VOILADIS : des relations lexicales aux structures de discours

Clémentine Adam, Cécile Fabre, Philippe Muller
CLLE/IRIT

16 Décembre 2008

PLAN

- Introduction - Objectifs du projet VOILADIS
- La cohésion lexicale
- La segmentation thématique
- Premières expériences
- Perspectives

VOILADIS - Objectifs du projet

- Contexte :
  - VOILADIS = VOIsinage Lexical pour l'Analyse du DIScours
  - Projet PRES - UTM (CLLE-ERSS) / UT3 (IRIT-LILaC)
  - Thèse Clémentine Adam
- Objectifs
  - S'appuyer sur des indices lexicaux pour identifier la structure des discours
  - Démarche complémentaire à celle poursuivie dans le projet Annodis :
    - Au niveau global : indicateurs lexicaux de continuité/rupture
    - Au niveau local : corrolaire sur le plan lexical de certaines relations de discours
  - Exploiter les ressources lexicales construites à partir de corpus (Upery)

VOILADIS - Objectifs du projet

Plan de réalisation annoncé :

1. mise en place d'un environnement permettant de projeter sur un corpus diversifié les résultats du programme Upery (voisins distributionnels)
2. mise au point d'une méthode permettant de s'appuyer sur cette annotation lexicale pour dégager des zones de cohésion lexicale et/ou des segments textuels proches associés par la présence de voisins distributionnels dans chacun d'eux
3. caractérisation de l'apport de la méthode sur le plan de l'organisation discursive.

Cohésion discursive - rappel

- procédés qui permettent de relier des segments les uns aux autres (cohésion lexicale, coréférence, ellipses, connecteurs...)
- un des indicateurs de surface de la structure du discours "Coherence defines the macro-level semantic structure of a connected discourse, while cohesion creates connectedness in a non-structural manner." [Barzilay and Elhadad1997]

Cohésion Lexicale

- "the cohesive effect achieved by the continuity of lexical meaning" [Halliday and Hasan1976]
- "the dominant mode of creating texture" [Hoey1991]

"Any structural theory of text must be concerned with identifying units of text that are about the same thing. When a unit of text is about the same thing there is a strong tendency for semantically related words to be used within that unit. By definition, lexical chains are chains of semantically related words. Therefore it makes sense to use them as clues to the structure of the text." [Morris and Hirst1991]

Procédés de cohésion lexicale

Synthèse de [Tanskanen2006]

- [Halliday and Hasan1976]
  - **Réitération** : répétition, recours à un synonyme, un hyperonyme
  - **Collocation** : "cohesion that is achieved through the association of lexical items that regularly co-occur"
- [Stubbs2001]
  - **Relations de réitération**
    - répétition simple
    - répétition complexe (proximité morphologique, proximité inter-catégorielle)
    - équivalence (synonymie dans le discours)
  - **Relations de collocation**
    - ensembles (couleurs, mois...)
    - collocations de type activité (*repas*/*manger*, *voiture*/*conduire*)
    - collocations élaboratives (frames - *Cambridge* => *University frame*)

Exemple

repetition
ordered sed
equivalence

En juillet 1961, **Cuba** signifie son appartenance au "bloc socialiste". Le 4 septembre 1962, le pays conclut un accord d'assistance militaire avec l'Union soviétique et, une semaine plus tard, Moscou déclare que toute attaque contre **Cuba** provoquerait une riposte nucléaire. Le Congrès américain pour sa part vote le 3 octobre une résolution qui met en demeure contre toute "action subversive dans l'hémisphère occidental". Kennedy interdit cependant l'opération Northwoods mise au point et proposée par l'état-major, laquelle prévoyait d'orchestrer une série d'attentats contre les États-Unis, puis d'en accuser **Cuba** afin de mobiliser l'opinion publique contre Castro.

