Rapport sur `pdftotext`

1 Examen du premier article avec ou sans options

Examen option par option pour le fichier `Boudin-Torres-2006.pdf`.

1.1 Sans option

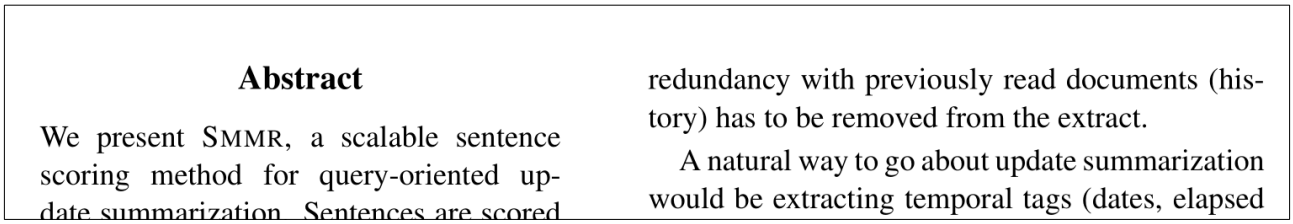
Le titre apparaît bien sur les deux premières lignes, mais il n'y a pas de ligne vide pour le séparer de l'en-tête qui contient les références des auteurs, ce qui rendra l'analyse compliquée :

```
A Scalable MMR Approach to Sentence Scoring
for Multi-Document Update Summarization
Florian Boudin \ and Marc El-Bèze \
\
Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, BP1228,
```

Le document original contient des bécarres (♮) et des bémols (♭) qui ne sont pas correctement retranscrits, mais ces symboles ont l'air d'être des erreurs. Les lettres accentuées sortent impeccablement.

Le titre de section « *Abstract* » a bien droit à sa ligne à part entre deux lignes vides, mais ensuite c'est la deuxième colonne de la page qui sort, au lieu de la première. Comparer :

```
Abstract


redundancy with previously read documents (history) has to be removed from the extract.
A natural way to go about update summarization
would be extracting temporal tags (dates, elapsed
```



Abstract

We present SMMR, a scalable sentence scoring method for query-oriented update summarization. Sentences are scored

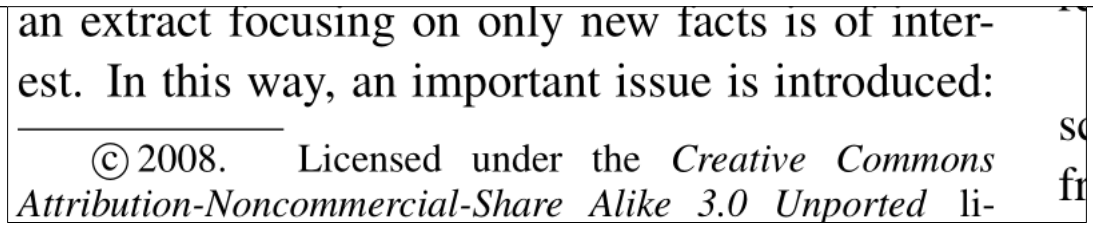
redundancy with previously read documents (history) has to be removed from the extract.

A natural way to go about update summarization would be extracting temporal tags (dates, elapsed

Comme le montre l'extrait ci-dessus, les lignes qui finissent par un mot divisé (-) dans le document original sont enchainées sans retour à la ligne dans le texte de sortie, tandis que les autres sont coupées de la même manière que dans le document original. Les alinéas ne sont pas pris en compte, ils apparaissent comme de simples retours à la ligne.

La colonne de gauche, après le titre « *Abstract* », ne s'affiche qu'ensuite ; elle est quand même séparée de ce qui précède par une ligne vide. Il semble néanmoins difficile de repérer le résumé automatiquement dans ces conditions.

La colonne se termine dans le document original par une note de bas de page, séparée du corps par un filet horizontal. Dans le texte de sortie, il s'enchaîne avec un simple retour à la ligne ; comparer :

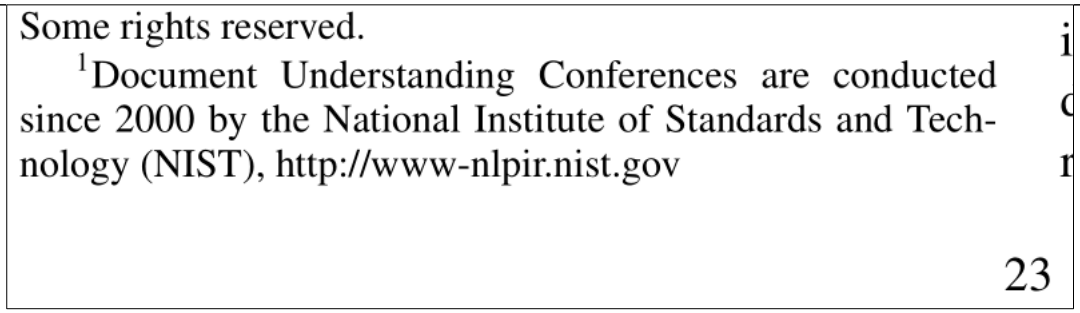

an extract focusing on only new facts is of interest. In this way, an important issue is introduced:

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* li-

```
an extract focusing on only new facts is of interest. In this way, an important issue is introduced:

c 2008.

Licensed under the Creative Commons

Attribution-Noncommercial-Share Alike 3.0 Unported license (http://creativecommons.org/licenses/by-
nc-sa/3.0/).
```

Difficile, là encore, de découper automatiquement le texte correctement.

