Statistical methods for retrieving most significant paragraphs in newspaper articles

José Abraços
Departamento de Informática, Faculdade de Ciências e Tecnologia / UNL
2825 Monte da Caparica, Portugal
jea@di fct unl pt

Gabriel Pereira Lopes
Departamento de Informática, Faculdade de Ciências e Tecnologia / UNL
2825 Monte da Caparica, Portugal
gpl@di fct unl pt

Abstract

Retrieving a most significant paragraph in a newspaper article can act as a kind of summarization It can give the human reader some hints on the contents of the article and help him to decide whether it deserves a full reading or not It may also act as a filter for a robust natural language understanding system, to extract relevant information from that paragraph in order to enable conceptual information retrieval

Taking a newspaper article and a base corpus, word co-occurrences with higher resolving power are identified These co-occurrences are used to establish links between the paragraphs of the article The paragraph which presents the larger number of links to other paragraphs is considered a most significant one

Though designed and tested for the Portuguese language, the statistical nature of our proposal should ensure its portability to other languages

1. Introduction

The advantages of using statistical methods when dealing with large volumes of text are known Namely, their capability of facing any kind of subjects, without fearing the most baroque syntactical structures, and always producing an answer which, though varying in liability, is always more useful than "fail"

The scope of the present work is the use of statistical methods to retrieve a most significant paragraph in a newspaper article The method we propose might help a reader in getting a quick glimpse of the contents of a newspaper and deciding which articles deserve a full reading It can besides facilitate searches through journalistic text bases But we are also interested on pruning the amount of text to be automatically processed for robust understanding of natural language This will enable conceptual based document representation and conceptual information retrieval (Mauldin 1991)

The process is based on retrieving the co-occurrences with higher resolving power in each document, using them to establish links between paragraphs, and selecting the paragraph with more links to other paragraphs

Tests performed with the support of a base corpus of about 500 thousand words were able to identify a most significant paragraph in 7 out of 10 newspaper articles We present, in annex, the results of some experiments concerning one of the articles

2. Antecedents

An idea borrowed from Information Retrieval, is that a term will be so more relevant in a document the more frequently it occurs in that document, and the less frequently it occurs in a base corpus

Maarek (1992), following other authors, considers that using pairs of words as an indexing unit is more adequate to information retrieval than using single words Intuitively, it is plausible to admit that, for instance, the pair [*file system*] is far more informative than the words *file* and *system* taken in isolation Maarek aims at retrieving pairs of lexically related words In English, 98% of the lexical relations occur between words within a span of 5 words in a sentence, i e, the window to consider when extracting words related to word *w*, should span from position *w*-5 to *w*+5 Maarek also defines the *resolving power* of a pair in a document *d* as

$$\rho = -P_d \log P_c$$

where P_d is the observed probability of appearance of the pair in document *d*, P_c the observed probability of the pair in corpus, and *-log* P_c the quantity of information associated to the pair It is easily seen that ρ will be higher, the higher the frequency of the pair in the document and the lower its frequency in the corpus, which agrees with the idea presented at the beginning of this section

Church and Hanks (1990) propose the application of the concept of mutual information

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

to the retrieval, in a corpus, of pairs of lexically related words They also consider a word span of ±5 words and observe that "interesting" pairs generally present a mutual information above 3

Salton and Allan (1995) focus on paragraph level Each paragraph is represented by a weighed vector, where each element is a term (typically, word stems, after excluding those in a *stop list*) The weight of each term reflects (as usual) positively its frequency in the document and negatively its frequency in the corpus Using a measure of similarity between vectors and applying a similarity threshold, one can define which paragraphs are linked They then consider of central importance the paragraph with the largest number of connections to other paragraphs

The idea underlying the present work was to integrate these 3 approaches and to apply the resulting method to newspaper articles, with the purpose of retrieving, in each article, a most significant paragraph

3. The proposed approach

As stated before, the method of Church and Hanks identifies pairs of lexically related words So, for instance, the pair [*conselho segurança*] (security council), with an associated mutual information of 5.3, can be considered as a potential indexing term, while the pair [*para a*] (to the), though 63 times more frequent in our corpus, having a mutual information of 0.7, can be excluded We have then a criterion for exclusion, that dispenses with the need for stop lists, and that aims at assuring the existence of a lexical relation between the words of the remaining pairs