Techniques de chaînage lexical

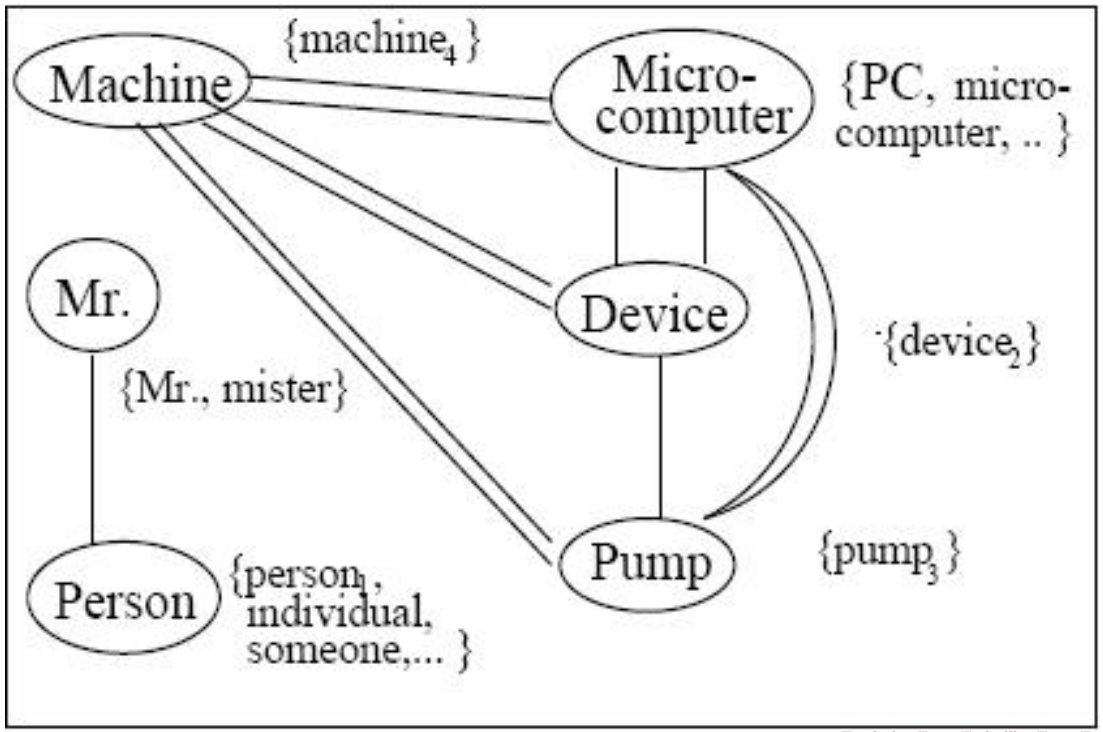
[Stairmand1996], [Hirst and St-Onge1998] ...
Exemple avec WordNet :

- Pour le terme i du texte, on génère l'ensemble des mots trouvés dans ses synsets
- Pour chaque autre terme du texte, s'il appartient à cet ensemble, on l'ajoute à la chaîne lexicale du terme i
- On répète le processus pour tous les termes
- On stocke les chaînes de plus de 3 termes

[Barzilay and Elhadad1997] : calcul d'un score pour les chaînes concurrentes générées, en fonction du nombre et du poids des relations qui les composent => désambiguïsation

Chaînage lexical - exemple [Barzilay and Elhadad1997]

Mr. Kenny is the **person** that invented an anesthetic **machine** which uses **micro-computers** to control the rate at which an anesthetic is pumped into the blood. Such **machines** are nothing new. But his **device** uses two **micro-computers** to achieve much closer monitoring of the **pump** feeding the anesthetic into the patient.



Pour quoi faire

Calcul de chaînes lexicales pour :

- Pondération des termes dans une tâche de RI/QR
- Résumé automatique (*lexical chain-based summarizers*)
  Chaines lexicales = étape intermédiaire pour la production de résumé
- Détection de thèmes (*topic tracking and detection*)
  On calcule des recouvrements entre chaînes pour repérer la récurrence d'un thème.
- Détection de mots erronés (*malapropisms*)
- Génération de liens hypertextes (*chaining across texts*)

cf. 2005 ELECTRA Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (Beyond Bag of Words)

Source d'information pour estimer le lien entre mots

- Statistiques de co-occurrence
- Ressources
  "The chains can be built using any lexical resource that relates word semantically" [Nicola Stokes *et al.*2002]
  En particulier :
  - Roget's Thesaurus [Morris and Hirst1991]
  - WordNet

  WordNet comme source de connaissance pour calculer les chaînes lexicales : [Hirst and St-Onge1998]
  - lexical repetitions = extra-strong relations
  - synonyms, near-synonyms, hypernyms = strong relations (1 relation directe)
  - autres parcours spécifiques dans l'arbre = *regular relations*

Exploiter des relations "non-classiques"

