

TOPIC CONTINUITY IN DISCOURSE AND AUTOMATIC CLASSIFICATION OF DIGITAL DOCUMENTS

Gabriel de Ávila Othero

ABSTRACT: In this paper, we will discuss the importance that studies in functional syntax dealing with topic continuity and flow in discourse can have for applications in NLP, specifically in relation to the automatic classifiers of digital texts and documents. We analyze the methods used in the works of automatic classification of digital texts presented in Langie (2003) and Langie & Lima (2003) and show that, if some resources were developed taking into account functional theories about discourse topic continuity in the process of automatic classification, the precision of the automatic text classifiers could be more accurate.

KEYWORDS: Topic continuity; reference; automatic classifiers.

RESUMO: Neste artigo, iremos discutir a importância dos estudos em sintaxe funcional que lidam com a continuidade tópica no discurso. Veremos que esses estudos são importantes para aplicações de PLN, especialmente no desenvolvimento de classificadores automáticos de documentos digitais. Analisaremos os métodos usados nos trabalhos de classificação automática de documentos digitais apresentados por Langie (2003) e Langie & Lima (2003). Iremos mostrar que, se fossem consideradas as teorias funcionais sobre a continuidade tópica dos referentes no discurso, a precisão dos classificadores automáticos poderia ser maior.

PALAVRAS-CHAVE: Continuidade tópica; referência; classificadores automáticos de texto.

Introduction

Companies, universities and research centers are using digital documents more and more – such as virtual books, on-line articles, virtual databases, banks of theses, etc. Besides resolving the problem of physical space, the use of digital texts enables the practically instantaneous exchange of documents between institutions from virtually any place on the planet. Entire libraries are appearing on the Internet, containing exclusively virtual books (or e-books)¹. Besides that, newspaper and magazine websites are starting to put all their texts available on-line².

Nevertheless, while more digital documents are emerging on the Internet and entering the

¹ Cf. Project Gutenberg (www.gutenberg.net), for example.

² Cf. www.bbc.com and www.lemonde.fr, for example.

day-to-day activities of companies and academic institutions, a great challenge emerges: the ability to classify and organize these digital documents in an efficient manner. Using manual classification can be quite costly, slow and demanding for the institutions. A task and – and a challenge – that presents itself to researchers from Natural Language Processing (NLP) and Computational Linguistics is that of the automatic classification of digital documents. After all, once on the computer, why not use *software* that does the arduous work of reading an entire document before classifying it according to its theme, area or subject? That is, since we have virtual books and digital documents, why don't we also have automated classifiers and programs that can work as actual “digital librarians”?

The automatic or semi-automatic classifiers seem to be the key to solve the problem of dealing with enormous quantities of digital texts and documents. Besides being agile and quick, these kinds of programs in general do not represent large costs to institutions, companies, or even to private users. However, automatic classification is still in its initial stages, since it has emerged to solve a relatively recent problem.

In this paper, we will discuss the importance that studies in functional syntax dealing with topic continuity in discourse can have for applications in NLP, specifically in relation to the automatic classifiers of digital texts and documents. In section 1, we analyze the methods used in the works of automatic classification of digital texts presented in Langie (2003) and Langie & Lima (2003). Our goal is to show that, if some resources were developed taking into account functional theories about discourse topic continuity in the process of automatic classification, the precision of the automatic text classifiers could be more accurate.

In section 2, we see how topic continuity works in discourse. We will be based mainly on the ideas of Givón (1979, 1992) and Ariel (1988). We will propose the adoption of these ideas by programmers, when developing programs that classify digital texts regarding their topic or subject. We show that important words in topic continuity (in particular personal subject pronouns) are underestimated or even ignored by programmers in the development of tools for automatic classification of texts.