But not all pairs of lexically related words are good indexing terms of a document The pair should also meet the requirement of being relevant in the considered document The method of Maarek proposes a measure of the resolving power of each pair in the concerned document, thus enabling the selection, among all the potential indexing terms, of those that are relevant in each document For instance, [*estados unidos*] (united states) has a high mutual information (8 1) but it can be of little relevance in an article about the liberation of prisoners by the Serbs of Sarajevo (ρ=0 007) The experiments carried out point to a threshold of the resolving power around 0 01 We consider as relevant in a document only the pairs with a resolving power above this threshold

When the same pair occurs in different paragraphs of the same document, links can be established between those paragraphs At this point, we only consider pairs that were not excluded in previous steps (mutual information ≥ 3 and resolving power $\geq 0\ 01$) Though, each link is not limited to pairs of words In fact, the wider the link, the higher its relevance After processing a document, we often get overlapping pairs For instance, in an article where the expression *dos três antigos beligerantes* (of the three former contenders) is used repeatedly, the following pairs were retrieved

[*três beligerantes*] [*antigos beligerantes*] [*dos beligerantes*][1]

By observing the overlap of these pairs in the very document, a single link can be retrieved, in the form of the tuple [*dos três antigos beligerantes*]

Adapting the method of Salton and Allan, we can formulate the hypothesis that the paragraph with the larger number of links to other paragraphs would be of central importance in the document

In summary, the steps of the proposed method are

- in a base corpus, compute the frequency of each word and the frequency of each co-occurrence, considering a window spanning from position w-5 to w+5,
- to each document compute, similarly, the frequency of each word and each co-occurrence,
- exclude, from the co-occurrences identified in the document, those presenting a mutual information or a resolving power under the defined thresholds ($I(x,y) < 3$ or $\rho < 0\ 01$),
- take the selected pairs and group the overlapping ones, the resulting tuples (pairs and groups of pairs) occurring repeatedly in different paragraphs establish links between those paragraphs,
- hypothetically, the paragraph presenting a larger number of links to other paragraphs will be of central importance in the document

It should be noted that this proposal, compared to Salton and Allan's, has the advantages (at least in theory) of avoiding the use, always arbitrary, of stop lists[2], and of basing the calculations exclusively on the tuples that are relevant in the document, instead of using the heavy vectors containing all the terms of each paragraph We don't have, so far, enough data to make any claim about the comparative quality of the links

---

1 pairs [*três antigos*] [*dos antigos*], though considered relevant, didn't score enough mutual information to be selected

2 the relevance of a word depends on the context, so, we prefer not to *a pnon* exclude any word, by sending it to a *stop list* In fact, some of the tuples we retrieved as relevant include words that would otherwise be part of such a list An example is the pair [*não alinhados*] (nonaligned) where the word *não* (not) though quite significant in context, would be excluded via *stop list*

4. Applying the proposal

The base corpus was initially built with news from Lusa news agency, in a total of 216 319 words Later, news from "O Público" newspaper (about 90 000 words) and more news from Lusa were added, and the total reached 537 085 words The consequences of this enlargement will be discussed in the next section

Experiments were made with 10 articles from "O Público", that didn't belong to the corpus

Both the corpus and the documents were subjected to a very elementary pre-processing, which basically consisted of

- converting all uppercase letters to lowercase
- converting all numbers to NUMERO (NUMBER)[3]
- eliminating all non-letter characters

Words or co-occurrences not present in the corpus, if occurring in a document, would lead, respectively in the computation of mutual information or resolving power, to dividing by 0 or to $\log_2 0$ To prevent this situation, in such cases, and only for calculation purposes, the document is added to the corpus By doing so, though, the mutual information becomes overestimated For instance, the pair [*na eslavónia*] (in slavonia) occurs 3 times in an article As *eslavónia* doesn't occur in the corpus, the article is added to the corpus, for calculation purposes only concerning this pair The result is the presupposition that, despite the quite low frequencies of *eslavónia* and [*na eslavónia*], almost every time the word *eslavónia* occurs it is preceded by *na*, the mutual information of the pair being then artificially raised

To overcome this overestimation, 2 additional mutual information thresholds were defined

- if one of the words (or both) doesn't occur in the corpus, it must be $I(x,y) \geq 10$,
- if both words occur in the corpus but they never co-occur, it must be $I(x,y) \geq 8$

These limits are not definitive They were suggested by the experiments carried out, which were though too few to ensure their definition with certainty

