Rapport pdf2text :

Exemple de l'article Boudin-Torres-2006.pdf

Sans option :

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

Manchester, August 2008

23

AScalableMMRApproachtoSentenceScoringforMulti-DocumentUpdateSummarizationFlorianBoudin\andMarcEl-B`eze\\LaboratoireInformatiqued'Avignon3
entconceptsarese-lected,orderedandassembledaccordingtotheirrelevancetoproducesummaries(alsocalledex-tracts)(ManiandMaybury,1999).Recently
onstructthetimelinefromdocuments(SwanandAllan,2000).Thesetemporalmarkscouldbeusedtofocusextractsonthemostrecentlywrittenfacts.However,mos
es.Inthispaper,weproposeascalablesentencescoringmethodforupdatesummarizationderivedfromMMR.Motivatedbytheneedforrelevantnovelty,candidate

introducesourproposedsentencescoringmethodandSection3presentsexperimentsandevaluatesourapproach.2MethodTheunderlyingideaofourmethodisthat
erm-spaceΓ,whereNisthenumberofdifferenttermsfoundinthecluster,isconstructed.SentencesarerepresentedinΓbyvectorsinwhicheachcomponentisthet
gthelinearcombination:Sim1(s,Q)=α·cosine(~s,~Q)+(1−α)·JWe(s,Q)(2)whereα=0.7,optimallytunedonthepastDUCsdata(2005and2006).Thesystemproduce
rob-abilitieseventhoughtheyarenot.Justasrewriting(3)as(NRstandsforNoveltyRelevance):NR=argmaxs∈S[λ·Sim1(s,Q)+(1−λ)·(1−maxsh∈HSim2(s,sh))]
```

On peut constater qu'aucun espace est présent dans l'ensemble du texte le rendant donc inexploitable, donc non pertinent pour en retirer des informations. On voit ici que le texte ne prend pas en compte les accents « Montr´eal ».

Pour remédier au fait qu'il n'y a aucun espace on peut y appliquer l'option -A qui permet un meilleur rendu du texte en le forçant.

Avec option -A :

Pour le texte suivant Das_Martins.pdf nous avons le titre sur une seule ligne

On nous informe du changement de page par l'élément : '' avec au-dessus le nombre de la page

```
1

summarization dialect:
```

Affichage compliqué des équations mathématique

Version pdf : $P(s \in \mathcal{S} \mid F_1, F_2, ..F_k) = \frac{\prod_{i=1}^{k} P(F_i \mid s \in \mathcal{S}) \cdot P(s \in \mathcal{S})}{\prod_{i=1}^{k} P(F_i)}$

Version txt :

```
(cid:81)k
i=1 P (Fi | s ∈ S) · P (s ∈ S)

(cid:81)k

i=1 P (Fi)

P (s ∈ S | F1, F2, ..Fk) =

(1)
```

Dans l'ensemble intéressant pour ce texte, reprend le texte dans l'ordre sans trop de modifications si ce n'est au niveau des équations

Commande pdf2text pour Gonzalez_2018_Wisebe.pdf :

Titre sur 2 lignes distinctes

```
WiSeBE: Window-Based Sentence

Boundary Evaluation
```

Table restituée avec le titre du premier suivi en dessous du titre de la deuxième partie de la table

PDF

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//

txt

```
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//
```

Sous les deux titres, le texte du premier puis en dessous séparé d'une ligne le texte de la deuxième partie

Table 2 plus complexe un peu moins exploitable

Commande pdf2text pour Torres.pdf :

Du mal à lire les polices d'écriture non standard comme le Text page 1 après introduction ou encore le texte en gras au-dessus de celui-ci

Aucun souci sur le fait que le texte soit scindé en 2 parties pour lire de gauche à droite et de haut en bas

Avantages :

- Le texte est lu qu'il soit scindé ou en bloc
- L'élément : '' informe du changement de page
- L'option -A permet un meilleur rendu du texte en le forçant

Inconvénients :

- Titre souvent sur 2 lignes distinctes
- Tables et formules retranscrite plus ou moins bien
- Accent non retranscrit sur la lettre mais à coté
- Police en gras ou spéciale rend un texte avec plein de saut de lignes pour une phrases