Vers un « niveau discours » en Traitement Automatique des Langues

Marie-Paule Péry-Woodley

Préambule: niveaux de traitement en TAL

- Niveaux morphologique, syntaxique, sémantique, pragmatique
- … et le niveau discursif?
- Quels objets? Quelles questions?
- Quel intérêt pour le TAL ?

Plan de l'exposé

1. Définir un « niveau discours »: bottom-up ou top-down?
2. Une linguistique du discours en corpus et outillée
3. TAL et discours

1. « Niveau discours »

- Discours et document
- De la phrase au texte, de l'énoncé au discours (ou l'inverse ?)
- Analyse ascendante vs descendante
- « Niveau discours »: approches

Discours et document

- “Work gets done through documents. When a negotiation draws to a close, a document is drawn up, an accord, a law, a contract, an agreement. When a new organization is established it is announced with a document. When research culminates, a document is created and published. And knowledge is transmitted through documents: research journals, text books and newspapers. Documents are information organized and presented for human understanding.” (Cole *et al.,* 1998:223).

- TAL 47/2 Discours et document : traitements automatiques http://www.atala.org/rubrique.php3?id_rubrique=46

De la phrase au texte, de l’énoncé au discours (ou l’inverse ?)

- “ Texts are not just simple sequences of sentences but rather complex artifacts that exhibit a sophisticated high-level, discourse/rhetorical organization/structure.” Marcu http://www.isi.edu/~marcu/discourse/

- “A text, as we are interpreting it, is a ***semantic unit***, which is not ***composed*** of sentences but ***realized*** in sentences”. (Halliday, 1977-2003:45-46)