The articles analyzed in those experiments are in average 500 words long Pre-processing and frequency calculations are obtained through *gawk* commands (Unix) The calculations of mutual information, resolving power and the filtering of co-occurrences through these criteria are implemented in C

Nevertheless, given the experimental nature of the system, optimization was no main concern Searches in the file containing the co-occurrences of the corpus (22 MB) are sequential, this source of inefficiency being only palliated by previously sorting the co-occurrences by decreasing order of probability In what concerns the article presented in Annex A (441 words), pre-processing, calculation of frequencies and sorting takes about 5 seconds The calculations involved in selecting and sorting co-occurrences take about 8 minutes[4] By the characteristics of the implementation, this last time is directly proportional, among other factors, to the amount of words in the corpus and to the amount of unknown words that occur in the document

Out of the 10 articles that were analyzed, the method we propose achieved the identification of the most significant paragraph in 7 and was clearly mistaken in 1 In the remaining 2 articles, there doesn't seem to be, intuitively, a most representative paragraph This intuition in the evaluation of the results is necessarily subjective

Notwithstanding the very small number of articles involved in this test, it may be curious to compare our results with those that would be obtained by just picking up the 1st paragraph of each article, or even both the 1st and the 2nd paragraphs

# of articles	our proposal	1st §	1st§+ 2nd§
retrieves a most significant §	7	5	6
existance of a most significant § is not clear	2	2	2
the § retrieved is not a most significant one	1	3	2

5. Discussion of the results

The proposed method ignores a series of basic questions, namely

Lemmatization

All the calculations are made without any attempt of unifying plural forms with singular forms, different conjugations of a same verb, etc Nevertheless, it doesn't look clear that new links, obtained by grouping words that, though sharing a common stem, were in fact used in distinct forms, will necessarily increase the performance of the system Would it make sense to unify *tribunal de família* (court that deals with family cases) with *tribunal familiar* (familiar court)? And

[3] the choice of reducing all numbers to NUMERO has to do with the kind of documents under study, in texts about law, for instance, the distinction between Law 12/86 and Law 47/95 may be important

[4] times measured in a DECstation 5000/200

direitos do homem (human rights) with *direito dos homens* (law of men)?

Anaphora resolution

Though the unification of the anaphor with the antecedent, in most cases, makes obviously sense, anaphora resolution would require a complete analysis of the text, totally outside the scope of this proposal Curiously, in the only experiment that was made of full anaphora resolution, the number of links between paragraphs substantially increased, but the paragraph retrieved as most significant - the first - was no longer the one obtained by intuition - the second[5] - (refer to results in the annex)

Unification of synonyms, hyponyms, hyperonyms

The same arguments presented about lemmatization can apply here The experiment of unifying initials with full names - e g *ONU* ⇔ *Nações Unidas* (UN ⇔ United Nations) - simple to do with the help of a thesaurus, gave similar results to those of applying anaphora resolution

Size of the co-occurrence window

The window spanning from position *w-5* to *w+5*, defined for English language, may be not the most adequate to Portuguese No further experiments were performed with other sizes of windows

Indexing terms

The resolving power criterion aims at assuring that the selected co-occurrences are relevant in the document being analyzed A manual indexing could, nevertheless, choose other terms, possibly even foreign to the document In fact, in an article describing a coup there may be references to *derrube de governantes* (overthrowing of rulers), *tomada do poder* (taking the power), without any explicit occurrence of the expression *golpe de estado* (coup)

We present, in annex, the results of processing a document using the initial corpus (216 319 words) and the augmented one (537 085 words) In what concerns the co-occurrences that were selected as establishing links, one can notice the exclusion of [*da ONU*] (of the UN) in the 2nd case (in the 1st case it already presented a mutual information very close to the threshold) All the other selected co-occurrences remain, and their ordering in terms of resolving power is also preserved The paragraph retrieved as central is the same Both corpora are though quite small Nothing indicates that the results would stand a more substantial increase of the corpus

We also tried to find out how far establishing links could help in identifying a structure of the text The structures obtained, by connecting iteratively each new paragraph to the one with more links in common, are not conclusive In some cases they are close to a possible intuitive structure of the text, while in other they diverge considerably The structure obtained for the text in annex was among the most plausible

6. Conclusion

The methodology we propose integrates the concepts of mutual information associated to a pair of words, resolving power of that pair in a document and establishing of links between paragraphs of a document, with the purpose of retrieving a most representative paragraph

