Détection d'explications dans un document

PLAN

Introduction

- ✓ Projet encadré par Nadine Lucas
- ✓ Détection des textes explicatifs
- ✓ Détection d'explications
- ✓ A partir de sources HTML en français ou en anglais

Introduction

- ✓ Qu'est ce qu'une explication ?
- ✓ Le couple « posé / explication »
  - •Le posé est unique et inattendu.
  - •L'explication peut être multiple.
- ✓ Détection des couples :
  - •Critère sur la MFM
  - •Critère sur des marques
  - •Critère sur la position des marques

Contraintes

- ✓ A partir de textes encodés en HTML
- ✓ En français ou en anglais
- ✓ Le balisage peut-être mal formé
- ✓ Le texte en entrée est-il explicatif?
- ✓ Détection des explications par coloriage.

Les lutins

Une mère avait eu son enfant enlevé du berceau par les lutins, qui avaient mis à sa place un petit monstre à grosse tête avec le regard fixe, occupé seulement de boire et de manger. Dans sa détresse, elle alla demander conseil à sa voisine, qui lui dit de porter le petit monstre à la cuisine, de l'installer devant la cheminée et d'allumer le feu pour faire bouillir de l'eau dans deux coquilles d'œuf : " le monstre ne pourra pas s'empêcher de rire, lui dit-elle, et dès l'instant qu'il rit, c'en est fini de lui."

La femme fit tout ce que sa voisine lui avait dit de faire, et Grosse Tête, en la voyant mettre l'eau à bouillir dans des coquilles d'œuf, parla :

Moi qui suis vieux pourtant
Comme les bois de Prusse
Je n'avais jamais vu cuisiner dans un œuf!

Et le voilà qui éclate de rire, et il riait encore quand déjà surgissait toute une foule de lutins qui rapportèrent le véritable enfant, l'installèrent devant le feu et emportèrent avec eux le monstre à grosse tête.

Contraintes

- ✓ A partir de textes encodés en HTML
- ✓ En français ou en anglais
- ✓ Le balisage peut-être mal formé
- ✓ Le texte en entrée est-il explicatif?
- ✓ Détection des explications par coloriage.

Etude de corpus

- ✓ Se familiariser avec les différents textes.
- ✓ Mise en forme matérielle : un rôle importan
- ✓ Etude du code source.
- ✓ Structure d'un texte explicatif.

***Les sables* MOUVANTS**
David Pouilloux

En bord de mer, sur les rives d'un fleuve ou près d'un marécage: les sables mouvants sont des PIEGES MORTELS. Explication de leur APPETIT.

La mort jaune rôde. Moult explorateurs, soldats, scientifiques, touristes et autres aventuriers pourraient en témoigner. S'ils n'avaient été engloutis. Les sables mouvants existent. Où ça? Quasiment partout. La planète n'en est certes pas couverte comme la lune de cratères. Mais les sables avaleurs sont légions. De la France à la Chine, de la Finlande au Cameroun. Qu'importe le climat (tempéré, continental, polaire ou tropical) pourvu qu'on ait les ingrédients de base: du sable et de l'eau. Néanmoins, vous avez pu le constater sur les plages, tout sable humide ne se goinfre pas de baigneurs. Car pour faire un bon sable mouvant, il faut des conditions bien spéciales.

Dans les années cinquante, le professeur Ernest Rice Smith, un géologue américain, prit sa pelle et son seau et remplit ce dernier d'une bonne louche de sables mouvants. Ses conclusions: ni la forme des grains, ni la présence de vase ne sont responsables du phénomène, tout est question d'eau. Et l'important, ce n'est pas que le sable soit humide — on peut rouler avec un 32 tonnes sur la majorité des plages sans risquer l'engloutissement —, mais c'est la façon dont l'eau mouille les grains.

Détection d'explications dans un document

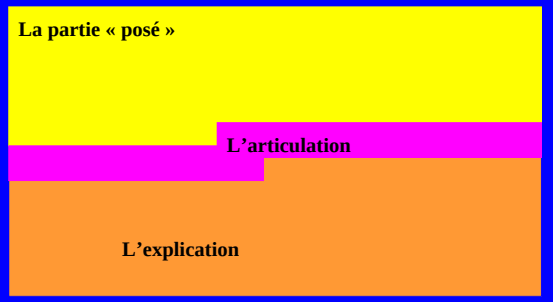
Etude de corpus

- ✓ Se familiariser avec les différents textes.
- ✓ Mise en forme matérielle : un rôle importan
- ✓ Etude du code source.
- ✓ Structure d'un texte explicatif.