- Dépasser les relations lexicales traditionnelles [Tanskanen2006], [Morris and Hirst2004]
- Tenir compte des relations qu'instaure le discours lui-même ([Mortureux1993])

"In recent lexical cohesion research in linguistics (...) **non-classical relations** are largely ignored, and the same is true in implementations of lexical cohesion in computational linguistics (...), as the lexical resource used is WordNet. It is notable, however, that the original view of lexical semantic relations in the lexical cohesion work of Halliday and Hasan (1976) was very broad and general ; the only criterion was that there had to be a recognizable relation between two words." [Morris and Hirst2004]

Les voisins distributionnels

- Ressource permettant d'exploiter des relations sémantiques "non-classiques"
- Principe d'acquisition des couples de voisins :
  - Construction de couples (prédicat, argument) sur la base de l'analyse en dépendance (en sortie de Syntex)
  - Un triplet syntaxique (gouverneur , relation , dépendant) fournit un couple (prédicat, argument)
    prédicat = gouverneur_relation
    argument = dépendant

Distinction arguments et prédicats

Rapprocher des prédicats qui partagent les mêmes arguments ET rapprocher des arguments qui partagent les mêmes prédicats

[attaque, invasion]	[attaque_de, combattre_obj]
imminence_de	iroquois
redouter_obj	sarrasins
vulnérable_à	envahisseur
protéger_de	pirate
repousser_obj	monstre
détruire_lors de	ennemis
...	...

Voisins en tant qu'argument

Catégorie	Lemme	Relation	Nb cooccurrents	Catégorie	Lemme	Relation	Nb cooccurrents	a	Prox Lin
N	attaque	_	505	N	opération	_	493	212	0.379
N	attaque	_	505	N	invasion	_	196	130	0.347
N	attaque	_	505	N	action	_	952	280	0.342
N	attaque	_	505	N	tentative	_	190	130	0.334
N	attaque	_	505	N	coup	_	494	191	0.328
N	attaque	_	505	N	combat	_	628	202	0.309
N	attaque	_	505	N	prise	_	343	154	0.305
N	attaque	_	505	N	traitement	_	367	155	0.301
N	attaque	_	505	N	intervention	_	258	132	0.294
N	attaque	_	505	N	changement	_	398	152	0.289
N	attaque	_	505	N	construction	_	694	198	0.276
N	attaque	_	505	N	formation	_	668	188	0.267
N	attaque	_	505	N	expérience	_	421	150	0.266
N	attaque	_	505	N	publication	_	309	128	0.264
N	attaque	_	505	N	pratique	_	440	148	0.26
N	attaque	_	505	N	programme	_	680	183	0.258
N	attaque	_	505	N	mesure	_	502	157	0.257
N	attaque	_	505	N	développement	_	827	206	0.256
N	attaque	_	505	N	campagne	_	468	142	0.252
N	attaque	_	505	N	effet	_	507	153	0.252
N	attaque	_	505	N	crise	_	276	113	0.252
N	attaque	_	505	N	maladie	_	412	133	0.251
N	attaque	_	505	N	recherche	_	778	190	0.251
N	attaque	_	505	N	activité	_	885	210	0.25
N	attaque	_	505	N	attentat	_	131	86	0.25
N	attaque	_	505	N	étude	_	765	189	0.246

Voisins en tant que prédicat

Catégorie	Lemme	Relation	Nb cooccurrents	Catégorie	Lemme	Relation	Nb cooccurrents	a	Prox Lin
N	attaque	de	196	V	attaquer	suj	155	76	0.39
N	attaque	de	196	V	attaquer	obj	274	95	0.344
N	attaque	de	196	V	contrôler	suj	122	49	0.24
N	attaque	de	196	V	occuper	suj	457	94	0.236
N	attaque	de	196	V	vaincre	obj	97	42	0.235
N	attaque	de	196	V	perdre	suj	231	62	0.233
N	attaque	de	196	V	combattre	obj	207	56	0.23
N	attaque	de	196	V	détruire	suj	179	52	0.228
N	attaque	de	196	V	se emparer	suj	99	41	0.228
N	attaque	de	196	N	guerre	contre	119	41	0.223
N	attaque	de	196	V	garder	suj	116	44	0.21
N	attaque	de	196	V	tirer	suj	127	44	0.208
N	attaque	de	196	V	entrer	suj	161	48	0.207
N	attaque	de	196	V	chasser	obj	145	41	0.202
N	attaque	de	196	V	subir	suj	133	44	0.201
N	attaque	de	196	N	arrivée	de	228	54	0.199
N	attaque	de	196	N	main	de	217	49	0.197
N	attaque	de	196	V	envahir	suj	98	36	0.196
N	attaque	de	196	V	disposer	suj	251	57	0.194
N	attaque	de	196	V	soutenir	suj	316	62	0.193
N	attaque	de	196	V	conserver	suj	177	48	0.192
N	attaque	de	196	V	conquérir	suj	71	31	0.19
N	attaque	de	196	V	tuer	suj	173	41	0.185
N	attaque	de	196	V	abandonner	suj	99	36	0.184
N	attaque	de	196	V	réussir	suj	127	39	0.183
N	attaque	de	196	N	attaque	contre	36	27	0.177

