Signalling well-written academic articles in an English corpus by text mining techniques

Nadine Lucas*— Bruno Crémilleux*— Leny Turmel **

GREYC, CNRS UMR 6072,Caen University
Campus Côte de Nacre
F- 14032 Caen Cedex France
Tel 33 (0)2 31 56 73 36   fax 33 (0)2 31 56 73 30
*{bruno, nadine}@info.unicaen.fr
** lturmel@etu.info.unicaen.fr

KEYWORDS: text mining, textual linguistics, corpus characterisation, rules characterising classes, academic style, detection of English mistakes

1. Introduction

In the last decades, most of the information supports have been transformed into electronic form. The value of these data has been fully recognised only recently: their potential use may greatly exceed the uses for which the data were actually collected and stored. Data mining and text mining address the interactive and iterative knowledge discovery processes from large databases and collections of texts. This research domain has been quite active in the last decade due to the pressure of corporate data owners who need added value from the huge volume of data they easily collect.

We meet such a challenge in classifying academic articles in English prior to copy editing. We have to classify a large corpus to discriminate well-written versus poorly-written academic articles in an English corpus. The aim is to inform computer-aided copy editing by focussing attention on poorly-written articles. We also wish to skip correct (i. e. with no mistake) segments and highlight incorrect segments in these articles in order to save time. The training corpus includes forty academic articles written by authors who use English as first or second language with varying degrees of competence. We use pairs of articles allowing comparison between two successive versions, before and after human copy editing.

The task is challenging, because correctness in style is difficult to define. This corpus of great interest requires an original text-mining method discovering rules from the linguistic corpus on correct and incorrect segments. The main feature presented here is the use of several different text segments, or textual measures, where stylistic mistakes or improper use can be detected. Each text is thus divided into parts, into paragraphs and into sentences, in order to circumscribe textual semantic units.

We claim that results stemmed from cross-fertilisation between novel text-mining and textual linguistics approaches. The level-specification of forms and the top-down positional inheritance (see Section 3) are essential to extract associations concluding on correct or incorrect segments. It would be hopeless to rely on words alone. Furthermore, owing to the large number of segments, efficient data mining tools are required. We use MVMiner, a prototype that is able to extract all associations in a data set above a frequency threshold. Associations are given in a special condensed representation (see Section 4) allowing relevant rules characterising classes to be exhibited.

The paper is organised as follows. First we present the problem and second introduce the linguistic approach, based on position and form. The main notions of level-specification and positional inheritance are explained. Third, the text mining method is presented as well as the techniques to extract associations and rules characterising classes. Broadly speaking, the text mining method captures reliable similarities and rules embedded in the articles. Then we examine results. Experiments show that extracted rules highlight features that do characterise classes. Last, we discuss the usefulness of this method.

2. Problem description

First, let us describe the task. The ultimate objective is to ensure "readability" or "better understanding" of academic articles. Corrections have to do with style mainly and sometimes with grammar, they have nothing to do with word-spelling. Thus words cannot be used as descriptors. Two classes are established, well-written and poorly-written articles. Descriptors characterising one of the classes (well-written articles) are established: they are designed for characterising correctness in academic articles in English. The aim is to extract association rules from the pairs of texts, and then to apply them to discriminate well-written from poorly-written texts in a new un-revised collection of academic articles. New texts can then be labelled automatically.

2.1. Segments

Current approaches in text mining equate descriptors with words. But since most corrective changes concern word position rather than words themselves, checking words would be hopeless. Comparison of the pairs of text would return no significant results. To overcome this difficulty, we established a hierarchy of textual measures that correspond to selection as described by Hockett (1958) or Harris (1952) and involve the linguistic notion of span for textual markers. We emphasise the relative positions of textual markers inside these observation windows. Thus, instead of handling simple units (text and words), we now consider correct and incorrect segments.

2.2 Main features

Our approach may be characterised by two concepts, called top-down positional inheritance and level-specification for forms. Level-specification for forms is devised to represent span. Top-down inheritance amounts to keep information obtained at the higher levels of hierarchy down to the smallest measures of text. This amounts to memorising information on the tree-like hierarchy provided by the layout, since any word can be related to its including comma-punctuated unit. A comma-unit found, say, in an introduction, is related to the sentence where it is located, to the paragraph including it and to the part including this paragraph (see Figure1).

3. Linguistic contextual approach

Descriptors are textual markers that come from linguistic knowledge on discourse structure. They fall in two main categories, forms and positions. Forms may be associated with words as a first approach (Péry-Woodley 1993). Words are here considered as small segments in the chain of characters actually handled by computers. Words are not informative by themselves; they belong to a context.