Analyse ascendante

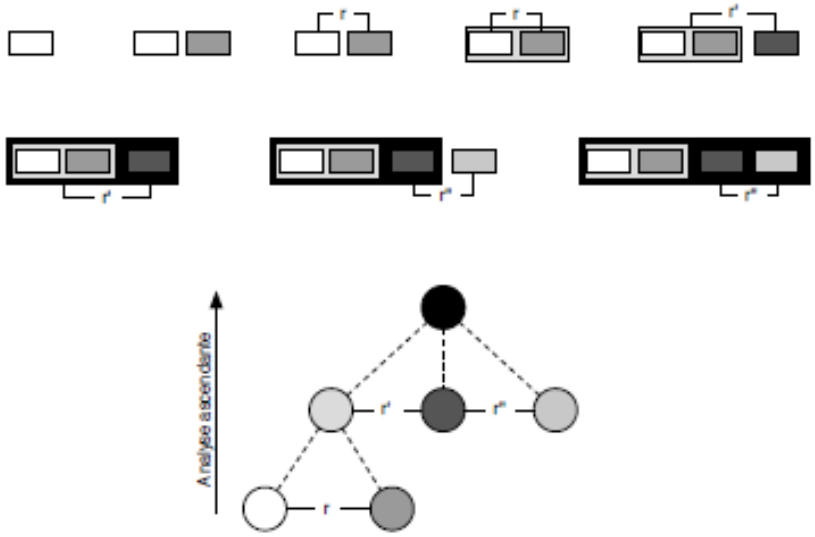


Fig. 5.19 – Analyse ascendante

Widlöcher 2008: 136

Analyse descendante

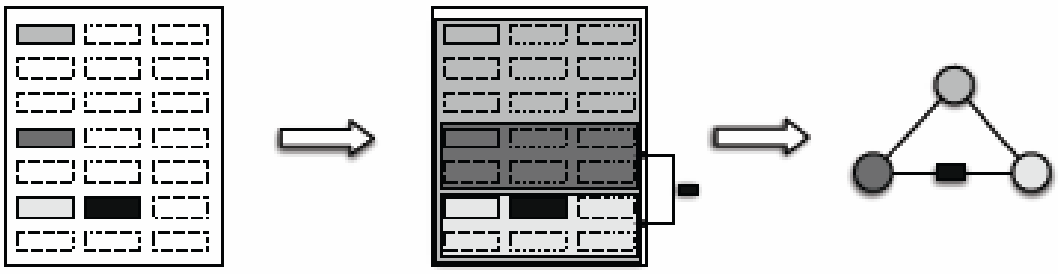


Fig. 5.20 – Analyse descendante et sporadique

Widlöcher 2008: 139

« Niveau discours »: approches

Un exemple: un texte, plusieurs approches de l'organisation discursive

- Approches descendantes
  - Structure de document
  - Cadres
  - Chaînes
  - Structures énumératives
- Approche ascendante
  - Relations de discours

Structure de document

Structure 1: Objets textuels, MFM

Titre 1er niveau

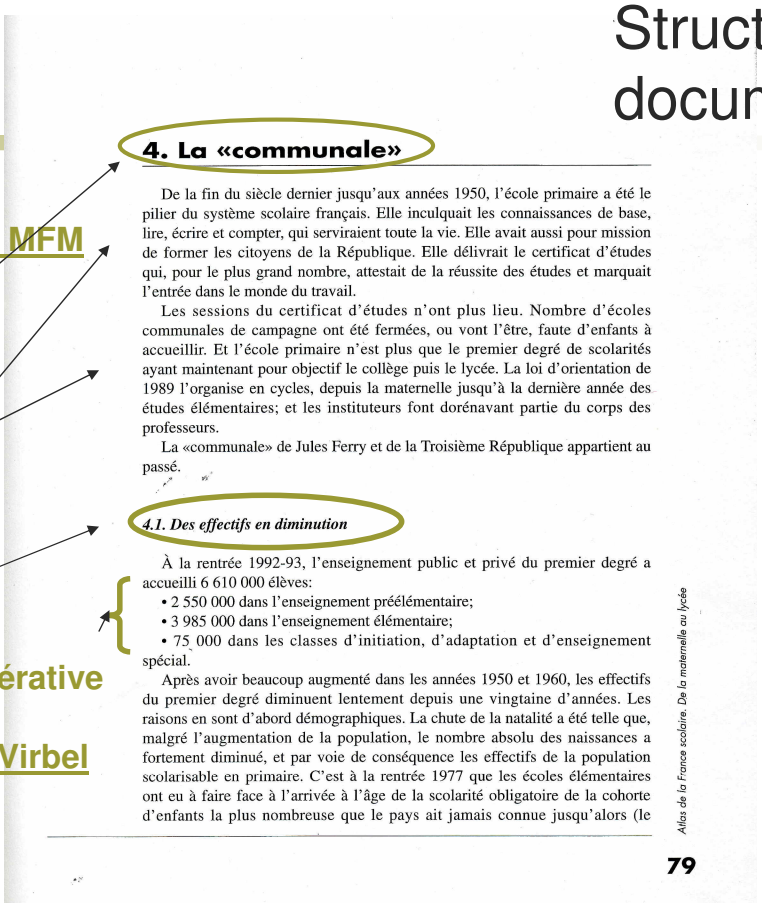
Section

Paragraphes

Titre 2nd niveau

Structure énumérative

Cf. MAT, Luc & Virbel



4. La «communale»

De la fin du siècle dernier jusqu'aux années 1950, l'école primaire a été le pilier du système scolaire français. Elle inculquait les connaissances de base, lire, écrire et compter, qui serviraient toute la vie. Elle avait aussi pour mission de former les citoyens de la République. Elle délivrait le certificat d'études qui, pour le plus grand nombre, attestait de la réussite des études et marquait l'entrée dans le monde du travail.