Détection d'explications dans un document

Etude de corpus

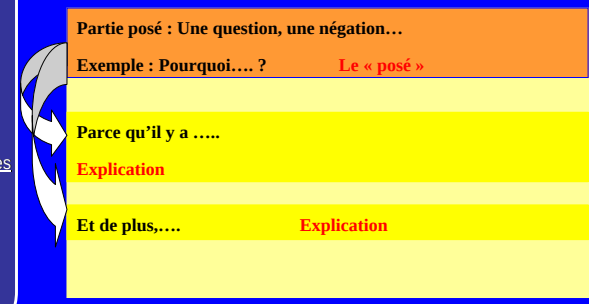
- ✓ Structure binaire :



Détection d'explications dans un document

Etude de corpus

- ✓ Structure d'enchâssement :



Repérage des parties

✓ Basé sur la mise en forme matérielle (MFM)

✓ Nous donne le premier niveau de segmentation du texte.

✓ Essentiel pour détecter les explications de premier niveau.

✓ L'algorithme doit accepter le plus grand nombre de sources HTML.

✓ Nous allons donc typer chaque partie.

***Les sables* MOUVANTS**
David Pouilloux

En bord de mer, sur les rives d'un fleuve ou près d'un marécage: les sables mouvants sont des PIEGES MORTELS. Explication de leur APPETIT.

La mort jaune rôde. Moult explorateurs, soldats, scientifiques, touristes et autres aventuriers pourraient en témoigner. S'ils n'avaient été engloutis. Les sables mouvants existent. Où ça? Quasiment partout. La planète n'en est certes pas couverte comme la lune de cratères. Mais les sables avaleurs sont légions. De la France à la Chine, de la Finlande au Cameroun. Qu'importe le climat (tempéré, continental, polaire ou tropical) pourvu qu'on ait les ingrédients de base: du sable et de l'eau. Néanmoins, vous avez pu le constater sur les plages, tout sable humide ne se goinfre pas de baigneurs. Car pour faire un bon sable mouvant, il faut des conditions bien spéciales.

Dans les années cinquante, le professeur Ernest Rice Smith, un géologue américain, prit sa pelle et son seau et remplit ce dernier d'une bonne louche de sables mouvants. Ses conclusions: ni la forme des grains, ni la présence de vase ne sont responsables du phénomène, tout est question d'eau. Et l'important, ce n'est pas que le sable soit humide — on peut rouler avec un 32 tonnes sur la majorité des plages sans risquer l'engloutissement —, mais c'est la façon dont l'eau mouille les grains.

Sur le balisage HTML

✓ Basé sur l'unique et le multiple.

✓ Travail à « gros grain »

Ex : <P>texte 1<I>texte 2 <B>texte3</B></I></P>

Ici on ne s'occupe que de la balise <P>, les balises <I><B> seront prises en compte plus tard.

✓ Beaucoup moins sensible au HTML « mal formé »

✓ Mais une balise apparaissant plusieurs fois peut être marqueur d'une forme spéciale en fonction de sa position.

Position des balises

✓ Un nouveau critère efficace.

✓ Deux balises identiques a des positions éloignées sont des marques spéciales.

✓ Le critère d'éloignement se base sur la répartition de l'ensemble des balises.

Critère de longueur du texte

✓ Un texte court est généralement un marqueur.

✓ Qui définit une partie spéciale.

✓ La longueur de référence se base sur la moyenne des longueurs des parties du texte.

Engendrées par le vent, les vagues qui se développent dans la zone où le vent souffle et qui peuvent se propager hors de cette zone, sont appelées ondes de gravité. Elles possèdent des caractéristiques spatiales et temporelles bien précises. Les paramètres d'une vague de forme sinusoïdale, forme épurée et théorique, sont la hauteur entre la crête et le creux, et la longueur d'onde, à savoir la distance qui sépare deux crêtes successives. La grandeur caractéristique du temps est la période : placez-vous en un point fixe par rapport au fond, vous voyez défiler chaque vague. Le temps qui sépare le passage de deux crêtes successives est la période de la vague.

Brisant de récif, à Hawaï.

…

Le Corps du document

✓ C'est dans cette partie que nous allons effectuer le repérage des explications.

✓ Le corps se situe entre deux séries de marques spéciales.

✓ Mais à l'intérieur du corps se trouvent aussi des marques spéciales nécessaires à la détection de nos explications.

