The chunk as the period of the functions *length* and *frequency* of words on the syntagmatic axis

Jacques Vergne
GREYC - Université de Caen - France
Jacques.Vergne@info.unicaen.fr

Abstract

Chunking is segmenting a text into chunks, sub-sentential segments, that Abney approximately defined as stress groups. Chunking usually uses monolingual resources, most often exhaustive, sometimes partial : function words and punctuations, which often mark beginnings and ends of chunks. But, to extend this method to other languages, monolingual resources have to be multiplied. We present a new method : endogenous chunking, which uses no other resource than the text to be segmented itself. The idea of this method comes from Zipf : to make the least communication effort, speakers are driven to shorten frequent words. A chunk then can be characterized as the period of the periodic correlated functions length and frequency of words on the syntagmatic axis. This original method takes its advantage to be applied to a great number of languages of alphabetic script, with the same algorithm, without any resource.

Introduction

Chunking is a frequent segmentation step in many processing types : robust parsers, parsers of linear complexity (Vergne, 2000), computing stress groups and linking them in tts systems, to compute macro-prosody (Vannier et al., 1999), in automatic indexing, the chunk as another indexed grain above the word in the grain hierarchy, and in sub-sentential alignment, the chunk as an aligned grain.

The method we propose is based on the properties of the functions length and frequency of words on the syntagmatic axis. These two functions are correlated : integer, periodic, synchronous, in phase opposition, and their period allows to define the chunk. On a period, the length function is non-decreasing, and the frequency function is non-increasing. These concepts continue in Zipf's direction : minimizing the communication effort drives the speaker to shorten frequent words (Zipf, 1949). The length metrics defined by Zipf is not the number of letters, but the number of syllables or the number of phonemes of the written form (Zipf, 1935); the metrics of our method is also the number of syllables, or more precisely the number of vowel nuclei, computable from the written form; this metrics takes its root into the oral origin of the chunk. The word frequency is measured in the segmented text.

This method of segmentation into chunks is based on digital properties, and is valid on languages with alphabetic script. It is endogenous, as it computes on the text to be segmented and does not use any resource external to the parsed text.

1 Structure model of the chunk according to Abney and according to Déjean

The concept of chunk has been proposed by Steve Abney (1991). It has been based on properties of speech : Abney defined the chunk as a stress group. As speech is constrained by the vocal system, we can see the chunk as a generic concept on natural languages, a concept of language. Hervé Déjean (1998) has proposed a structure model for the chunk : beginnings and ends of chunk (words or morphemes) around a kernel (Déjean, 1998, page 117); our method uses this model.

For instance, the written form "Commission" has been found in the following chunks in the same text :

[Commission européenne]
[la Commission]
[la Commission européenne]
[dans la Commission]

And here is the synthesis :

[dans [la [Commission] européenne]
[beginnings [kernel] ends]

Proceedings of the 11th International Conference on Parsing Technologies (IWPT), pages 202–205,
Paris, October 2009.

2 Local deductions and their generalization at text level

Properties of the chunk are used locally at occurrence level : an occurrence of a written form is locally a beginning or an end of a chunk. An important question is to decide how to articulate local deductions at occurrence level and their global merging at text level.

We know that occurrences of the same written form may be occurrences of more than one word, in different contexts. For instance, "*on*" in English is the beginning of a chunk in "*on the contrary*", but it is the end of a chunk in "*it is going on*". These two occurrences correspond to two different words, which have different positions and different contexts, and their local deductions cannot be merged. So, we can merge local deductions for occurrences of the same word. In practice, we merge local deductions for occurrences of a written form if there is no beginning - end contradiction.

We tried full merging, as if all occurrences were of the same word. This solution remains valid for monofocused short texts (some thousands words). But, to be able to chunk longer texts, we have chosen now the solution of a partial generalization (see below in 4).

3 Two properties of a chunk

The algorithm exploits two properties of the chunk.

3.1 Property 1 : the chunk is a constituent of the virgulot

Hervé Déjean (1998) has defined the "entreponctuations" as a constituent delimited by two punctuations. Nadine Lucas (Lucas, 2001) has proposed the term "virgulot", that we will use now. We define the following constituent hierarchy : the text is constituted of virgulots, themselves constituted of chunks, themselves constituted of occurrences of written forms.

Property exploited by the algorithm :
- a written form attested at the beginning of a virgulot is a beginning of a chunk,
- a written form attested at the end of a virgulot is an end of a chunk.

Here are some instances of virgulots :

*, **in** denen Aale **leben** ,*
*, **bis** die Bewirtschaftungspläne **vorliegen** .*

*. **It** also intends to explore **measures** ,*
*, **before** migrating upstream to spend most of their **lives** .*

*, **en** las aguas centro-occidentales del Océano **Atlántico** .*
*, **donde** transcurre la mayor parte de su **vida** .*

*. **Lasciandosi** trasportare dalla corrente e **nuotando** ,*
*, **dove** si riproducono una sola volta e poi **muoiono** .*

First written forms of virgulots are beginnings of chunks (prepositions, pronouns, …), and their last written forms are ends of chunks (nouns, verbs, adjectives, …).