1. Automatic Classification of Texts

The Automatic Classification of Texts can help institutions that deal with large quantities of digital documents. Automatic or semi-automatic classification programs can help virtual and

conventional libraries. According to Langie (2003, p. 6), the Automatic Classification of Texts can be defined in the following way³:

The categorization of texts is the task of attributing a Boolean value {T, F} for each pair $(d_j, c_i) \in D \times C$, where $D = \{d_1, \dots, d_{|D|}\}$ is a set of documents and $C = \{c_1, \dots, c_{|C|}\}$ is a predefined set of categories.

In other words, automatic classification consists in classifying a digital document in a determined category, according to criteria established *a priori* by human programmers. This classification generally respects a hierarchical structure, using “a tree of categories, allowing documents to be classified on the leaves as well as on the intermediary nodes” (Langie & Lima, 2003, p. 3).

Langie & Lima (2003, p. 2) explain that

With the use of a hierarchical category structure, the process of classification can be decomposed in smaller processes, in which the quantity of variables involved is reduced. [According to Koller & Sahami (1997)], categories that are near, inside the hierarchical structure, have in general more similar features than other categories.

Thus, a word like *linguist* may not be very clear to help classify a text between the categories *Syntax*, *Semantics* or *Phonology*, but it can be a good attribute to differentiate texts between larger categories, like *History*, *Medicine* or *Linguistics*.

See an example of a category tree adapted from Langie & Lima (2003, p. 5):

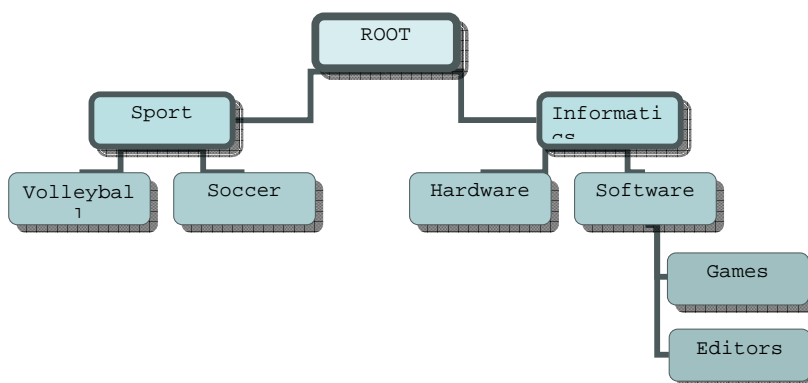


Figure 1: Example of a category tree

³ Langie based his definition on Sebastiani (2002).

Basically, the task of an automatic classifier is to “read” the texts and classify them into categories or subcategories pre-established by human programmers. During the “reading” of the text, the program will compare a series of words that appear in the target-text with the words it already knows. From there on, it will try to classify this target-text within a category, based on the number of occurrences of the keywords that appear in the document. Within these words that the classifier knows, there are words with more semantic value and others with less semantic value. This means that some words receive higher values than others on a scale of importance in the classification of the text: “the importance of the composition of vectors is the combination of factors known as TFC [Salton e Buckley, 1988]. In this combination, more important tokens receive importance-weight near to 1 and less important tokens receive values close to zero” (Langie & Lima, 2003, p. 8).

Still according to Langie & Lima (2003, p. 8), these are the main phases through which digital documents pass until they are duly classified:

Before their representations are generated, documents pass through a *pre-processing* phase. In this phase, the words of the document are changed to small characters, and punctuation characters and numbers are eliminated. Besides, a selection of attributes with the removal of a list of stopwords (articles, adverbs, conjunctions, numerals, prepositions, pronouns, and copula verbs) is applied. The stopwords list contains 365 words. Attributes are also selected with the removal of words whose frequency of the document (FD) is inferior to 3. The FD of a t token corresponds to the number of documents in which t appears at least once. [According to Yang & Pedersen (1997)], the principle of attribute selection using FD is that rare words are not informative to predict the category of a given document and they do not act on the performance of the classifier.

We will not concentrate on the functioning of the automatic classifiers of texts themselves. What we want to call attention to here are basically two factors about the procedure of the classifiers: (i) the so called *stopwords list*; and (ii) the criteria used by the programmers to classify a word as “relevant” or “non-relevant” in the determination of the subject of the text.