Les voisins distributionnels

- Voisinage sémantique calculé sur la base des contextes partagés => rapprochements discursifs
- Rapprochement de prédicats => relations intercatégorielles
  Ex : *octroi de* / *concéder obj* - *résister à* / *protection contre*
- Bien au-delà des relations canoniques qui structurent le lexique
  Expérience de comparaison avec dicosyn (Galy et Bourigault)
  20% de dicosyn dans les voisins / 5% des voisins dans dicosyn

Projection des voisins pour le repérage de chaînes - exemple

En juillet 1961, Cuba signifie son appartenance au “bloc socialiste”. Le 4 septembre 1962, le pays conclut un accord d’assistance militaire avec l’Union soviétique et, une semaine plus tard, Moscou déclare que toute attaque contre Cuba provoquerait une riposte nucléaire. Le Congrès américain pour sa part vote le 3 octobre une résolution qui met en demeure contre toute ” action subversive dans l’hémisphère occidental ”. Kennedy interdit cependant l’opération Northwoods mise au point et proposée par l’état-major, laquelle prévoyait d’orchestrer une série d’attentats contre les États-Unis, puis d’en accuser Cuba afin de mobiliser l’opinion publique contre Castro.

Amorcer le repérage de relations

- Repérer des relations spécifiques
- Combiner parenté distributionnelle (*in absentia*) et cooccurrence (*in praesentia*)
- Caractériser contextuellement la relation de voisinage distributionnel

De l'antonymie à la relation d'opposition

- *Ancillary Antonymy* [Jones2002]
  "The antonymous pair (...) contributes to a larger contrast ; that the antonyms themselves are not the primary contrast of the sentence, but are actually responsible for signalling a more important opposition (usually instantial) between another pair of words, phrases or clauses." (p.45)
  "Charles, unskilfully, is playing for the popular vote ; Diana, very skilfully, is doing the same."
  "corpus evidence suggests that [the adversative conjunction] is perhaps the most dispensable contrast-generating device." (p.60)
- Combiner parenté distributionnelle et marqueurs de relation : Mémoire de François Morlane-Hondère sur l'antonymie

De la synonymie à la relation d'élaboration

- Expérience sur les couples V-N [Fabre and Bourigault2006] Sélection des couples de voisins V-N qui apparaissent au moins 1 fois au sein d'1 même paragraphe avec le même dépendant (1441 couples sur corpus LM10)

(...) un éventuel raccourcissement du mandat présidentiel. (...) c'était à lui seul d'apprécier s'il devait volontairement écourter son mandat

Les Finlandais et les Suédois sont en effet de fervents adeptes de la marche à pied. [...] Dans chacun de ces deux pays, 76 % et 74 % des plus de 55 ans, notamment, pratiquent la marche au moins une fois tous les quinze jours, contre 40 % en moyenne dans l'Union.

- Travail entamé par Edith Galy : outil de visualisation de la cooccurrence des voisins

References

R. Barzilay and M. Elhadad.
Using lexical chains for text summarization.
In *ACL'97/EACL'97 Workshop on Interlligent Scalable Text Summarization*, Madrid, 1997.

C. Fabre and D. Bourigault.
Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques.
In *TALN*, pages 121–129, Leuven, 2006.

M.A.K. Halliday and R. Hasan.
Cohesion in English.
Longman, London, 1976.

G. Hirst and D. St-Onge.
WordNet : An Electronic Lexical Database and Some of its Applications. Cambridge, MA : The MIT Press., chapter Lexical chains as representation of context for the detection and correction of malapropisms.