Il faut remarquer aussi qu'une note de bas de page numérotée se présente à peu près comme un numéro de page :


Some rights reserved.

[1]Document Understanding Conferences are conducted since 2000 by the National Institute of Standards and Technology (NIST), http://www-nlpir.nist.gov

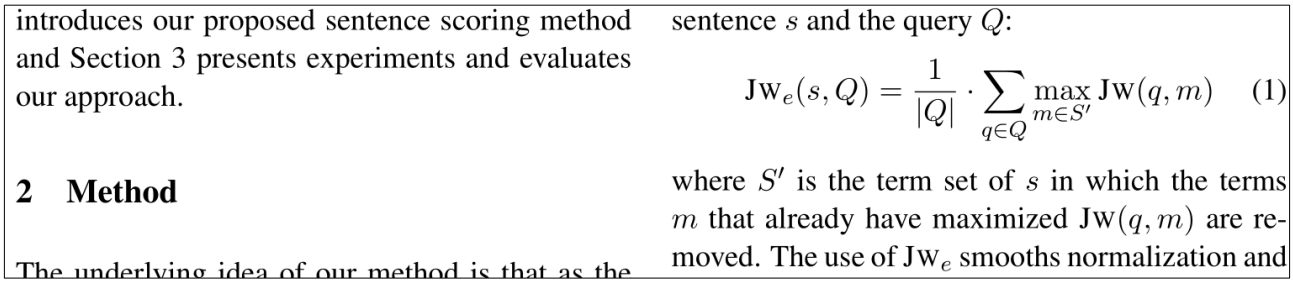
23

```
Some rights reserved.

1

Document Understanding Conferences are conducted

since 2000 by the National Institute of Standards and Technology (NIST), http://www-nlpir.nist.gov


23
```

Le numéro de la page 23 est séparé par une ligne vide avant et après, mais ce n'est pas le cas du numéro 24 (simple retour à la ligne avant, ligne vide après), par exemple. On peut tout de même repérer les numéros de page par ce qu'ils sont sur chaque page la dernière séquence de chiffres ayant une ligne à elle toute seule, et qu'ils sont incrémentés pour chaque page.

Le pied-de-page de la première page (qui n'existe que sur celle-ci) est encadré par des lignes vides en-dessous du numéro, et au-dessus du caractère de saut de page *(form feed)*; il ne pose pas de problème d'identification.

La présence de formules mathématiques n'arrange pas les choses en termes d'ordre de lecture :


introduces our proposed sentence scoring method and Section 3 presents experiments and evaluates our approach.

2 Method

The underlying idea of our method is that as the

sentence s and the query Q:

$$\text{Jw}_e(s,Q) = \frac{1}{|Q|} \cdot \sum_{q \in Q} \max_{m \in S'} \text{Jw}(q,m) \quad (1)$$

where S' is the term set of s in which the terms m that already have maximized $\text{Jw}(q,m)$ are removed. The use of Jw_e smooths normalization and

sort comme :

```
introduces our proposed sentence scoring method

and Section 3 presents experiments and evaluates
```

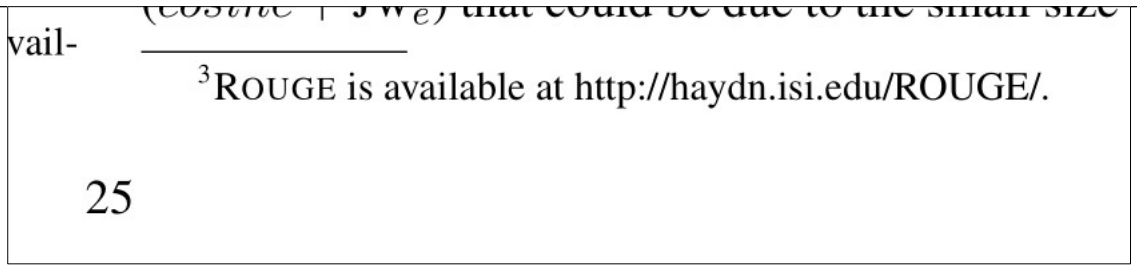
```
our approach.


sentence s and the query Q:
1 X
J We (s, Q) =
.
max J W(q, m)
m∈S 0
|Q|


2


where S 0 is the term set of s in which the terms
m that already have maximized J W(q, m) are removed. The use of J We smooths normalization and
misspelling errors. Each sentence s is scored using
the linear combination:


q∈Q


Method


The underlying idea of our method is that as the
```

Problèmes :
— Comme dit précédemment, le bloc de texte de droite s'enchaîne après les premières lignes de gauche (heureusement avec une ligne vide de séparation) ;
— La formule mathématique est illisible, et il est difficile d'expliquer pourquoi « $q \in Q$ » se retrouve après le bloc de texte qui le suit et avant le titre de la section ;
— Le numéro de section se retrouve quant à lui entre la formule et le bloc de texte qui le suit, ce qui est à peu près cohérent avec le comportement par défaut de `pdftotext`, mais séparé par deux blocs de textes du *titre* de section, alors qu'il devrait être sur la même ligne !
— Le numéro de la formule mathématique n'est pas perdu, mais il se retrouve coincé entre le numéro de la première sous-section (elle aussi loin de son titre) et une autre formule mathématique :

```
measures and the scalable MMR scoring method.
2.1


(1)


~
Sim1 (s, Q) = α · cosine(~s, Q)
+ (1 − α) · J We (s, Q)
```

Le reste est du même acabit. Il faut reconnaître que les caractères spéciaux sont bien rendus, à part le symbole de somme (∑) qui semble avoir disparu.