In the stopwords list (check the appendix), one finds all those words that apparently do not play an important role in the automatic classification of the texts. They are words that, in principle, should not present a relevant semantic value, having, in general, a merely functional role in the text. On this list we can see words like articles, adverbs, conjunctions, numerals, prepositions, *pronouns*, and copula verbs. However, as we will see in the next section, we firmly

believe that the pronouns are the main mechanism for topic continuity in discourse. They have a fundamental value for any classifier that is based on the number of occurrences of a determined term in the text to proceed with its classification.

Another interesting factor, that we will also analyze in the next section, has to do with the classification criteria of the *relevant* or *non-relevant* words. This question is intrinsically related to the previous one, since, as we will see, a discursive topic is rarely repeated: normally it is (co-)referred to by way of lexical anaphora, or, more frequently, by way of pronominal or elliptical anaphora.

2. On Topic Continuity

Within the functionalist framework – as presented in Givón (1979, 1993, 1995), DeLancey (2001), among others – syntax is seen as a strategy of phrasal organization for a determined communicative purpose. We, speakers, organize the information of a text phrase by phrase, by way of nominal expressions, which can serve as topics, and their anaphoric expressions, which function normally to maintain the topic continuity. In this textual organization, nominal expressions help us to store, file and organize the new information which “comes in” from text; while the anaphoric expressions (mainly via ellipsis or pronouns) will serve to maintain the file of the discursive topic open in the working memory.

While there are still some controversies about what it really means to be a topic, or about whether topicality is scalar or not (cf. Givón, 1979, 1992, Pontes, 1986, and Ariel, 1988), we will define the topic as *that which is talked about*. To this succinct definition, however, we should add another, based on the ideas of Givón: *a topic is said to be a topic if it is maintained as such in a succeeding series of sentences*.

According to Barbisan & Machado (2000, p. 72), “topics are labels of files to storage in the episodic memory. These labels are seen as perceptually prominent by the speaker, who incrust important information there to the listener”.

While the nominal expression “opens” these new files in our memory, introducing new topics, pronouns normally serve to refer (or co-refer) to the topic, maintaining the topic chain, and guaranteeing *topic continuity and maintenance* in the discourse.

Below we see a graphic on the process of referencing and topic continuity:

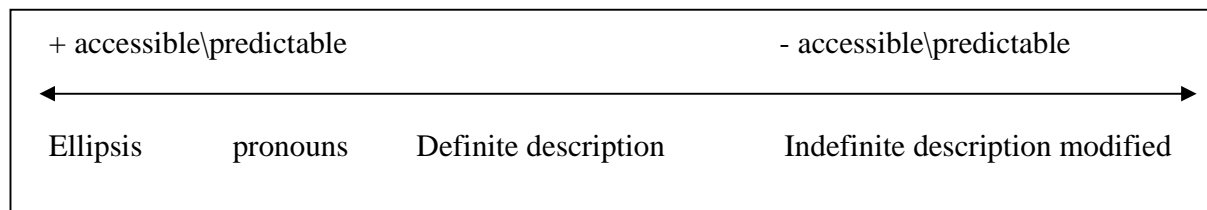


Figure 1: Quantitative scales of referential identification (contextual resources to designate the referent)

In the process of topic continuity, we use different strategies of topic maintenance. Ariel (1988) shows that, despite the diversity of strategies for topic maintenance, the distribution of topic markers in anaphoric resolution is not exactly flexible. She argues that the pronominal anaphora resolution is widely used, especially when the topic has maximum accessibility and minimum distance. Several studies (cf. Ariel, 1988, Barbisan & Machado, 2000, Givón, 1992, among others) have shown that pronominal anaphora resolution is definitively the most used strategy in discourse topic continuity. This lead us immediately to answer the two questions we tackled in section 1:

(i) The list of stopwords contains pronouns, not taking them into consideration during the process of text classification.