The methods we use are purely statistical Nevertheless, notwithstanding their simplicity, the rough simplifications referred in the previous section and the exiguousness of the corpus, the results seem quite interesting The liability of these results is though limited by the amount of tests that were performed and by an evaluation based on the intuition of the authors

Probably, an increase of the corpus and the refinement of the process with some, even elementary, linguistic criteria, would benefit the performance

Though designed and tested for the Portuguese language, the statistical nature of this methodology should ensure its portability to other languages

References

Church, K and Hanks, P (1990) Word association norms, mutual information, and ' lexicography *Computational Linguistics, 16 (1)*, p 22-29

Maarek, Y (1992) Automatically constructing simple help systems from natural language representations In P Jacobs Ed , *Text-based intelligent systems current research and practice in information extraction and retrieval*, Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey, p 243-256

Mauldin, M (1991) *Conceptual information retrieval - a case study in adaptative partial parsing*, Kluwer Academic Press, Dordrecht

Salton, G and Allan, J (1995) Selective text utilization and text traversal *International Journal of Human-Computer Studies, 43*, p 483-497

[5] In fact, in this article, central information seems to concentrate in the 2 initial paragraphs, the 2nd reiterating most of the information introduced by the 1st The 2nd refers the 2 actors (UN and NATO) whose actions will be analysed latter This may suggest some preference against the 1st Anyway, each one of these 2 paragraphs can be considered as representative of the text

Annex A - Results of the experiments, relative to one of the target newspaper articles

Full text of the article, the selected co-occurrences (relative to the larger corpus) are underlined

Capacetes azuis vão ser protegidos pela NATO (O Público, 17Jan96)

ONU aprova missão na Eslavónia

O Conselho de Segurança da ONU decidiu segunda-feira à noite estabelecer uma administração transitória, apoiada por uma operação de manutenção da paz, na região da Eslavónia Oriental, último território no interior das fronteiras administrativas da Croácia ainda controlado pelos independentistas sérvios

Para a missão, com a duração prevista de um ano, vão ser disponibilizados inicialmente até cinco mil capacetes azuis, com regras de envolvimento limitado mas que poderão beneficiar de uma protecção da NATO Esta operação, já designada "Administração Transitória das Nações Unidas para a Eslavónia Oriental" (UNTAES), foi aprovada por unanimidade pelos 15 membros do Conselho de Segurança

Trata-se da primeira decisão concreta para a aplicação do plano de paz destinado a esta sensível região da Croácia que faz fronteira com à Voivodina sérvia, após o acordo do passado dia 12 de Novembro entre a Croácia e representantes dos sérvios locais, concluído à margem das conversações de Dayton (Estados Unidos) sobre a Bósnia-Herzegovina

A administração transitória da Eslavónia Oriental - atravessada pelo Danúbio, limite natural entre a Sérvia e a Croácia e o grande eixo fluvial da região, incluindo nas ligações com a Hungria - vai ser confiada ao diplomata norte-americano Jacques Klein, antigo oficial da Força Aérea dos EUA e que se tornará numa espécie de "governador" deste fértil território, cerca de cinco por cento da superfície da Croácia Na qualidade de administrador provisório, Klein possui autoridade máxima sobre as componentes civil e militar da missão da ONU

Desde meados de Dezembro que a ONU e a NATO decidiram "repartir" a sua intervenção na ex-federação jugoslava A Aliança Atlântica desempenha actualmente uma função determinante na Bósnia, ao dirigir a operação "Esforço Concertado", enquanto a ONU mantém o comando das operações na Croácia e Macedónia Esta decisão do Conselho de Segurança põe oficialmente termo à fracassada Operação das Nações Unidas para o Restabelecimento da Confiança na Croácia (ONURC), cujos efectivos foram retirados na sua quase totalidade durante o ano passado, na sequência das ofensivas militares croatas na Eslavónia Ocidental, em Maio, e na Krajina, em Agosto

Nos termos do acordo assinado em Dayton, esta região deverá ficar totalmente desmilitarizada 30 dias após a instalação efectiva no terreno da força da ONU, e prevê-se um período de transição máximo de dois anos, findo o qual a região deverá regressar ao controlo efectivo da Croácia Para os especialistas militares das Nações Unidas envolvidos nesta operação, a tarefa mais difícil consistirá em convencer os cerca de 20 mil milicianos sérvios fortemente armados a entregarem as suas armas e aceitarem o regresso da autoridade de Zagreb à região