Au bas de la troisième page, le numéro de page est encadré par un numéro de note et la note elle-même, gênant pour distinguer avec certitude le sens des chiffres :


vail-

[3]ROUGE is available at http://haydn.isi.edu/ROUGE/.

25

```
3


25


ROUGE is available at http://haydn.isi.edu/ROUGE/.
```

Le tableau de la dernière page n'est pas rendu tel quel mais ses données restent groupées de façon logique. En revanche le titre est renvoyé après quelques références bibliographiques de la colonne de droite. Mais, encadré par des lignes vides, il peut être repéré par son format :

```
Table 1: ROUGE average recall scores computed

on the DUC 2007 update corpus.
```

1.2 Option `-raw`

Avec cette option, le texte est lu dans l'ordre dans lequel il est enregistré dans le document P.D.F. Si ceux-ci sont bien faits, cela devrait résoudre une bonne partie des problèmes du sans-option.

Problème remarquable : Si cette fois-ci, chaque ligne est telle que sur le document original (pas de lignes regroupées là où un mot était divisé), aucune ligne n'est laissée vide pour séparer les blocs de texte. Il s'agirait donc de repérer des champs tous à la suite les uns des autres.

Le champ qui apparaît au début du document est le pied de la première page, et il semble difficile de le distinguer du titre, et celui-ci des références des auteurs :

```
Coling 2008: Companion volume – Posters and Demonstrations, pages 23–26

Manchester, August 2008

A Scalable MMR Approach to Sentence Scoring

for Multi-Document Update Summarization

Florian Boudin \

and Marc El-Bèze \

\
```

Les bandes (\) ne sont que des bécarres parasites qui n'apparaitraient pas dans d'autres documents et ne sauraient donc servir de repère pour la délimitation.

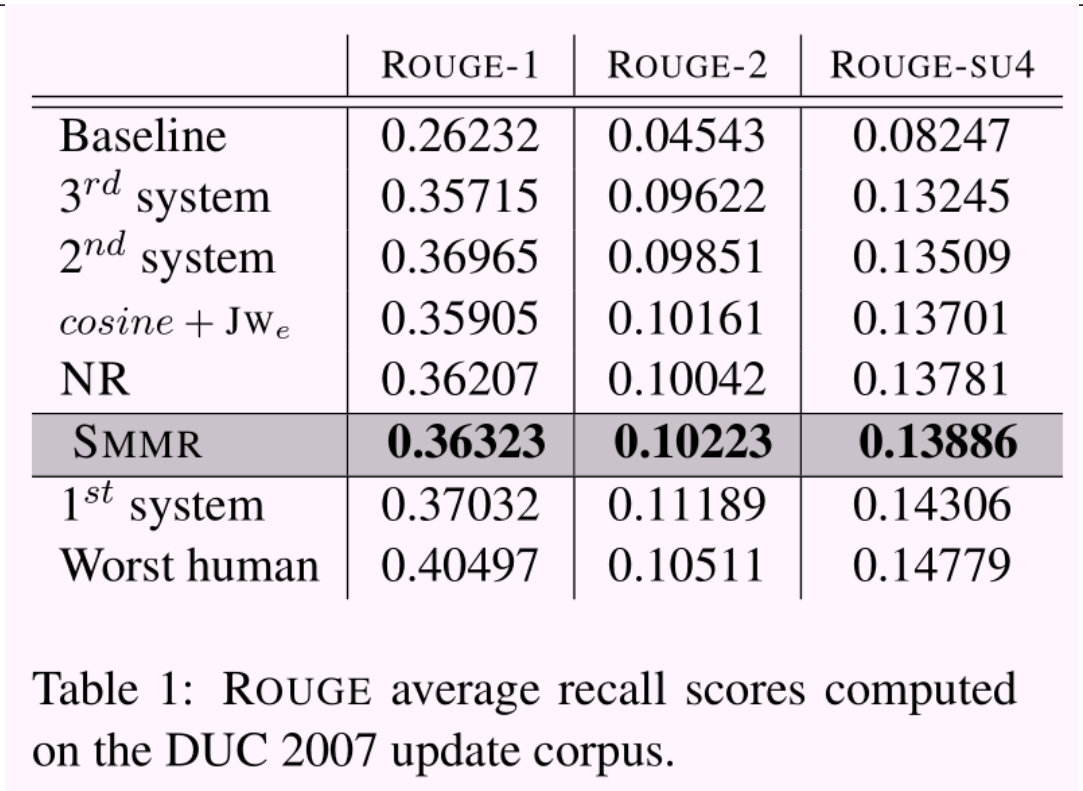
Progrès : Après la ligne « *Abstract* » qui identifie clairement le bloc de résumé, celui-ci se trouve effectivement tel quel, suivi d'une ligne « *1 Introduction* » qui précède effectivement l'introduction.

Bémol : les notes de bas de page sont toujours dans le flux du texte et ne sont pas plus dinstinguible qu'avant. Dièse : les notes de bas de page ne font pas forcément partie des *sections* de l'article qu'il nous faudra distinguer, et si quelqu'un lit le corps du document parsé, il devrait être en mesure de comprendre qu'il ne s'agit pas du texte principal.

Du reste, le texte est dans l'ordre, les numéros de section précèdent bien les titres sur une ligne à part, certains numéros de note peuvent ressembler à des numéros de page comme dans l'exemple déjà cité, mais cela ne devrait pas être insurmontable. La bibliographie est annoncée par une ligne « *References* ».

L'affichage des formules mathématiques est désastreux, mais on ne peut pas espérer beaucoup d'un format texte brut de ce côté, et elles ont au moins le mérite d'être sur des lignes consécutives avec leur numéro d'équation.