Les sessions du certificat d'études n'ont plus lieu. Nombre d'écoles communales de campagne ont été fermées, ou vont l'être, faute d'enfants à accueillir. Et l'école primaire n'est plus que le premier degré de scolarités ayant maintenant pour objectif le collège puis le lycée. La loi d'orientation de 1989 l'organise en cycles, depuis la maternelle jusqu'à la dernière année des études élémentaires; et les instituteurs font dorénavant partie du corps des professeurs.

La «communale» de Jules Ferry et de la Troisième République appartient au passé.

4.1. Des effectifs en diminution

À la rentrée 1992-93, l'enseignement public et privé du premier degré a accueilli 6 610 000 élèves:

- 2 550 000 dans l'enseignement préélémentaire;
- 3 985 000 dans l'enseignement élémentaire;
- 75 000 dans les classes d'initiation, d'adaptation et d'enseignement spécial.

Après avoir beaucoup augmenté dans les années 1950 et 1960, les effectifs du premier degré diminuent lentement depuis une vingtaine d'années. Les raisons en sont d'abord démographiques. La chute de la natalité a été telle que, malgré l'augmentation de la population, le nombre absolu des naissances a fortement diminué, et par voie de conséquence les effectifs de la population scolarisable en primaire. C'est à la rentrée 1977 que les écoles élémentaires ont eu à faire face à l'arrivée à l'âge de la scolarité obligatoire de la cohorte d'enfants la plus nombreuse que le pays ait jamais connue jusqu'alors (le

Atlas de la France scolaire. De la maternelle au lycée

79

P 79 de l'Atlas de la France Scolaire,

un extrait de document

Cadres de discours

Structure 2 :

cadre de discours

Cf. Charolles

4. La «communale»

De la fin du siècle dernier jusqu'aux années 1950, l'école primaire a été le pilier du système scolaire français. Elle inculquait les connaissances de base, lire, écrire et compter, qui serviraient toute la vie. Elle avait aussi pour mission de former les citoyens de la République. Elle délivrait le certificat d'études qui, pour le plus grand nombre, attestait de la réussite des études et marquait l'entrée dans le monde du travail.

Les sessions du certificat d'études n'ont plus lieu. Nombre d'écoles communales de campagne ont été fermées, ou vont l'être, faute d'enfants à accueillir. Et l'école primaire n'est plus que le premier degré de scolarités ayant maintenant pour objectif le collège puis le lycée. La loi d'orientation de 1989 l'organise en cycles, depuis la maternelle jusqu'à la dernière année des études élémentaires; et les instituteurs font dorénavant partie du corps des professeurs.

La «communale» de Jules Ferry et de la Troisième République appartient au passé.

4.1. Des effectifs en diminution

Chaînes

Structure 3 :

Chaînes

Cf. Cornish

Cohésion (reference, lexical cohesion)

Cf. Halliday & Hasan

4. La «communale»

De la fin du siècle dernier jusqu'aux années 1950, l'école primaire a été le pilier du système scolaire français. Elle inculquait les connaissances de base, lire, écrire et compter, qui serviraient toute la vie. Elle avait aussi pour mission de former les citoyens de la République. Elle délivrait le certificat d'études qui, pour le plus grand nombre, attestait de la réussite des études et marquait l'entrée dans le monde du travail.