3.2 Property 2 : the chunk is the period of the correlated functions *length* and *frequency* of words on the syntagmatic axis

We define two integer functions of words on the syntagmatic axis (inside a virgulot) : their length, defined as their number of syllables, and their frequency in the text to be segmented.

Here is an instance of a virgulot :

, would migrate from the rivers on their territories ,

length:	1	3	1	1	2	1	1	4
frequ.:	10	3	6	65	2	6	4	1

On the length function, we have the following non-decreasing sequences : [1 3] [1 1 2] [1 1 4].

On the frequency function, we have the following non-increasing sequences : [10 3] [6] [65 2] [6 4 1].

For these two functions, a period corresponds to a sequence; in other words, these sequences give a way to segment; these 2 functions are synchronous : sequences of both functions (nearly) define the same periods; on a (synchronous) period, both functions are in phase opposition : on a period (which defines a chunk), the length function is non-decreasing, and the frequency function is non-increasing; the common properties of these two functions allow us to call them correlated; it is an other way to say that short words are frequent and that long words are rare.

We notice, following Zipf (1949) in "Human Behavior and the Principle of Least-Effort" that writing and speech are an optimal compression; it reminds the principles of file compression in computer science : frequent data are short coded, and rare data are long coded. Let us make an observation on the Zipf law, as it is known today : this law makes a relation between frequency and rows of words sorted by decreasing frequency; if we knew only this law, we would forget length of words; but Zipf proposed to consider length and frequency together, in a correlated way, as an optimization (the Least-Effort). As we use length and frequency together, in a correlated way, we go back to the origin of Zipf's concepts.

To compute word length from the written form, length is defined as the number of syllables, i.e. the number of vowel nuclei (a sequence of contiguous vowels corresponds to a vowel nucleus, and to a length equal to 1). This calculation needs as input the vowels of the alphabet (Latin or Greek). There is a particular case : is the *y* vowel or consonant. The *y* is vowel in "*system*" (length 2) and consonant in "*rayon*" (length 2); *y* is consonant by default; *y* is vowel at the beginning or the end of a word, or alone (*usually, by, y*); *y* is vowel between 2 consonants; these rules are enough to process all cases for the 20 natural languages of the corpus. Acronyms (sequences of uppercases) have a length equal to twice their numbers of letters (tendency to be in the end of chunk). A number (sequence of figures) has a length equal to 1, whatever its number of figures (tendency to be in the beginning of chunk).

4 An algorithm based on these properties

The frequency and the length of every written form are computed.

For the property 1, based on the virgulot, the text is processed, and occurrences of written forms at the beginning or end of virgulot are noted as beginning or end of chunk.

For the property 2, based on monotonous sequences, the text is processed, while noting borders between 2 monotonous sequences, that gives for each border an end and a beginning of chunk. A Boolean function "in the same sequence" returns whether 2 contiguous words are in the same monotonous sequence (i.e. in the same chunk). Four solutions are experimented : on length only, on frequency only, on length AND frequency (then shorter chunks), or on length OR frequency (then longer chunks). Results are very comparable, because both functions are strongly correlated[1]. For example, this function, in "length OR frequency" mode, on words i and i+1, to express the fact that these two words are in the same sequence, has the following form :

words i and i+1 are not separators of virgulot
AND
(length(i+1) ≥ length(i) OR
frequency(i+1) ≤ frequency(i))

[1] Using length alone allows, not using frequency, to get a method usable on a very short text, as a search engine query.

The generalization of local deductions is done the following way : for all occurrences of a written form, a synthesis of local deductions is done. There are 8 cases : 2 properties, 4 cases for each (2 Booleans : beginning, end). If all local deductions are compatible, they are merged, i.e. occurrences without any local deduction take the tag of occurrences with the same local deduction : either beginning or end of chunk.

Here is the trace of the process on our instance of virgulot :

virgul. b	virgul. e	sequ. b	sequ. e	general. b	general. e	result b	result e	len.	freq.	
[1,0]	[1,0]	[0,0]	[2,0]	1	10	*would*
[0,0]	[0,1]	[0,1]	[0,2]	3	3	*migrate*
[0,0]	[1,0]	[1,0]	[2,0]	1	6	*from*
[0,0]	[0,0]	[1,0]	[1,0]	1	65	*the*
[0,0]	[0,1]	[0,1]	[0,2]	2	2	*rivers*
[0,0]	[1,0]	[1,0]	[2,0]	1	6	*on*
[0,0]	[0,0]	[1,0]	[1,0]	1	4	*their*
[0,1]	[0,1]	[0,0]	[0,2]	4	1	*territories*

From the first property (the first column of Booleans), *would* is the beginning, and *territories* is the end of the virgulot, therefore beginning and end of a chunk ([marks a beginning of chunk,] marks an end of chunk) :

, [would migrate from the rivers on their territories],

The second property (the second column of Booleans) which exploits the monotonous sequences, here in "length OR frequency" mode, gives the following chunking :

, [would migrate] [from the rivers]
[on their territories] ,

The generalization of local deductions (the third column of Booleans) adds the fact that *the* and *their* are beginnings of a chunk elsewhere in this text.