FSV12
Le journal du CNRS, septembre 2001, p. 25

géodynamique

Le point chaud de l'Afar sous surveillance

Près de 90% des volcans naissent en bordure des plaques tectoniques, au niveau des dorsales et des plaques de subduction. Mais il existe un deuxième type de volcanisme, beaucoup moins répandu, dont l'origine ne semble pas être liée aux mouvements tectoniques : le volcanisme de point chaud. *" Certains volcans apparaissent au milieu des plaques lithosphériques et résultent de la remontée rapide de matière chaude provenant des profondeurs du manteau,* explique Jean-Paul Montagner, directeur du Département de sismologie de l'Institut de physique du globe de Paris (IPGP). *Ces panaches mantelliques percent la croûte terrestre et à mesure du défilement des plaques au-dessus du point chaud, se forment des chapelets d'îles volcaniques parfaitement alignées (Hawaï, La Réunion…) "* . Mais comment et à quelle profondeur naissent-ils? Parviennent-ils tous en surface? Quelle est leur structure intime? Pour répondre à ces questions, un programme d'étude géophysique coordonné par Michel Cara, directeur de l'École et observatoire des sciences de la terre (Éost) de Strasbourg a été mis en place. Deux équipes de l'IPGP et de l' Éost se sont ainsi rendues au Yémen et en Ethiopie, régions où se trouve l'un des rares points chauds émergés. Organisée dans le cadre du programme " Corne de l'Afrique " de l'Insu, leur mission avait pour but de densifier le réseau de sismomètres large bande afin " d'échographier « le globe en profondeur. *" Au lieu d'utiliser les ultrasons, nous nous servons des ondes sismiques pour imager les points chauds,* explique Jean-Paul Montagner. *Ces ondes se propagent plus lentement dans les milieux chauds. En repérant les anomalies de vitesse, nous pouvons ainsi cartographier les panaches mantelliques en 3 dimensions. "* Pendant une semaine, les chercheurs parisiens ont sillonné le Yémen à la recherche de zones épargnées par le " bruit culturel " (les vibrations produites par l'activité humaine). C'est finalement au nord d'Aden qu'une nouvelle station a été mise en place, venant enrichir le dispositif de surveillance déjà installé dans l'année écoulée — une station au Yémen, et trois sur la rive éthiopienne de la Mer Rouge. *" Nous attendons à présent que les données s'accumulent,* explique le chercheur. *Fin 2001, nous devrions être en mesure de fournir une image détaillée du sous-sol de la corne africaine."*

Jacques Gozzo
Contact: **Jean-Paul Montagner**.
Département de sismologie IPGP, UMR 7580, Paris.
Tél.: 01 44 27 48 95.
jpm@ipgp.jussieu.fr
[texte sur 3 colonnes avec une illustration carrée
Légende de la photo]

La segmentation

✓ Nous permettra de détecter des explications locales.

✓Segmentation jusqu'au virgulot.

✓ A l'aide d'expressions régulières.

✓Stockage de l'information dans des tables

Niveau_n
Id
Pere
Pos
Longueur
Texte

Des marques spécifiques

✓ Qui vont permettre de marquer le texte : négation, interrogation, phrase incomplète...

Ex : Pourquoi a-t-on .... ? Texte d'explication...

✓ Des marques non valables pour toutes les unités typographiques.

✓Exemple : Quel est le pourcentage de.... ?

La recherche de l'interrogation ne se fera pas au niveau virgulot.

Extrait de la table contenant les marques à repérer :

Id	Description	Expression régulière	type
16	Voilà comment	(Voilà\scomment)	phrase explicative
17	Qui ... ?	(Qui(.*?)?(\?))	phrase interrogative
18	n' ... pas	\sn\'(.*?)?pas\s	proposition négative
19	Comment... ?	(Comment(.*?)?(\?))	phrase interrogative
22	Quel...?	(Quel(.*?)?(\?))	phrase interrogative
21	Pourquoi...	(Pourquoi(.*?)?(\?))	phrase interrogative

Des marques spécifiques

✓ Qui vont permettre de marquer le texte : négation, interrogation, phrase incomplète...

Ex : Pourquoi a-t-on .... ? Texte d'explication...

✓ Des marques non valables pour toutes les unités typographiques.

✓Exemple : Quel est le pourcentage de.... ?

La recherche de l'interrogation ne se fera pas au niveau virgulot.

Procédure en « appel écho »

✓ On souhaite relier les marques entre elles.

✓ Ce qui va nous permettre de trouver les parties explicatives et les parties posées.

✓ On a donc des relations possibles entre deux éléments distants.

Extrait de la table relation :

Appel, A	Echo, E
phrase définitoire	phrase conclusive
phrase interrogative	phrase explicative
phrase négative	phrase conclusive
proposition négative	phrase conclusive
proposition définitoire	proposition conclusive

Travail sur le titre

✓ Il nous renseigne sur la valeur explicative du texte :

Si la partie du titre possède un caractère interrogatif ou négatif et que le dernier segment du corps est résultatif alors le corps est une explication pour la partie titre.