(ii) A word is considered relevant if it is not rare and is not found in the list of stopwords. However, if a topic (which is relevant to the text) is referred exclusively by pronouns, will it be considered relevant? The answer is an incredible “NO” for any automatic text classification tool that cannot count on anaphoric pronoun resolution.

The question that comes to mind of the linguist is the following: how does an automatic classifier of digital documents intend to classify texts basing themselves on the frequency of the most important and/or relevant words if, at the same time: (i) the expressions considered most important – the topics – are normally referred to by pronominal anaphora (at least in short and medium distances); and (ii) the tool, besides not recognizing the pronominal anaphoric relation, leaves all the pronouns in a list of rare and irrelevant words, called the *stopwords list*.

3. Possible Paths

It escapes the specific proposal of this article to test the hypotheses discussed here in an empirical manner. However, we truly believe that this is the necessary path which the automatic classification tools for digital documents should take. After all, as we know, a text is not composed of nominal expressions and their exhaustive repetition. It is elaborated in a different manner: topic is normally introduced in the form of a nominal expression (usually an indefinite nominal phrase) and referred to by different types of anaphora, from ellipsis, to pronominal anaphora (most frequent cases), to the repetition of the nominal expression itself (often, through a definite expression).

Despite appearing to be the ideal path, our considerations take us to another big problem: that of pronominal anaphoric processing⁴. This is certainly a true challenge for the automatic tools that work with texts in natural language. However, joint forces between linguists, computational linguists and programmers are in the way of solving these issues.

References

- ARIEL, M. Referring and accessibility. *Journal of Linguistics*, 24, 1988.
- BARBISAN, L. B.; MACHADO, R. F. O tópico no texto argumentativo. *Letras de Hoje*, v. 35, n. 3, 2000.
- COULSON, Mark. Anaphoric reference. In: GREENE, Judith; COULSON, Mark. *Language understanding: current issues*. Buckingham: Open University Press, 1996.
- De ROCHA, M. A corpus-based study of anaphora in English and Portuguese. In: BORTLEY, S. P. & McENERY A. M. (eds.). *Corpus-based and computational approaches to discourse anaphora*, London: UCL Press, 1996.
- DeLANCEY, S. *On functionalism*. Lecture. LSA Summer Institute. Santa Barbara, 2001.
- FLIGELSTONE, S. Developing a scheme for annotating text to show anaphoric relations. In: LEITNER, G. (ed). *New directions in English language corpora: methodology, results, software*. Berlin: Mouton de Gruyter, 1992.
- GARSDIE, R.; FLIGELSTONE, S.; BOTLEY, S. Discourse annotation: anaphoric relations in

⁴ Cf. Fligelstone (1992), De Rocha (1996), and Garside, Fligelstone & Botley (1997).

corpora. In: GARSIDE, R.; LEECH, G.; McENERY, A. *Corpus annotation: linguistic information from computer text corpora*. London / New York: Longman, 1997.

GIVÓN, T. From discourse to syntax: grammar as a processing strategy. In: GIVÓN, T. (ed.) *Discourse and syntax*. New York: Academic Press, 1979.

GIVÓN, T. The grammar of referential coherence as mental processing instructions. *Linguistics*, 30, 1992.

GIVÓN, T. *English grammar – a function-based introduction*. Amsterdam / Philadelphia: John Benjamins, 1993.

GIVÓN, T. *Functionalism and grammar*. Amsterdam / Philadelphia: John Benjamins, 1995.

KOLLER, D.; SAHAMI, M. Hierarchically classifying documents using very few words. *Proceedings of ICML-97, 14th International Conference on Machine Learning*, 1997.

LANGIE, L. C. *Um estudo sobre a aplicação do algoritmo kNN à categorização hierárquica de textos*. M. A. Thesis. Computer Science Department, Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS, 2003.

LANGIE, L. C.; LIMA, V. L. S. Classificação hierárquica de documentos textuais digitais usando o algoritmo kNN. *I Workshop em Tecnologia da Informação e Linguagem Humana*. São Carlos, Brazil, 2003.