3. 1 Textual measures

Our postulate is that the context can be approached through textual measures, as expressed by the material layout of the document. Texts are automatically divided into text-zones, and the text-body is subdivided in parts, sections, paragraphs, sentences and comma-punctuated units. When checking modifications in the original and revised versions of articles, these units are compared at each level of the tree-like hierarchy provided by the layout. For the two examples produced below, the clause span will be associated with a text chain delimited by comma, while text segments characterised by contrasted use of voice will be best approximated by sections or parts. These text measures are observation windows. They play the same role as selection, the difference being that we consider *all* available segments in the text hierarchy.

Two examples can be used to illustrate this point. One common mistake is the misplacement of adverbs; this is checked within the span of the clause and inside the verb group. It concerns word position in a narrow context, the verb phrase, itself belonging to a wider context, which is a period or comma-delimited unit roughly encompassing a clause.

Example 1 a (original text)

> They concur to identify an area, in a certain way homogeneous, where, besides the beauty of places, the Cultural Heritage represents an extraordinary richness worldwide *recognized*, which animates a strong need of safeguard and valorisation.

Example 1 b (revised text)

> They concur to identify an area, in a certain way homogeneous, where, besides the beauty of places, the cultural heritage represents an extraordinary richness recognized worldwide, which animates a strong need of safeguard and valorisation.

In this example, note the last position in the fifth comma-punctuated unit is filled with the form *recognized* in the original text and with the form *worldwide* in the revised text.

Words are obviously not the only type of segments useful for the current study. As will be explained below, most stylistic mistakes involve lack of coherence in paragraphs and parts. Example 2 shows an incorrect use of infinitive mode. It appears in the last but one sentence, in a context where indicative mode and passive voice are required. The narrow context is the paragraph, and the larger one is the part (given only for a) below). Letting aside the confusion between infinitive and imperative, we would say that parts may be marked by mode, but not at the particular position filled by this paragraph. Here, sentences have been shrunk and each is on a line.

Example 2 a (original text)

§1 Cancer risk in mutation carriers has been evaluated by a …by 60 years of age.
The age-specific penetrance was higher in women than in men …of breast cancer [14].
In another hospital-based series…, the risk was estimated to be …at the age of 45.
Lifetime risk was estimated to be 73% in males…, …by breast cancer [30].
To keep in mind that individuals with LFS …, …in childhood [80].
It could be firmly stated that penetrance is high, but incomplete particularly in males.

§2 However, cancer incidence …did not seem… of ADCC in childhood [34; 81].
Paradoxically, …with a wide spectrum of mutations [27; 29; 35; 56; 63; 66; 82].
§3 Mutations outside the core …could be associated…in the core DNA-binding domain [73].
Some cases … have been described…, they could be not so rare [30; 83; 84].
This is not unexpected …of mutation carriers.
§4 One case of mosaïcism was reported [85], it must be kept in mind for genetic counseling.

Example 2 b (revised text)

Cancer risk in mutation carriers has been evaluated by a …by 60 years of age.
The age-specific penetrance was higher in women than in men …of breast cancer [14].
In another hospital-based series…, the risk was estimated to be …at the age of 45.
Lifetime risk was estimated to be 73% in males…, …by breast cancer [30].
It should be kept in mind that individuals with LFS …, …in childhood [80].
It could be firmly stated that penetrance is high, but incomplete particularly in males.

Representing span is not obvious. The next sections detail the main concepts of text observation windows or textual measures and their descriptors.

3.2 Descriptors

3.2.1 Forms

Knowledge on the academic genre as such was acquired prior to this experiment (Lucas 1991) and special attention was given to academic English focussing on English as second language (Grabe 1987, Péry-Woodley 1989). Specific words or features are detected, as having functional value (Péry-Woodley 1993). Such forms are discourse connectors in a very broad meaning. Our postulate was to use different descriptors for each measure of text, according to its level, in an effort to represent span. Special effort was made to select textual markers for higher levels. This is called level-specification of descriptors. Note that a full syntactic analysis could not be run prior to our experiment, because tags would interfere with text-mining. Moreover, classic syntactic parsing does not provide the kind of information we needed. Hence, only marks that could be found by a very shallow parsing through regular expressions were selected.