Si le titre est neutre et que le dernier segment est une subordonnée résultative alors le corps est une explication pour la partie titre.

Travail sur le titre

✓ Il nous renseigne sur la structure du texte :

- Si la partie du titre possède un caractère interrogatif ou négatif et que le dernier segment du corps est résultatif alors la structure du texte est une structure d'enchâssement.
- Si le titre est neutre et que le dernier segment est une subordonnée résultative alors la structure du texte est une structure binaire articulée.

Travail sur le titre

✓ Il nous aide à délimiter le posé de premier niveau :

- On regarde le nombre de parties que comporte la section « titre »
- Valable si la partie « titre » comporte au moins 2 éléments.

La mécanique des vagues

Des lames tubulaires d'Hawaï à la puissante barre de Grand Bassam en Côte d'Ivoire, les vagues semblent échapper à toute description figée. Elles obéissent pourtant à un cycle de vie identifiable.

Par Loys Schmied

Qui n'a pas été fasciné par le spectacle des vagues, se faisant et se défaisant, répétitif, parfois violent et chaotique ? L'image symbolique des vagues est la passivité, passivité de la houle qui festonne le fluide sans un souffle de vent, passivité dans la violence, la puissance, l'inertie qui se communique à la masse fluide. Une vague est une onde qui se propage à la surface de l'eau. Comme toutes les ondes, les vagues transportent de l'énergie, mais ne transportent pas de matière. La masse d'eau, mise en mouvement au passage d'une vague près du rivage reste près du rivage. De même, la masse d'eau, mise en mouvement au large par cette même onde, reste au large. L'image du drap que l'on secoue de haut en bas en est une bonne illustration : si vous secouez fortement un drap à une extrémité, vous engendrez des ondulations qui se propagent à l'autre extrémité. Vous pouvez suivre la déformation du tissu qui se propage d'un bout l'autre du drap, mais la matière, le coton du drap, n'est pas deplace.

Mais lorsqu'une vague déferle, n'y-a-t-il pas transport d'eau ? Certes, mais il s'agit de déplacements locaux dont les dimensions sont sans commune mesure avec la distance de propagation d'une onde, qui peut atteindre plusieurs centaines de kilomètres.

Engendrées par le vent, les vagues qui se développent dans la zone où le vent souffle et qui peuvent se propager hors de cette zone, sont appelées ondes de gravité. Elles possèdent des caractéristiques spatiales et temporelles bien précises. Les paramètres d'une vague de forme sinusoïdale, forme épurée et théorique, sont la hauteur entre la crête et le creux, et la longueur d'onde, à savoir la distance qui sépare deux crêtes successives. La grandeur caractéristique du temps est la période : placez-vous en un point fixe par rapport au fond, vous voyez défiler chaque vague. Le temps qui sépare le passage de deux crêtes successives est la période de la vague.

Brisant de récif, à Hawaï.
Sur la côte nord de Hawaï, du fait de l'absence de plateau continental, les lames heurtent les récifs de l'île avec la puissance quasi intacte du plein océan.
Anatomie d'une vague.
Une lame est caractérisée par sa longueur d'onde, la distance horizontale séparant deux crêtes ou deux creux successifs, sa hauteur, la distance verticale entre le sommer de la crête et la base du creux, et sa période, le temps mis par une crête pour parcourir une longueur d'onde.
...
...

Algorithme

✓ Basé sur les procédures en « appel-echo » à partir des marques et de leurs relations.

✓ Le titre permet de nous guider

1) Si le titre et le corps ne sont pas marqués au premier niveau alors je dis que le texte n'est pas explicatif.
2) Si le titre est marqué et que le dernier segment de texte l'est alors je colorie tout le corps de texte.
3) Si le titre est neutre et que le corps est marqué alors je cherche un couple «posé/explication » dans le corps du document.
4) J'utilise la table de relations pour trouver des couples. Je descends d'une unité typographique (n-1) quand j'ai exploré celle de niveau n.

Difficultés rencontrées

✓ Le HTML mal formé.

✓ Qui ne permet pas de mettre en place tous les algorithmes pensés.

✓Des segmentations difficiles.

✓Détection de fin de corps de texte

Conclusion

✓ Le travail des linguistes est délicat.

✓Car on observe une multitude de règles

✓Qui peuvent donner lieu à de nouvelles règles.

✓Mais il faut savoir faire des algorithmes les plus souples possibles.