Then these three sources of deduction are merged, and we obtain the following segmentation (the forth column) :

, [would migrate] [from [the rivers]
[on [their territories] ,

5 Some sentences segmented into chunks

The validation corpus of the method is composed of 12 press releases (about 1000 words each for one language), every release is written into 6 to 20 languages, and of the part 1 of the "Treaty establishing a Constitution for Europe" in 11 languages (about 10 000 words for one language), from the website of the European Union (http://europa.eu/).

The following sentences are extracted from the release IP/05/1018 of 2005 (and processed in "length OR frequency" mode) :

[Die Laichgründe] [der Aale] befinden] [sich [im Sargassosee] [im mittleren Westatlantik] .

[Eels spawn] [in [the Sargasso Sea [in [the western central Atlantic] Ocean] .

[Las anguilas] desovan] [en [el Mar [de [los Sargazos] , [en [las aguas] centro-occidentales] [del Océano Atlántico] .

[La zone [de frai] [de l'anguille] [se situe [en mer] [des Sargasses] , [dans [la partie centre-ouest] [de l'océan Atlantique] .

[Le anguille] [si riproducono] [nel mar [dei Sargassi] , [nell'Atlantico centro-occidentale] .

The following sentences are extracted from the part 1 of the "Treaty establishing a Constitution for Europe" (and processed in "length OR frequency" mode) :

[Die Union] steht allen europäischen] Staaten offen] , [die [ihre Werte] achten] [und [sich verpflichten] , [sie gemeinsam] [zu fördern] .

[The Union] [shall be open] [to [all [European States] [which respect] [its values] [and [are committed] [to promoting] them] together] .

[La Unión] [está abierta] [a todos] [los Estados] europeos] [que respeten] [sus valores] [y [se comprometan] [a promoverlos] [en común] .

[L'Union [est ouverte] [à [tous [les États] européens] [qui respectent] [ses valeurs] [et [qui s'engagent] [à [les promouvoir] [en commun] .

[L'Unione [è aperta] [a tutti] [gli Stati europei] [che rispettano] [i suoi valori] [e [si impegnano] [a promuoverli congiuntamente] .

Conclusion

While characterizing the chunk in a purely digital way, from properties of length et frequency functions of words on the syntagmatic axis, this original method consists in calculations on the text to segment; it has the advantage to be applied to a great number of languages, with the same algorithm, without any monolingual resource : languages with alphabetic script, with a written word which separates function words from content words (it is not the case in Finnish), and compatible with a structure model of the chunk where function words generally are before content words; the method is promising for the 22 languages of the European Community[2].

This method can be applied in automatic indexing, for search-engines (as Exalead does, to be able to output the most frequent terms associated to the documents of the answer), and in subsentential alignment, to constraint the statistical alignment (as in Similis, the alignment software of Lingua et Machina, but this software uses monolingual resources for every language). The interesting feature of this method is not to need any resource for a new language to process[3].

As it is independent from specificities of each language, this method is not "multilanguage", neither "multi-monolanguage", but as it exploits generic properties of natural languages, that is properties of language, as an abstraction of natural languages, we could perhaps simply call it a "linguistic" method.

References

Steven Abney. 1991. *Parsing By Chunks*. in Principle-Based Parsing, 257-278, Kluwer Academic Publishers.

Hervé Déjean. 1998. *Concepts et algorithmes pour la découverte des structures formelles des langues*. Thèse de doctorat de l'université de Caen, France.

Nadine Lucas. 2001. *Étude et modélisation de l'explication dans les textes*. Actes du Colloque "L'explication: enjeux cognitifs et communicationnels", Paris.

Gérald Vannier, Anne Lacheret-Dujour, Jacques Vergne. 1999. *Pauses location and duration calculated with syntactic dependencies and textual considerations for t.t.s. system*. ICPhS 1999, San Francisco, USA, August 1999.

Jacques Vergne. 2000. Tutorial : *Trends in Robust Parsing*. Coling 2000.

George K. Zipf. 1935. *The psychobiology of language : An introduction to dynamic philology*. Boston, Mass., Houghton-Mifflin.

George K. Zipf. 1949. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley.

[2] See results on : http://www.info.unicaen.fr/~jvergne/chunking_multilingue_endogene/

[3] But a problem for this large scale multilingual method is to evaluate the results on so many languages : we need a speaker for every language. For the moment, it is done for German, English, Spanish, French and Italian.