Characteristic word endings such as *-ly* or words of few letters are used as descriptors for comma-punctuated units and sentences, pairs of words for paragraphs, groups of three words are used to describe sections or parts. Whenever possible, words or contiguous words are used, for instance co-ordination marks *and, as well*, and *as well as* are descriptors respectively for sentence, section, and part level. Features such as passive voice were checked at the pr-processing stage through very rough criteria including *be* auxiliary and *-ed*. Such features type paragraphs, sections or parts when they are present in all sentences in the considered measure. However, gaps inside the measure are allowed, that is to say that beginning and end are considered remarkable positions as explained below. Special punctuation marks are also used as descriptors. All these textual markers were selected on the basis of a careful examination of the corpus.

In the same way, different forms belonging to the same grammatical or rhetorical category were grouped into one class as descriptor for higher levels, for instance *the*, *this, these, that, those,* which stand for themselves at sentence level, were merged as anaphoric determiners at paragraph level. Some examples are given and classes are shown between {} in Table 1. Classes were fused or rearranged as a single class as descriptor for part level, for instance anaphoric determiners and anaphoric adverbs were merged in an anaphoric general class. These level-specific descriptors are also matched with first and last positions clues.

3.2.2 Positions

Specific position in a given measure is a descriptor, independent from forms. Special positions refer to beginning or end of any measure of text, e.g. beginning of a sentence or end of a paragraph. The beginning of a comma-punctuated unit is the first word, the beginning of a sentence is the first comma-punctuated unit, if any, otherwise the first word. Similarly the beginning of a paragraph is the first sentence, the beginning of a section is the first paragraph, and the beginning of a part is the first section, if any, otherwise, the first paragraph, while the beginning of the text-body is the first part. The same principle applies for end.