Les alinéas et retraits divers sont bien sûr ignorés, l'affichage du tableau est cependant peut-être plus compréhensible car les lignes sont conservées comme telles :



	ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline	0.26232	0.04543	0.08247
3^{rd} system	0.35715	0.09622	0.13245
2^{nd} system	0.36965	0.09851	0.13509
$cosine + \text{JW}_e$	0.35905	0.10161	0.13701
NR	0.36207	0.10042	0.13781
SMMR	**0.36323**	**0.10223**	**0.13886**
1^{st} system	0.37032	0.11189	0.14306
Worst human	0.40497	0.10511	0.14779

Table 1: ROUGE average recall scores computed on the DUC 2007 update corpus.

```
ROUGE-1 ROUGE-2 ROUGE-SU4
Baseline 0.26232 0.04543 0.08247
3rd system 0.35715 0.09622 0.13245
2nd system 0.36965 0.09851 0.13509
cosine + JWe 0.35905 0.10161 0.13701
NR 0.36207 0.10042 0.13781
SMMR 0.36323 0.10223 0.13886
1st system 0.37032 0.11189 0.14306
Worst human 0.40497 0.10511 0.14779
Table 1: ROUGE average recall scores computed
on the DUC 2007 update corpus.
```

Il faut garder à l'esprit que les lignes présentées ici sont encadrées par le reste du texte, dans l'ordre, mais sans séparation. On peut faire la même remarque que pour les notes de bas de page, que le lecteur saura reconnaître qu'il ne s'agit pas du corps du texte à proprement parler.

Conclusion provisoire pour cette solution. — Si l'on ne veut pas entrer dans les détails moins importants qui sont perdus ou perturbés à la traduction, cette option est correcte, mais le plus difficile risque d'être d'identifier le titre et les noms des auteurs, enchainés à un pied-de-page qui se retrouve en en-tête. Il me reste à voir le résultat sur les autres articles, pour ne pas avoir de mauvaise surprise.

1.3 Option `-layout`

Ici, le texte est disposé, autant que possible, comme dans le document original ; l'ajustement se fait par des retours à la ligne et des espaces.

Le titre est clairement distingué des noms et coordonnées des auteurs :

```
                    A Scalable MMR Approach to Sentence Scoring
                     for Multi-Document Update Summarization


       Florian Boudin \ and Marc El-Bèze \                          Juan-Manuel Torres-Moreno \,[
        \
          Laboratoire Informatique d’Avignon                         [
                                                                       École Polytechnique de
Montréal
        339 chemin des Meinajaries, BP1228,                        CP 6079 Succ. Centre Ville H3C
3A7
           84911 Avignon Cedex 9, France.                               Montréal (Québec), Canada.
           florian.boudin@univ-avignon.fr                          juan-manuel.torres@univ-
avignon.fr
             marc.elbeze@univ-avignon.fr
```

Sur le présent document, certaines lignes sont divisées car trop longues, mais il y a bien deux blocs bien distincts présentant les deux auteurs, et séparés par une ligne vide du titre.

Les alinéas sont conservés et les deux colonnes sont bien visibles dans chaque page.

Difficulté : Si les lignes des deux colonnes sont bien alignées dans le texte original, en face-à-face, la sortie est bien alignée comme dans l’extrait ci-dessus ; mais si elles sont en quinconce, on voit une séquence de lignes qui ne contiennent que l’une des deux colonnes, alternativement :

```
Decreasing λ in (3) with the length of the sum-
                                                                    Most existing automated evaluation
methods work
mary was suggested by (Murray et al., 2005) and
                                                                    by comparing the generated
summaries to one or
```

Il est facile de repérer et de corriger automatiquement ces alternances en remettant les lignes face à face, mais elles pourraient masquer le fait que quelques lignes sont réellement vides d’un côté ou de l’autre.

Le pied de la première page pourrait être repéré par ce que, même sans le filet de séparation, il commence par un alinéa différent des autres. Mais ça demandera beaucoup d’analyse dans tous les cas.

Les titres sont correctement alignés avec leurs numéros et relativement faciles à identifier.

Les formules sont loin d’être parfaitement représentées, mais sans doute plus clairement :

```
                1 X
J We (s, Q) =     ·   max J W(q, m)        (1)
              |Q|     m∈S 0
                    q∈Q
```

Le tableau est présentable tel quel :

```
                    ROUGE -1   ROUGE -2   ROUGE - SU 4

    Baseline        0.26232    0.04543      0.08247
    3rd system      0.35715    0.09622      0.13245

    2nd system      0.36965    0.09851      0.13509

    cosine + J We   0.35905    0.10161      0.13701
    NR              0.36207    0.10042      0.13781
     S MMR          0.36323    0.10223      0.13886

    1st system      0.37032    0.11189      0.14306
    Worst human     0.40497    0.10511      0.14779
```

```
Table 1: ROUGE average recall scores computed
on the DUC 2007 update corpus.
```

La détection des colonnes peut être problématique mais je pense qu'elle est assez facile en repérant les longues séquences de lignes sans caractères sur les mêmes colonnes. Cette détection doit être faite sur chaque page indépendamment, car la gouttière (colonne vide) peut être déplacée selon le contenu de la colonne de gauche. `pdftotext` semble faire en sorte que la disposition soit régulière sur chaque page sans tenir compte des autres.

Problème : risque de confusion entre un texte en colonnes et un tableau ?

