the case - i.e. the more W is present in the full-text F , the more one would tend to associate with F the abstract A within which W is frequent - one can reason even

[1] The reason why there are no references in this article is that the method hereby presented does not follow any antecedent study but was based upon pure intuition.

further: if W is frequent in abstract A and full-text F_1 but not frequent in full-text F_2, one would tend to couple A with F_1 and not with F_2.

We have supposed that such “surface information” as relative word frequencies of diverse words are the only “features” needed in order to obtain the list of most plausible [abstract, full-text] candidate couples.

2 Method

The method hereby proposed is based upon very simple frequency counting which occurs in two passes. In the first, so-called “article pass”, the total frequency $F_{W,total}$ for every word W in all articles, as well as $F_{W,A}$ denoting a number of occurences of the word W in article A are calculated. Therefore, after this first pass, our algorithm can calculate the relative frequency:

$$P_{W,A} = F_{W,A} / F_{W,total}$$

i.e. “relative probability of word W occuring in article A when compared with the rest of the corpus”. It is evident that the highest $P_{W,A} = 1$ will be obtained in case of hapaxes (i.e. $F_{W,A} = 1$; $F_{W,total} = 1$; $P_{W,A} = 1 / 1 = 1$) and “relative hapaxes” (i.e. the words like proper nouns which occur in one article only; $F_{W,A} = N$; $F_{W,total} = N$; $P_{W,A} = N / N = 1$).

In the second, so-called “abstract pass”, the algorithm calculates the overall score for every possible (abstract, article) couple by taking, one after another, every word W from every abstract, and, **adding** the to the $P_{W,A}$ value for every possible article A. Thus all the mathematical operations of our method are contained within this line of simple PERL code:

```
$abstract_article{$abstract}{$article}+= ($word_freq_in_article{$word}{$article} / ($word_total{$word})) if $word_total{$word};
```

In other words, the final score for a possible [abstract, article] couple is obtained as **a sum of all $P_{W,A}$** where W is every word present in the abstract-being-scored, and A is any article whatseover. Finally, [abstract, article] couples are sorted in descending order and every abstract is coupled with the article from the top of the list, i.e. having the highest score.

3 Results

The overall results of our tentatives are presented in the Table 1. After having being motivated by encouraging results (>97%) obtained by application of our method upon the training corpus (N=300) , we have applied the same method upon the testing corpus (N=200). While the test yielded 90% hit rate, it earned us the last place in DEFT competition.

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

Unfortunately, it was only after the announcement of the oficial DEFT results that we have found time to vary our method. Firstly, one can notice (c.f. row 4 of Table 1) that the keeping of the total word frequencies learned from the training corpus can increase the efficacity of our additive scoring when applied upon testing corpus. Secondly, by implementing a list of stopwords from CPAN's Lingua ::StopWords package, we raised the hit rate of our simple formula to 97% which we consider to be a satisfying result, given the simplicity of our approach.

4 Discussion

The theoretical framework of our approach can be characterized as follows: the summarization process during which an author condenses the knowledge contained in the full text characterized by a word-frequency histogram F into much shorter abstract characterized by a word-frequency histogram A can be understood as a surjection of F onto A. In other terms, there exists a mapping function between F and A and the approximative knowledge of this function could allow us to find & evaluate the best candidate (F, A) couples.

Our naïve supposition stating that a very simple unweighted addition of relative probabilities of all the words present in the abstract could be considered as a sufficiently adequate approximation of such a mapping function, allowed us to obtain more than 90% of correct couplings within the scope of the test corpus, nonetheless our team ended up as last in this task of the DEFT competition for which 3 teams have attained 100% hit rate.

In spite of being last in the DEFT competition, we think that our proposal have certain properties which make it worth of interest. Firstly, the method proposed hereby is fully deterministic, no stochastic or quasistochastic element is involved in the scoring nor in the subsequent [article, abstract] coupling. Secondly, diversification of results after exclusion of "stopwords" indicates that the method hereby proposed possibly disposes of various parameters which could be possibly tuned in order to obtain 100% accuracy. Thirdly, the fact that no external corpus was needed in order to obtain the results which were obtained indicates that at least for a certain class of text-summarization problems, one is not obliged to implement "state of the art" machine learning techniques in order to construct robust statistical models of semantic spaces – , but can stick to more empiric "surface features" like relative word frequencies. It is evident that the calculation of such surface features demands less computational resources than more sophisticated techniques.

Lastly, we think that the method, *the idea* proposed hereby could be successfully applied not only upon corpora written in french language but could be used to couple abstracts and articles written in any non-inflectional language like English, for example. We think that this is possible because, as our results indicate, there exist certain quantitative correlations, certain observable morphisms, between distribution of symbols in full-text and distribution of symbols in related abstract.

5 The code

```
#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";
```