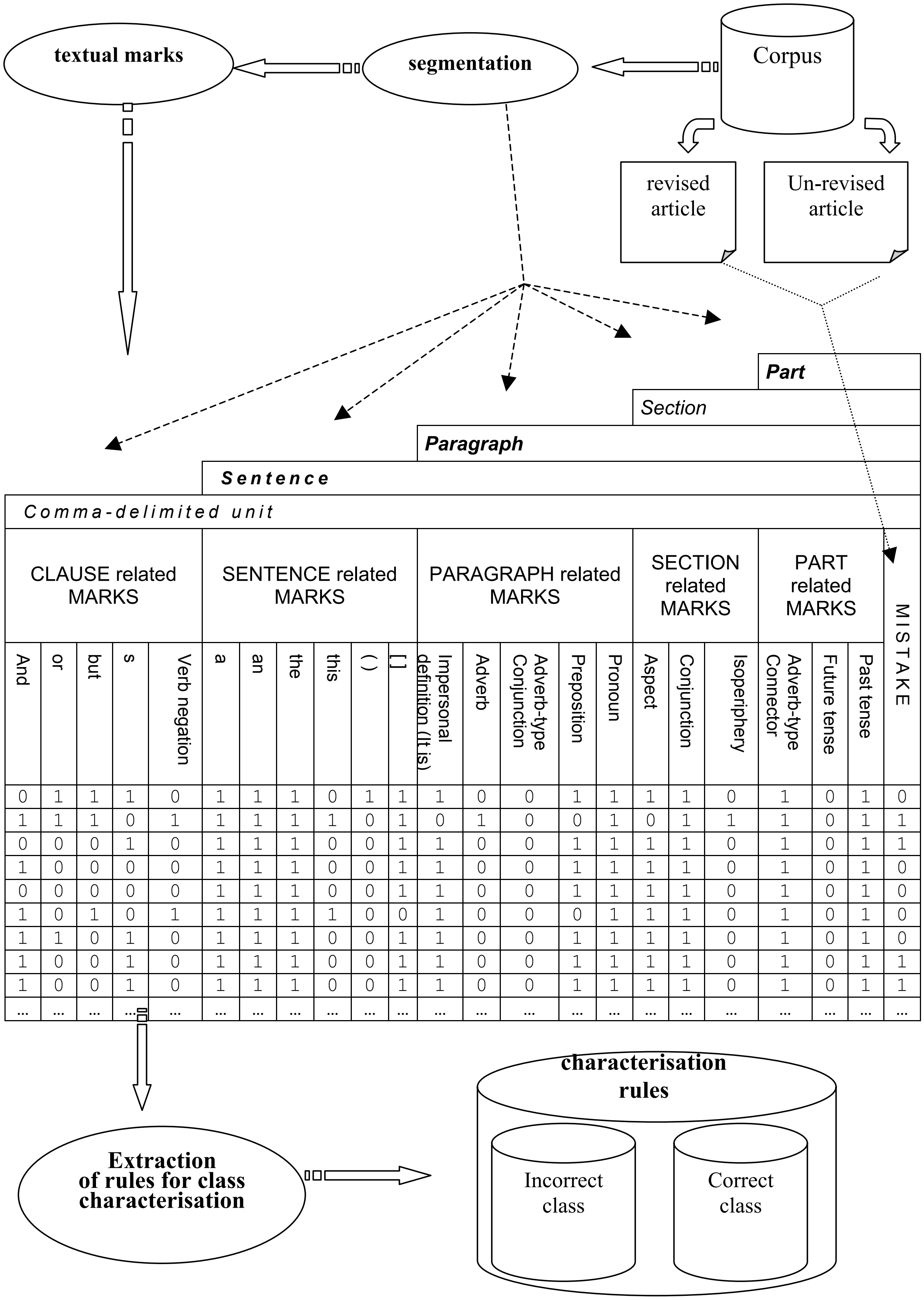
Table 1. *Examples of descriptors*

Measures	Descriptors morphological features
Comma-units	Word endings (-ed, -ing, -ly) Connectors 1 (Despite, Indeed, Because) Co-ordination 1 (and) Adjective anaphoric (Its, Their, Such…) Typography ()[]""
sentences	*Same as above*
paragraphs	Connectors 2 (There is, In fact…) Co-ordination 2 As well, in addition Adverb {-ly words, + despite, indeed…} Preposition Pronoun Determiner definite {this, these, those} Determiner indefinite {one, some…} Subordination Conjunctions …
sections	Ordinal Cardinal Voice Aspect Major Conjunctions {If, although, when, because, …}
parts	Connectors 3 (in spite of, for this reason…) Co-ordination 3 as well as Tense Future, Past Aspect Perfect Voice Passive Adverbial Connectors {however, nonetheless…} Anaphoric words {this, these…thus} Wh conjunction {which, what, when, why…} Other conjunctions {if, although, because…}

Thus, we carefully describe where textual marks may occur, instead of using just words as in classical text mining approaches. Examples of remarkable co-occurrence of descriptors, as related to specific position and specific forms, is for instance the high number of so-called anaphoric pronouns at the beginning of sentences themselves located at the end of paragraphs.

A special position descriptor describes repetition of the same textual mark at the beginning *and* end of a text measure. It is called "isoperiphery" and signals a repetition e.g. at the beginning and end of a paragraph there is a sentence bearing the same feature, for example, a "definition" sentence containing the word *is*. We also call this *parallelism of forms*.

Figure 1. Overview of the system



CLAUSE related MARKS					SENTENCE related MARKS						PARAGRAPH related MARKS					SECTION related MARKS			PART related MARKS			MISTAKE
And	or	but	s	Verb negation	a	an	the	this	( )	[ ]	Impersonal definition (It is)	Adverb	Adverb-type Conjunction	Preposition	Pronoun	Aspect	Conjunction	Isoperiphery	Adverb-type Connector	Future tense	Past tense	
0	1	1	1	0	1	1	1	0	1	1	1	0	0	1	1	1	1	0	1	0	1	0
1	1	1	0	1	1	1	1	1	0	1	0	1	0	0	1	0	1	1	1	0	1	1
0	0	0	1	0	1	1	1	0	0	1	1	0	0	1	1	1	1	0	1	0	1	1
1	0	0	0	0	1	1	1	0	0	1	1	0	0	1	1	1	1	0	1	0	1	0
0	0	0	0	0	1	1	1	0	0	1	1	0	0	1	1	1	1	0	1	0	1	0
1	0	1	0	1	1	1	1	1	0	0	1	0	0	0	1	1	1	0	1	0	1	0
1	1	0	1	0	1	1	1	0	0	1	1	0	0	1	1	1	1	0	1	0	1	0
1	0	0	1	0	1	1	1	0	0	1	1	0	0	1	1	1	1	0	1	0	1	1
1	0	0	1	0	1	1	1	0	0	1	1	0	0	1	1	1	1	0	1	0	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

4. Associations characterising English correctness

Broadly speaking, the text mining method captures reliable similarities and rules embedded in the articles. In our experiments, articles are translated into patterns based on the textual markers. We get between 728 and 22,041 patterns according to the window covered by the textual markers. Rules are built on these patterns and are evaluated by the classical measures of frequency and confidence. As the task is the discovery of all rules, the frequency and confidence of which are user-specified thresholds, there are a huge number of candidate rules and efficient algorithms are required. We use recent results on condensed representations based on δ-free patterns to mine δ-strong rules characterising classes.