Conclusion provisoire pour cette solution. — Mon rendu préféré, car il fait perdre le moins d'information par rapport au document d'origine, et ne me paraît pas si difficile à traiter en fin de compte. À étudier là encore sur les autres documents, pour voir si les colonnes sont bien distinctes et les tableaux bien reconnaissables.

2 Examen des autres articles avec les options retenues

C'est-à-dire avec l'option `-raw`, puis avec l'option `-layout`. Sauf mention contraire, ce qui a été dit pour `Boudin-Torres-2006.pdf` reste valable pour les autres documents.

2.1 Avec l'option `-raw`

2.1.1 En-tête

La diversité est de mise sur cette partie des articles. Essayons de repérer des éléments récurrents ou problématiques.

Pour certains articles, les premières lignes sont occupées par le pied de la première page.

Boudin – Torres :

```
Coling 2008: Companion volume – Posters and Demonstrations, pages 23–26

Manchester, August 2008
```

Iria – Juan-Manuel Gerardo :

```
Proceedings of the Fifth Law Workshop (LAW V), pages 1–10,

Portland, Oregon, 23-24 June 2011. c 2011 Association for Computational Linguistics
```

Mikheev :

```
c

# 2002 Association for Computational Linguistics
```

Nasr :

```
Proceedings of the ACL-HLT 2011 System Demonstrations, pages 86–91,

Portland, Oregon, USA, 21 June 2011. c 2011 Association for Computational Linguistics
```

On peut observer que dans trois cas sur les quatre, la première ligne indique un intervalle de pages et la seconde contient notamment une date, ce qui peut constituer des indices permettant de repérer ces pieds-de-page. Pour Mikheev, le `c` seul peut être interprété comme un symbole de copyright, qui ne peut normalement pas faire partie du titre. Autre constat : si le texte est sur une seule colonne, le pied-de-page se trouve bien en pied de page — du moins dans les exemples donnés.

Il n'y a plus qu'à espérer que d'autres articles scientifiques ne fassent pas exception aux généralités énoncées ci-dessus.

L'article Torres – Moreno est un cas particulier parce qu'il commence par une ligne d'en-tête `LETTER Communicated by Scott Fahlman`. Rien ne permet de le distinguer facilement d'un titre…

Dans tous les documents, ces lignes éventuelles de pied-de-page ou d'en-tête sont immédiatement suivies du titre, toujours écrit correctement mais qui peut prendre une ou deux lignes — la deuxième commençant ou non par une majuscule —, contenir des sigles, des points d'abréviation… tout comme les noms des auteurs, qui peuvent être abrégés ou non.

On trouve deux structures différentes pour ces noms d'auteurs :

- Chaque nom ou groupe de noms peut remplir une ligne, suivie d'une ou plusieurs lignes de coordonnées, parmi lesquelles on peut reconnaître des adresses, mais où d'autres lignes semblent plus difficiles à analyser :

```
Florian Boudin \
and Marc El-Bèze \
\
Laboratoire Informatique d'Avignon
```

  Dans cet extrait, comment deviner sans dictionnaire que les deux premières lignes sont des noms, et la dernière non ?

  Les deux Kessler ont un avantage : chaque auteur a son bloc annoncé par un nombre ordinal :

```
1st
Rémy Kessler
Université Bretagne Sud
CNRS 6074A
56017 Vannes,France
remy.kessler@univ-ubs.fr
2nd
Nicolas Béchet
Université Bretagne Sud
CNRS 6074A
56017 Vannes,France
nicolas.bechet@irisa.fr
3rd
Giuseppe Berio
Université Bretagne Sud
CNRS 6074A
56017 Vannes,France
giuseppe.berio@univ-ubs.fr
```

  Mais ça ne marche que pour les Kessler.

- Les noms peuvent se retrouver tous ensemble sur une ligne, ou sur deux :

```
Dipanjan Das André F.T. Martins
Language Technologies Institute
```

  Impossible de distinguer les différents auteurs, et il peut être difficile de distinguer aussi les coordonnées.

2.1.2 Résumé

Deux écoles : dans la plupart des articles, le résumé est annoncé par le mot *Abstract*, soit seul sur une ligne, soit en début de ligne et séparé du contenu par un point ou un tiret. Il faudra seulement prendre garde aux cas où le mot *abstract* se trouverait dans un titre. Mais dans Mikheev et Torres Moreno, le résumé suit

directement le bloc de coordonnées de l'auteur, sans titre. Il faut donc prévoir de pouvoir distinguer les lignes de coordonnées du début du résumé.

2.1.3 Titres de sections particulières

L'introduction ne pose pas de problème, elle est systématiquement annoncée par une ligne *Introduction*, mot éventuellement précédé d'un nombre en chiffres arabes ou romains, éventuellement un point, et toujours un espace.

La bibliographie est également facile à repérer, car elle est annoncée par une ligne *References*, mot en majuscules ou en minuscules.

Les autres titres de sections, notamment discussion et conclusion, quand elles sont présentes, ne posent généralement pas de problème, car ils sont toujours sur une ligne à part, commençant par un numéro en chiffres arabes ou romains. Comme en plus ces numéros se suivent dans l'ordre, il ne devrait pas être difficile d'identifier ces champs.

2.1.4 Figures, tableaux et formules mathématiques

Les formules mathématiques souffrent beaucoup, mais elles sont assez faciles à repérer car elles ont toujours un numéro entre parenthèses à la fin, *sauf* les formules sur une ligne, par exemple dans un algorithme, qui le rendent à peu près incompréhensible :

```
• Initialize
h = 0;
set the targets τ
μ
h+1 = τμ for μ = 1, . . . , P;
[Page suivante]1012 J. Manuel Torres Moreno and Mirta B. Gordon
• Repeat
1. /* train the hidden units */
h = h + 1; /* connect hidden unit h to the inputs */
learn the training set {E
ξμ, τ
```

De même, les figures et tableaux sont précédés *ou* suivis (mais toujours de la même façon dans un même article) par une ligne commençant par `Table`, `Figure` ou `Fig.` Suivi d'un numéro.