PONTES, E. *Sujeito: da sintaxe ao discurso*. São Paulo: Ática; Brasília: INL, Fundação Nacional Pró-Memória, 1986.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information Processing e Management*, Vol. 24, n. 5, 1988.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1, 2002.

SILVA, M. T. A codificação do tópico como SN DEF em textos narrativos e em textos argumentativos. *ReVEL*. Vol. 1, n. 1, 2003.

YANG, Y.; PEDERSEN, J. A comparative study on feature selection on text categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*, 1997.

Appendix: the stopwords list used by Langie (2003)

abaixo	cedo	disto
acima	cem	do
acola	certamente	dois
agora	cinco	dos
ai	cinquenta	doze
ainda	claramente	durante
al	cm	duzentos
alem	com	e
algo	comigo	ela
alguem	como	elas
algum	conciso	ele
alguma	conforme	eles
algumas	conosco	em
alguns	consequentemente	embora
ali	consigo	enquanto
alias	contigo	entao
ambos	contra	entre
andar	contudo	entretanto
anteontem	correntemente	especialmente
anteriormente	cr	essa
antes	cuja	essas
ao	cujo	esse
aonde	d	esses
aos	da	esta
aparentemente	daquilo	estar
apenas	das	estas
apesar	de	este
apos	debaixo	estes
apropriado	definitivamente	et
aquela	defronte	etc
aquelas	dela	eu
aquele	dele	ex
aqueles	demais	exatamente
aqui	dentro	exceto
aquilo	depois	exclusivamente
as	desde	f
assim	dessa	facilmente
ate	dessas	fazer
atraves	desse	felizmente
atualmente	desses	ficar
automaticamente	desta	finalmente
b	destas	fora
basicamente	deste	frequentemente
bastante	destes	g
bem	dez	geralmente
bilhao	dezeseis	h
bilhoes	diante	ha
bonito	difícilmente	haver
c	diretamente	i
cada	disponível	inicialmente
catorze	disso	inteiramente

isso	nenhuma	porque
isto	nessa	portanto
j	nessas	possivelmente
ja	nesse	posteriormente
jamais	nesses	pouca
jr	nesta	poucas
junto	nestas	pouco
k	neste	poucos
km	nestes	praticamente
l	ninguem	primeiramente
la	nisso	primeiro
las	nisto	principalmente
lo	no	prontamente
logo	nono	provavelmente
longe	normalmente	q
los	nos	quais
m	nossa	qual
maioria	nossas	qualquer
mais	nosso	quando
mal	nossos	quanta
mas	novamente	quantas
me	nove	quanto
mediante	novecentos	quantos
melhor	noventa	quarenta
menor	nunca	quase
menos	o	quatorze
mesmo	oitenta	quatro
meu	oito	que
meus	oitocentos	quem
mil	onde	quinto
milhao	ontem	quinze
milhares	onze	r
milhoes	opcionalmente	rapidamente
mim	os	realmente
minha	ou	s
minhas	outra	se
ml	outro	segundo
muita	p	seis
muitas	para	seiscentos
muito	parecer	seja
muitos	pela	sem
n	pelas	sempre
na	pelo	senao
nada	pelos	ser
nao	perante	sessenta
naquilo	permanecer	sete
necessario	pior	setecentos
nela	pois	setenta
nele	por	seu
nem	porem	seus
nenhum	porquanto	si

sim	W
so	X
sob	Z
sobre	zero
sp	a
sua	o
suas	ÿ
subito	
t	
tais	
tal	
talvez	
tambem	
tanta	
tantas	
tanto	
tantos	
tao	
te	
tel	
ter	
terceiro	
teu	
ti	
toda	
todas	
todavia	
todo	
todos	
tras	
tres	
treze	
trezentos	
trinta	
tu	
tua	
tudo	
u	
ultimamente	
um	
uma	
umas	
uns	
v	
varias	
varios	
vem	
vez	
vezes	
vinte	
voce	