4.1 Data mining and text mining

There are numerous text and data mining techniques that have often been developed on core techniques from statistics or machine learning. Data mining and text mining share common methods and algorithms (e.g., the search of frequent associations of patterns with a level-wise approach) (Mannila and Toivonen, 1997). Text mining has been introduced in order to offer tools for utilising text resources in data mining driven decision support. Text mining has some specificity: documents can be more or less structured, the sequence of words has to be taken into account, the pre-processing stage has a great role and it is recognised that linguistic and natural language resources are required. Many text mining methods include text operations such as pre-lexical analysis (e.g., treatment of digits, hyphens, punctuation marks), lexical analysis such as elimination of stop-words, stemming and morphology, selection of index terms, syntactic analysis (e.g., determination of noun phrases), semantic constructions including linguistic aspects as well as domain-specific components. Sequential patterns have been used for discovering trends among patents (Lent, Agrawal. and Srikant, 1997) and efficient algorithms to find co-occurring of text phrases have been developed (Ahonen-Myka *et al.* 1999). One popular data mining technique concerns knowledge discovery from frequent association rules. This kind of process has been thoroughly studied since the definition of the mining task in (Agrawal, Imielinski, Swami, 1993). Association rules can tell something like "It is frequent that when descriptors A1 and A2 are true within an example, then descriptor A3 tends to be true" where A1, A2 and A3 are for instance descriptors of texts (Feldman, abd Dagan and Klösgen, 1996). In the following sections, we define association rules more precisely and δ-strong rules characterising classes which are the simplest rules characterising classes with respect to their left-hand sides.

4.2. Association Rules

Let us provide a simple formalisation of association rules mining task. In the following, definitions are relating to a set of *examples* (e.g. paragraphs, sentences, comma-punctuated units).

[item, itemset]

Let *D* = {D1,…, Dn} a set of descriptors. An element of *D* is called *item* and a subset of *D* *itemset*.

For instance, an item corresponds to the presence of mark *-ly* at the end of a comma-punctuated unit. From a technical point of view, algorithms require binary items to be efficient but a set of initial non-binary descriptors can always be translated into a set of items. Hence, in our experiments, the whole set of descriptors (see Table 1 for a subset) corresponds to 270 items.

[association rule]

An association rule is an expression X => B, where X belongs to *D* (i.e. X is an itemset) and B ∈ *D*\ X.

For instance, *parenthesis the an that* => incorrect is an association rule meaning that when the items *parenthesis the an that* are true (present), then the segment is incorrect. X is also called the left-hand side of the rule. The fact that B must not belong to X is only to avoid producing trivial rules. Note that it is easy to generalise this definition to allow rules with several items in their right-hand side (i.e. conclusion) (Agrawal, Imielinski, Swami, 93). The classical measures of *frequency* and *confidence* capture the semantics of the "representativity" and the "strength" of the rule (Agrawal, Imielinski, Swami, 93).

[frequency, confidence]

Given X belonging to *D*, F (X) (or frequency of X) is the number of examples for which each item of X is true. The frequency of X=>B is defined as F (X ∪ B) and its *confidence* is F (X ∪ {B})/ F (X). Frequency is also called support.

The standard association rule mining task concerns the discovery of *all* rules the frequency and confidence of which are greater than a user-specified threshold. In other words, one wants rules that are frequent "enough" and valid. The main algorithmic issue concerns the computation of every frequent itemset.

[frequent itemset]

Let γ be a frequency threshold lower or equal to the number of examples. An itemset X is said frequent if F (X) < γ.

The complexity of frequent itemset mining is exponential with the number of descriptors. Many research works (e.g. Pasquier *et al.* 1999, Boulicaut Bykowski Rigotti 2000) concern the contexts for which such a discovery remains tractable, even though a trade-off is needed with the exact knowledge of the frequencies and/or the completeness of the extractions.

From a technical point of view, the search of association rules can be divided in two parts: first, the extraction of all frequent itemsets and then the generation of rules. Itemsets correspond to associations between descriptors. In this work, we use the discovery of frequent itemsets to validate hypothesis set out in Section 3. Furthermore, we use a special kind of association rules (δ-strong rules characterising classes) that we have developed (Crémilleux and Boulicaut 2002) to infer rules about the presence or not of English mistakes in segments.

4.3 δ-strong rules characterising classes

Let us consider a classification task with k class values. Assuming C1, … ,Ck are the k items that denote class values. Here, we have k = 2 (two classes, correct and incorrect). A δ-strong rule characterising classes concludes on a class value with a rather high confidence.

[δ-strong rule characterising classes]

A *δ-strong rule characterising classes* is an association rule with a minimal left-hand side that allows at most ∂ exceptions and that concludes on one class value (i.e., Ci).