Il n'est pas forcément essentiel de les repérer en tant que tels, mais cela peut être utile pour mieux analyser la cohérence du document, notamment pour repérer le premier titre après l'introduction et ainsi délimiter cette dernière.

2.1.5 Appels et notes de bas de page et numéros de page

Les numéros de notes de bas de page sont soit seuls sur une ligne, soit accolés à la note elle-même. Ils pourraient être confondus avec les numéros de page, mais ces derniers sont toujours sur la toute dernière ligne de chaque page… sauf pour Torres Moreno où ils sont sur la première ligne. Une possibilité serait de repérer la façon dont sont organisés les numéros de page, soit en début soit en fin de page, grâce au fait qu'il se suivent, et, les numéros de page isolés, de considérer les numéros de notes par ce qu'ils se suivent également.

Si l'on veut distinguer les notes de bas de page, il faut aussi les distinguer leurs numéros des chiffres des formules mathématiques qui pourraient se retrouver isolés sur une ligne.

Conclusions sur `-raw`

Une bonne partie du travail pour le parseur pourrait consister à d'abord analyser la mise en page utilisée, en supposant qu'elle est constante dans l'article, et en prévoyant pour chaque détail un certain nombre de

mises en pages possibles. Une fois que la mise en page est comprise par le parseur, en analysant la cohérence des numéros et appels de note, des séquences de numéros de pages, de titre, etc., il devient plus facile d'extraire les différents champs.

2.2 Avec l'option `-layout`

Le rendu colle autant que possible à la mise en page du document original. Pour le problème de la reconnaissance des colonnes, se reporter au paragraphe de la première partie *Option -layout*.

2.2.1 En-tête

Le titre principal est généralement facile à identifier, car il se trouve seul sur les premières lignes, et est séparé des noms des auteurs par une ligne vide. Exceptions :

- Iria – Juan-Manuel Gerardo : Les noms des auteurs suivent directement le titre, mais ils sont séparés par de larges espaces qui les rendent impossibles à confondre avec celui-ci :

```
                On the Development of the RST Spanish Treebank

      Iria da Cunha                Juan-Manuel Torres-Moreno                                Gerardo
Sierra
```

- Plus problématique, Torres, car à part la différence de longueur des lignes, rien ne permet de distinguer celles du titre de celle des auteurs :

```
                                    Summary Evaluation

                                  with and without References

   Juan-Manuel Torres-Moreno, Horacio Saggion, Iria da Cunha, Eric SanJuan, and Patricia Velázquez-
Morales
```

Pour ce qui est de distinguer les noms des auteurs entre eux et des autres informations de l'en-tête, ces noms sont en général tous sur la même ligne, et séparés soit par un grand espace, soit par de la ponctuation. Cas particuliers :

- Mikholov : Les deux premiers auteurs sont sur la même ligne, chacun au-dessus d'un bloc de coordonnées, et la même structure est répétée pour les deux autres auteurs, après quelques sauts de ligne.
- Un peu plus embêtant, Nasr : tous les noms sont sur la même ligne, séparés par des espaces simples :

```
    Alexis Nasr Frédéric Béchet Jean-François Rey Benoît Favre Joseph Le Roux∗
```

- Dans Torres-Moreno, les noms des deux auteurs, en gras dans l'original, sont sur des lignes consécutives juste devant les coordonnées, ce qui rend leur identification plus difficile car dans tous les autres documents, la ligne suivant un nom est réservée à ces coordonnées :

```
J. Manuel Torres Moreno

Mirta B. Gordon

Département de Recherche Fondamentale sur la Matière Condensée, CEA Grenoble,

38054 Grenoble Cedex 9, France
```

2.2.2 Résumé

Si l'on considère à part la détection des colonnes et la correction des lignes présentées en quinconce, le résumé, annoncé ou non par le mot *Abstract*, est toujours d'un seul bloc, toujours précédé et suivi d'une ligne vide, et toujours suivi du titre *Introduction*. Le détecter ne paraît donc pas très difficile.

2.2.3 Titres de sections particulières

Un petit problème apparaît qui n'existait pas avec l'option `-raw` : les grandes capitales suivies de petites capitales prennent un espace, par exemple dans Torres, l'introduction s'intitule `I. I NTRODUCTION`. On peut essayer de traiter cette erreur en reconaissant sa répétition au fil des titres. Rien d'autre n'est à

signaler, sinon que ces titres sont toujours précédés et presque toujours suivis d'un saut de page ou d'une ou plusieurs lignes vides (du moins dans leur colonne), et qu'à l'intérieur d'un article ils suivent le même modèle, ce qui est bon pour leur reconnaissance.

2.2.4 Figures, tableaux et formules mathématiques

Pour les formules mathématiques, comme indiqué plus haut, si elles ne sont pas retranscrites parfaitement (les symboles complexes comme la somme, la racine carrée, la barre de fraction passent à la trappe), elles sont beaucoup plus lisibles et cohérentes que sans l'option `-layout`.