Co-occurrences repeated in the document, with $I(x,y) \geq 3$ (possible links)

ρ	I(x,y)	$f_d(x,y)$	$f_c(x)$	$f_c(y)$	$f_c(x,y)$	x	y	
0 031037	3 001719	5	4474	99	82	da	onu	*of the UN*
0 025119	11 344917	3	0	10	0	eslavónia	oriental	*East Slavonia*
0 025119	*9 703371	3	52	0	0	administração	transitória	*transitory administration*
0 025119	*3 887126	3	1758	0	0	na	eslavónia	*in Slavonia*
0 022752	5 355113	3	151	70	10	conselho	segurança	*Security Council*
0 020184	4 854552	3	1006	55	37	das	nações	*of the Nations*
0 019714	8 907315	3	55	77	47	nações	unidas	*United Nations*
0 019371	4 967027	3	1006	77	56	das	unidas	*of the United*
0 017277	10 076007	2	16	5	0	croácia	sérvios	*Croatia Serbs*
0 017277	*4 257744	2	61	74	0	região	deverá	*region should*
0 017277	*4 107468	2	1006	0	0	das	eslavónia	*of Slavonia*
0 017277	*3 266078	2	561	16	0	entre	croácia	*between Croatia*
0 015837	12 078945	2	6	10	6	capacetes	azuis	*blue helmets*
0 015168	4 065371	2	73	354	10	vão	ser	*will be*

corpus 216 319 words

0 026902	10 807156	3	0	36	0	eslavónia	oriental	*East Slavonia*
0 026902	*9 554176	3	143	0	0	administração	transitória	*transitory administration*
0 026902	*3 889618	3	4352	0	0	na	eslavónia	*in Slavonia*
0 023081	5 309673	3	325	175	21	conselho	segurança	*Security Council*
0 020268	5 054372	3	2351	121	88	das	nações	*of the Nations*
0 020018	9 199451	3	121	151	100	nações	unidas	*United Nations*
0 019778	5 095578	3	2351	151	113	das	unidas	*of the United*
0 019371	6 565658	2	54	21	1	croácia	sérvios	*Croatia Serbs*
0 018465	*4 193060	2	2351	0	0	das	eslavónia	*of Slavonia*
0 015589	12 012423	2	18	26	18	capacetes	azuis	*blue helmets*
0 015387	3 817080	2	169	947	21	vão	ser	*will be*

corpus 537 085 words

ρ - resolving power
I(x,y) - mutual information
$f_d(x,y)$ - frequency of pair x,y in the document
* pairs eliminated by overestimate adjustment (cf section 4)

$f_c(x)$ - frequency of word x in the corpus
$f_c(y)$ - frequency of word y in the corpus
$f_c(x,y)$ - frequency of pair x,y in the corpus

Matrixes showing the number of links between paragraphs, and structures obtained by connecting each new paragraph to the preceding one with more links in common The connections used in the construction of the structure (cf section 5) are signaled in bold face

without anaphora resolution

§	1	2	3	4	5	6	Σ
1		2	1	1	1	0	5
2	**2**		0	1	2	1	6
3	1	0		0	0	0	1
4	1	**1**	0		0	0	2
5	1	**2**	0	0		1	4
6	0	1	0	0	1		2

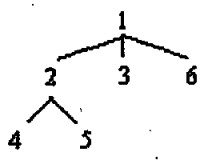
da ONU ⇔ das Nações Unidas
(of the UN ⇔ of the United Nations)

§	1	2	3	4	5	6	Σ
1		3	1	2	2	1	9
2	**3**		0	2	2	1	8
3	**1**	0		0	0	0	1
4	2	**2**	0		1	1	6
5	2	**2**	0	1		1	6
6	1	1	0	1	**1**		4

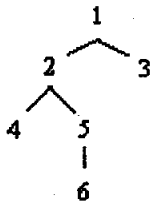
with full anaphora resolution

§	1	2	3	4	5	6	Σ
1		5	3	5	4	4	19
2	**5**		1	4	3	3	16
3	**3**	1		2	3	3	12
4	**5**	4	2		2	4	17
5	**4**	3	3	2		2	14
6	4	3	3	**4**	2		16

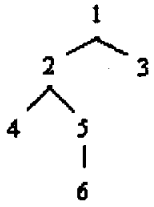
intuitive structure



without anaphora resolution



da ONU ⇔ das Nações Unidas



with full anaphora resolution