In other words, the confidence of a δ-strong rule characterising classes is at least equal to 1−(δ/ γ). It is out the scope of this paper to explain precisely how such rules are built (details are in Crémilleux and Boulicaut 2002). Let us simply say that we use recent results on condensed representations based on δ-free patterns (Boulicaut Bykowski Rigotti 2000, 2003). δ-strong rules are built from δ-free patterns that constitute their left-hand sides (Boulicaut, Bykowski and Rigotti, 2000). The property of freeness checked by δ-free patterns enables to give a safe pruning criterion in the search of frequent itemsets. It means that we are able to design effective algorithm even in the case of huge, dense and/or highly correlated learning data sets where usual approaches do not fit (Crémilleux and Boulicaut 2002). Furthermore, a δ-free pattern is a minimal conjunction of items to know the frequencies of a set of items. It means that we are able to extract the simplest rules (i.e. with the minimal left-hand sides) to conclude on an item, the uncertainty being controlled by δ. We argue that this property of minimal left-hand side is a fundamental issue for class characterisation. Not only it prevents from over-fitting (i.e. over-specified rules acquired on a learning set and leading to miss-classified new examples) but also it makes the characterisation of an example easier to explain. It provides a feedback on the application domain expertise that can be reused for further analysis. Moreover, if the extraction of patterns is done under a sensible sufficient condition, important classification conflicts (identical body conflicts, included bodies conflicts) are avoided, which is useful to classify new examples (Crémilleux and Boulicaut 2002).

Experiments are performed with our prototype MVMiner, which has been implemented by F. Rioult at GREYC laboratory. Given δ and γ, MVMiner extracts all δ-free patterns and rules characterising classes. The following section shows examples of such rules.

5. Results

Experiments show that extracted rules highlight features that do characterise classes (correct and incorrect segments). These features enable to classify new segments.

First let us note that the use of a top-down inherited context allows extracting rules even from a relatively small number of articles because it enables to define a sufficiently large number of patterns. This could not be obtained from data ignoring context, as is shown in Table 2. The last line shows a comparison with an experiment run without top-down inheritance. It is clear that context is essential to extract frequent itemsets.

Second, in order to test the appropriateness of the rules to qualify well-written versus poorly-written articles, we ran an experiment in which articles were tested irrespective of their primitive labelling (before and after copy-editing), on the whole corpus. Results showed that we succeeded in classifying segments in articles belonging to the revised corpus as correct.

Table 2. *Frequent itemsets extracted with and without context*

	Comma-units	Sentences	Paragraphs	Parts
Nb. of measures	22041	10643	3520	728
Frequency	60	80	50	50
δ	10	10	0	0
Number of frequents itemsets with top-down inherited context	48507	815	953	29
Nb. of frequents itemsets without top-down inherited context	5	8	14	0

δ-strong rules characterising classes provide a list of correlated textual markers, which are considered as being representative of a well-written text. This validates the established model for automatic editing and helps to define sound contexts for triggering correction rules. This is especially true to capture large contexts through large span patterns. It should be noted that disappointedly few rules were extracted from high level segments, none for parts and fewer than expected for paragraphs.

Positional inheritance balanced these results. It is particularly useful, because it is linked with high order coherence of text. Rules extracted on comma-units but exhibiting a large span of inherited features do characterise well-written articles, as shown in Table 3. The notation is somewhat difficult to read and needs to be translated, for example the first line reads "if a comma-unit belongs to a (complex) sentence including *where* and belongs to a part including a complex co-ordination syntagm *as well as*, then it probably is a correct segment". Here we can see that level-specification for marks also plays a role in characterising correctness. Example 3 illustrates such a case (see last sentences of the paragraphs for the cited marks).

Example 3a (original)

3.2 The MutL-related complexes

The existence of four MutL homologs, MLH1-3 and PMS1 have been reported in yeast (Table 2) [72, 73]. In human, in addition to the four MutL homologs, namely hMLH1, hMLH3, hPMS1 and hPMS2, a cluster of hPMS2-like genes have been localized on chromosome 7 [74, 75]. […] Because MLH1 is a common subunit to all three complexes, its deficiency leads to a severe phenotype, similar to that of MSH2-deficient cells. The MSI phenotype resulting from MLH1-deficiency is characterized by a tremendous increase in base-base mismatches, as well as frameshift mutations resulting from unrepaired IDL.