Il en va de même des tables, qui ont un rendu très régulier et sont précédées et suivies de lignes vides. Comparer dans Gonzalez 2018 :

```
                  Table 1. Sentence Boundary Detection example


       Speech transcript                    SBD applied to transcript
       two two women can look out after     two // two women can look out
       a kid so bad as a man and a          after a kid so bad as a man and a
       woman can so you can have a you      woman can // so you can have a
       can have a mother and a father       // you can have a mother and a
       that that still don't do right with  father that // that still don't do
       the kid and you can have to men      right with the kid and you can
       that can so as long as the love      have to men that can // so as
       each other as long as they love      long as the love each other // as
       each other it doesn't matter         long as they love each other it
                                            doesn't matter//
```

contre (`-raw`)

```
Table 1. Sentence Boundary Detection example
Speech transcript SBD applied to transcript
two two women can look out after
a kid so bad as a man and a
woman can so you can have a you
can have a mother and a father
that that still don't do right with
the kid and you can have to men
that can so as long as the love
each other as long as they love
each other it doesn't matter
two // two women can look out
after a kid so bad as a man and a
woman can // so you can have a
// you can have a mother and a
father that // that still don't do
right with the kid and you can
have to men that can // so as
long as the love each other // as
long as they love each other it
doesn't matter//
```

L'algorithme de tout à l'heure prend un peu plus de sens, si l'on interprète correctement les exposants.

```
   • Initialize
     h = 0;
                        μ
     set the targets τh+1 = τ μ for μ = 1, . . . , P;
[Page suivante]1012                                  J. Manuel Torres Moreno and Mirta B. Gordon



   • Repeat


         1. /* train the hidden units */
            h = h + 1; /* connect hidden unit h to the inputs */
                                              μ
            learn the training set {ξE μ , τh }, μ = 1, . . . , P;
```

Les figures sont toujours incompréhensibles mais elles sont d'autant mieux balisées avec des lignes vides.

2.2.5 Appels et notes de bas de page, numéros de page

Les notes de bas de page sont assez facilement reconnaissable dans la plupart des documents, à leur disposition, par exemple dans Mikolov :

```
     3
         We thank Geoff Zweig for providing us the test set.
```

Dans d'autres documents, le numéro est sur la même ligne que la note, toujours avec une ligne de séparation ou un alinéa par rapport au texte principal. La règle ne change pas à l'intérieur d'un même document, mais la taille de l'alinéa peut changer légèrement d'une page à l'autre, puisque `pdftotext` adapte les espaces au contenu de chaque page.

Les numéros de page sont aussi reconnaissables qu'avec l'option précédente, et même un peu mieux parce qu'elles sont souvent centrées.

2.2.6 Note supplémentaire sur les colonnes

Un survol rapide permet de voir que pour la plupart des articles écrits sur deux colonnes — il est peu probable qu'on tombe sur plus —, ces colonnes sont assez clairement délimitées, et seuls l'en-tête de l'article, le pied-de-page et les numéros de page peuvent s'insérer au milieu. Dans Torres, on trouve aussi des tableaux centrées, insérées au milieu d'un texte sur deux colonnes dans la même page :

```
                                                    TABLE V
                    S PEARMAN ρ OF CONTENT- BASED MEASURES WITH ROUGE IN THE Medicina Clínica C
ORPUS (S PANISH )


                            Mesure      ROUGE -1       p-value     ROUGE -2       p-value
ROUGE -SU4         p-value
                            JS            0.56       p < 0.100       0.46       p < 0.100
0.45          p < 0.200
                            J S2          0.88       p < 0.001       0.80       p < 0.002
0.81          p < 0.005
                            J S4          0.88       p < 0.001       0.80       p < 0.002
0.81          p < 0.005
                            J SM          0.82       p < 0.005       0.71       p < 0.020
0.71          p < 0.010
```

```
     found weak correlation among different rankings in                          a representation of
the task/topic in the calculation of

     complex summarization tasks such as the summarization                       measures. To carry
out these comparisons, however, we are

     of biographical information and the summarization of                        dependent on the
existence of references.
```

Un parsage intelligent devrait permettre de détecter ces tables et de les distinguer du flux principal.

Un document pose problème : Das Martin, dont les graphiques provoquent un grand désordre. On s'en rend compte en sélectionnant le texte dans le document P.D.F. :