3.3 The discrimination of the newly-synthesized strand and the processing steps of MMR

Once bound to mismatches, MutLα complex is able to interact with numerous factors, consistent with the assembly of a higher-order complex that is involved in the excision of a large fragment of the newly-synthesized DNA strand containing the mismatch (rev. in [8, 11, 85]). In organisms other than *E. coli*, the signal that allows to discriminate the newly-synthesized strand from the template is still doubtful, but does not involve DNA methylation [86]. The proliferating cell nuclear antigen (PCNA) which is essential for DNA replication, where it acts as a processivity factor has also been implicated in MMR before and during the DNA synthesis step and could be involved in this process. […]

Example 3b (revised)

3.2. The MutL-related complexes

The existence of four MutL homologs, MLH1–MLH3 and PMS1 has been reported in yeast (Table 2) [72,73]. In human, in addition to the four MutL homologs, namely hMLH1, hMLH3, hPMS1 and hPMS2, a cluster of hPMS2-like genes have been localized on chromosome 7 [74,75]. […] Because MLH1 is a common subunit to all three complexes, its deficiency leads to a severe phenotype, similar to that of MSH2-deficient cells. The MSI phenotype resulting from MLH1 deficiency is characterized by a tremendous increase in base–base mismatches, as well as frameshift mutations resulting from unrepaired IDL.

3.3. The discrimination of the newly synthesized strand and the processing steps of MMR

Once bound to mismatches, the MutLα complex is able to interact with numerous factors, consistent with the assembly of a higher-order complex that is involved in the excision of a large fragment of the newly synthesized DNA strand containing the mismatch (reviewed in [8,11,85]). In organisms other than *E. coli*, the signal that allows us to discriminate the newly synthesized strand from the template is still doubtful, but does not involve DNA methylation [86]. The proliferating cell nuclear antigen (PCNA) which is essential for DNA replication,

> where it acts as a processivity factor has also been implicated in MMR before and during the DNA synthesis step and could be involved in this process. […]

There are a fair number of rules involving repetition or parallelism, at the paragraph level: the label "isoperiph" in Table 3 signals a paragraph beginning and ending with sentences bearing the same textual marker. This stylistic feature appears with a fairly high confidence. Line two reads "if a comma-unit belongs to a sentence showing the presence of *by*, and belongs to a paragraph that can be characterised by parallelism, namely "definition" sentences containing the mark *is* at the beginning and end, and also belongs to a part well-characterised by special features, i.e. by the passive voice and the presence of impersonal pronouns, then there is a good chance that it is a correct segment". All these marks indeed describe coherent "passive" parts, and these parts occur in well-written articles. In the same way, line 3 reads "if a comma-unit belongs to a sentence marked by the special punctuation semicolon, and belongs to a paragraph that can be typed by parallelism of "argumentative" sentences marked by a conjunction, and also belongs to a part that is characterised by the presence of a strong adverbial connector, then its probably is a correct segment". All these features concur to characterise a well-marked "argumentative" part and such parts characterise well-written texts.

It is particularly interesting to note that comma-units are not characterised by descriptors *sui generi* in well-written texts, but because they belong to a well-marked context.

In Table 3, the textual measures where features appear and from which the comma-unit inherits are noted P for Part, § for paragraph, S for sentence, and CU for comma-punctuated unit.

Table 3. *Examples of class characterisation rules concerning comma-delimited units (22,041 comma-units in the data)*

Characterisation rules	Class	Frequency	Confidence
P As well as S where	CORRECT	229	1
P PASSIVE IMPERSO § ISOPERIPH S by	CORRECT	837	0,96
P ADV § CONJ § ISOPERIPH S semicolon	CORRECT	295	0,89
P FUTURE § ISOPERIPH S THERE	CORRECT	499	0,93
S PARENTHESIS CU an CU the CU that	INCORRECT	47	0,37
S PARENTHESIS CU an CU the CU with	INCORRECT	36	0,37
S PARENTHESIS CU the CU for CU and	INCORRECT	27	0,36
S PARENTHESIS CU the CU to CU if	INCORRECT	31	0,30
S PARENTHESIS CU the CU to CU []	INCORRECT	40	0,30
S PARENTHESIS CU the CU with CU and	INCORRECT	22	0,26
S PARENTHESIS CU the CU with CU or	INCORRECT	22	0,23
S PARENTHESIS CU the CU it CU []	INCORRECT	57	0,26
S PARENTHESIS CU the CU or CU []	INCORRECT	33	0,26

By contrast, the produced rules exhibit correlation of textual markers used in poorly written papers. The main result is that in poorly-written texts, there is little or no correlation between parts and paragraphs and sentences. Lack of coherence is the contrary of clearly contrasted text measures that characterised well-written articles. Rules show correlation only in a very narrow window, sentences and comma-punctuated units. Mined rules are interesting although no strong tendency emerges to type families of correlated mistakes. Our sample contains articles from various geographic origins, and articles written with varying degrees of competence. The confidence threshold is set at a low level in order to catch some regularities, for instance the fact that in a sentence with parenthesis, the co-occurrence of *the* and *an* signals awkwardness. Such information obtained from patterns is useful because the potential mistakes or mismatches in academic English are extremely numerous and we cannot even figure them out. Instead of asking a professional corrector to list frequent mistakes by hand, we retrieve information both on mistakes and on their context from the corpus itself.