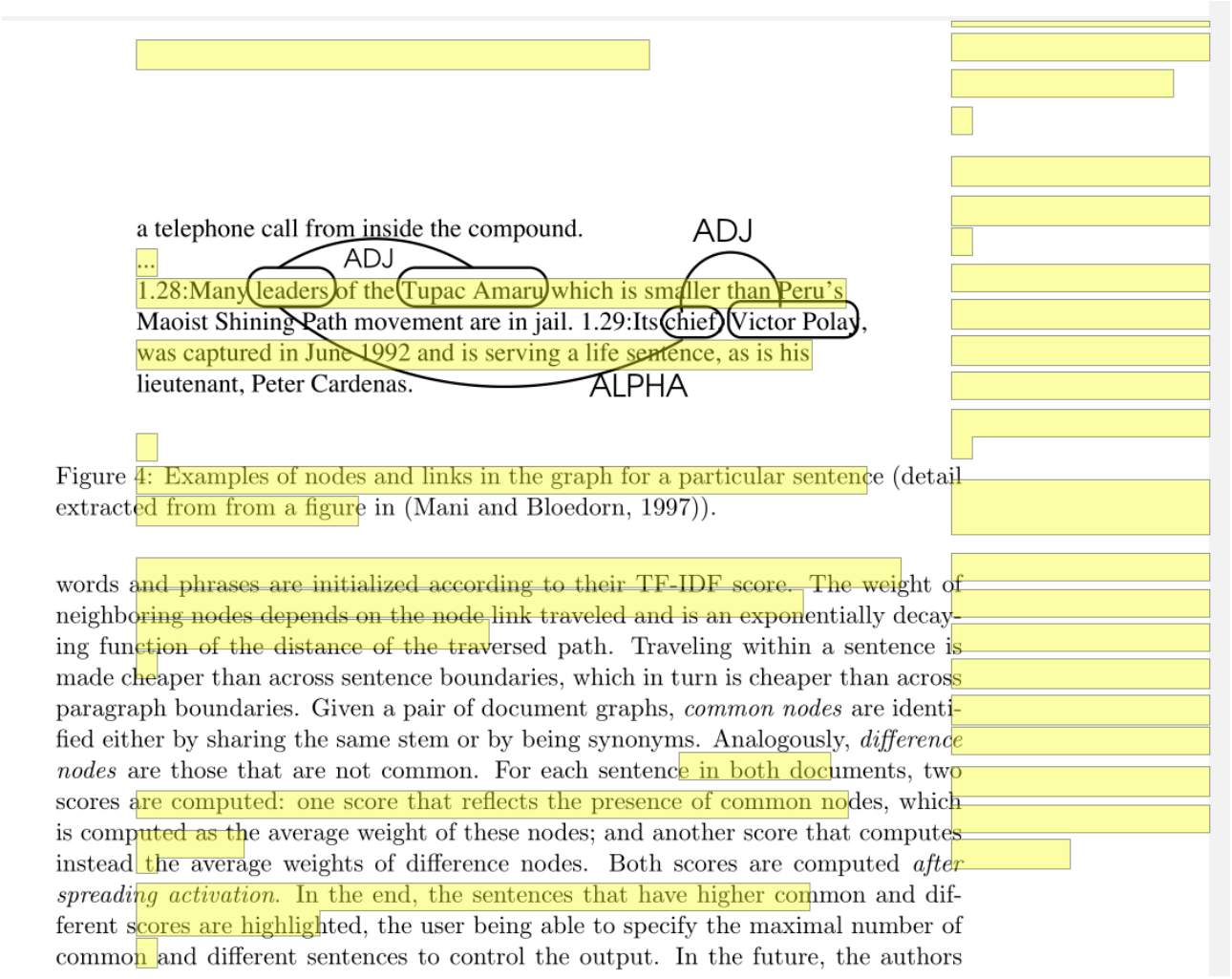


Figure 4: Examples of nodes and links in the graph for a particular sentence (detail extracted from from a figure in (Mani and Bloedorn, 1997)).

words and phrases are initialized according to their TF-IDF score. The weight of neighboring nodes depends on the node link traveled and is an exponentially decaying function of the distance of the traversed path. Traveling within a sentence is made cheaper than across sentence boundaries, which in turn is cheaper than across paragraph boundaries. Given a pair of document graphs, *common nodes* are identified either by sharing the same stem or by being synonyms. Analogously, *difference nodes* are those that are not common. For each sentence in both documents, two scores are computed: one score that reflects the presence of common nodes, which is computed as the average weight of these nodes; and another score that computes instead the average weights of difference nodes. Both scores are computed *after spreading activation*. In the end, the sentences that have higher common and different scores are highlighted, the user being able to specify the maximal number of common and different sentences to control the output. In the future, the authors

Le texte de sortie fait apparaître des extraits invisibles au milieu de la colonne, et même une deuxième colonne virtuelle :

```
       lieutenant, Peter Cardenas.                    ALPHA                                  rebels
belonging to the M

       1.30:Other top commanders conceded defeat and surrendered in July 1993. and better-known
Maoist

       ...                                                                                   ...

       1.32:

Figure 4:     President

           Examples   of Alberto Fujimori,
```

```
                        nodes and          who
                                    links in theisgraph
                                                of Japanese  ancestry, has
                                                      for a particular     had (detail
                                                                                sentence
```

Cette fois-ci, le rendu avec `-raw` est plus lisible, bien que le graphique ne soit pas distingué clairement du texte principal :

```
definite: yes
charge
class: verb voice :passive
polarity: +
tense: past
Figure 4: Dependency grammar representation of the sen-
tence “McVeigh, 27, was charged with the bombing”.
sentence that meets some criteria (e.g., a threshold number
of common content words). In practice, however, any repre-
sentative sentence will usually include embedded phrase(s)
containing information that is not common to all sentences
```

Peut-être faudrait-il se servir des deux conversions à la fois pour repérer et traiter proprement ce cas problématique.

Conclusions sur `-layout`

Je reste sur ma position que c'est la meilleure option à prendre, bien qu'elle réclame un certain travail pour détecter la présence de deux ou d'une seule colonne. En contrepartie, la plupart des champs devraient être plus faciles à extraire dans la plupart des cas, sans recourir à une analyse complète des normes de mise en page dans l'article. De plus, certains éléments seront bien plus lisibles si présentés tels que dans le document d'origine, plutôt qu'analysés et affichés linéairement.

Conclusion générale

Si nous choisissons l'outil `pdftotext` plutôt que `pdf2txt` pour la réalisation du parseur, je préconise de recourir principalement à l'option `-layout`, et, si l'on veut traiter certains cas problématiques comme celui de Das Martin, qui est susceptible de revenir dans d'autres articles, de mettre au point un algorithme qui puisse le détecter et corriger la sortie en la croisant avec le résultat obtenu avec d'autres options.

Si `pdf2txt` permet d'obtenir le texte dans l'ordre *et* avec des champs séparés par des lignes vides ; et si cela permet, notamment, d'extraire assez facilement les champs de l'en-tête, qui sont les plus problématiques ; alors ce sera probablement la meilleure solution.