Interestingly, a fair number of characterisation rules highlight the role of special punctuation marks, such as parenthesis. This is linked with the conventional use of bibliographical references and other deictic devices, which play a very important role in structuring academic articles (Lucas 2003). Misplacement of references is a common mistake, and human correctors frequently check them.

This work is an example of cross-fertilisation between textual linguistics and text mining techniques and call for further exploration on a much larger sample of texts. Future work is planned on high-level descriptors involving more elaborate pre-processing for a better liability. A larger number of forms should also be used to catch regularities in poorly-written articles.

At this stage, results prove the soundness of our linguistic description of well-written academic articles. Using segments also enables to produce a finer description of our corpus. Incorrect segments can be highlighted in any text, which proves to be a valuable practical help for computer-assisted copy-editing.

Bibliographical references

Ahonen-Myka H, Heinonen O, Klemettinen M, Verkamo A I 1999 Finding co-occurring text phrases by combining sequence and frequent set discovery. In *Proceedings of the workshop on text mining: foundations, techniques and applications*, pp. 1-9.

Agrawal R, Imielinski T, Swami A 1993 Mining Association Rules between Sets of Items in Large Database. In *Proceedings of ACM SIGMOD 93*, pp. 207-216.

Boulicaut JF, Bykowski A, Rigotti C 2000 Approximation of frequency queries by means of free-sets. In *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD'00*, Lyon, Springer, pp. 75-85.

Boulicaut JF, Bykowski A, Rigotti C 2003 Free-sets: a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery journal,* 7(1):5-22.

Crémilleux B, Boulicaut JF 2002 Simplest Rules Characterizing Classes Generated by ∂-Free Sets, In *Proceedings of 22nd Annual International Conference of the British Computer Society's Specialist Group on Artificial Intelligence (ES 02)*, Cambridge, Springer, pp. 33-46.

Feldman R, abd Dagan I, Klösgen W 1996 Efficient algorithms for mining and manipulating associations in texts, In *Thirteenth European Meeting on Cybernetics and Systems Research, Cybernetics and Systems*, Vienna, Volume II.

Grabe W 1987 Contrastive Rhetoric and Text-type Research. In Connor, Kaplan (Eds) *Writing across languages: Analysis of L2 Text.* Reading, MA, Addison-Wesley, pp. 115-137.

Harris Z 1952 Discourse analysis. *Language* (28): 1-30.

Hockett C 1958. *A Course in Modern Linguistics*. New York.

Lent B, Agrawal R, Srikant R 1997 Discovering trends in text databases. In *Proceedings of Third International Conference on Knowledge Discovery and Data Mining KDD 97*, Newport Beach, California, AAAI Press, pp. 227-230.

Lucas N 1991 Syntaxe discursive du japonais scientifique: à propos du thème. In Centro de estudos japoneses da Universidade de Sao Paulo (Ed.), *II encontro nacional de professores universitários de língua, literatura e cultura japonesa.*São Paulo, São Paulo University, pp. 77-98.

Lucas N 2003 La citation et l'appel à référence bibliographique dans les articles académiques. In Marnette, Rosier, López Muñoz (eds) *Le discours rapporté dans tous ses états Question de frontières*. Bruxelles, Duculot (in press).

Mannila H, Toivonen H, 1997 Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3), pp. 241-258,

Pasquier N, Bastide Y, Taouil R, Lakhal L 1999 Efficient Mining of Association Rules Using Closed Itemset Lattices. *Information Systems*, 24(1):25-46.

Péry-Woodley MP 1989 *Textual designs: signalling coherence in first and second language academic writing*. Lancaster University thesis, Notes et documents LIMSI 91-1.

Péry-Woodley MP 1993 Une pragmatique à fleur de texte: marques superficielles des opérations de mise en texte. In Moirand. *et al.* (eds), *Parcours linguistiques de discours spécialisés*, Bern, Peter Lang, pp. 337-348.