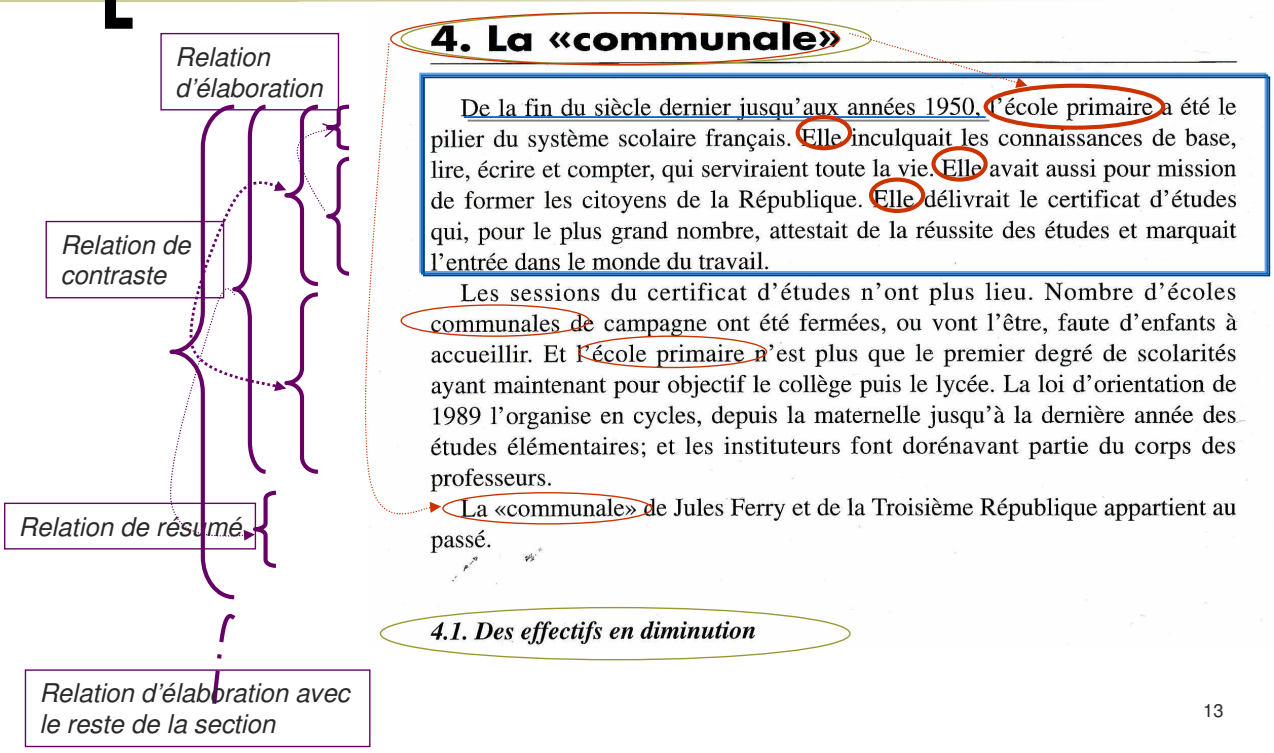
Les sessions du certificat d'études n'ont plus lieu. Nombre d'écoles communales de campagne ont été fermées, ou vont l'être, faute d'enfants à accueillir. Et l'école primaire n'est plus que le premier degré de scolarités ayant maintenant pour objectif le collège puis le lycée. La loi d'orientation de 1989 l'organise en cycles, depuis la maternelle jusqu'à la dernière année des études élémentaires; et les instituteurs font dorénavant partie du corps des professeurs.

La «communale» de Jules Ferry et de la Troisième République appartient au passé.

4.1. Des effectifs en diminution

Relations de discours

Structure 4 : relations de discours (Cf. Rhetorical Structure Theory)



Plan de l'exposé

1. Définir un « niveau discours »: bottom-up ou top-down?
2. Une linguistique du discours en corpus et outillée
3. TAL et discours

2. Une linguistique du discours en corpus

- Principes de base des linguistiques de corpus
- Une linguistique outillée
- Corpus et discours
- Réalisations: corpus annotés

Principes de base des linguistiques de corpus

- Authenticité des données
  La langue telle qu'elle est parlée ou écrite : primauté à l'attesté
- Volume des données
  - observations statistiques
  - phénomènes marginaux *vs* phénomènes massifs
- Diversité des données
  - ouverture à différents types de discours
- Fiabilité des données

Une linguistique outillée

Habert 2005:1 : “multiplication (…) des outils, des instruments et des ressources modifiant les conditions de constitution d’observables et d’analyse de données en sciences du langage”

=> **Linguistique outillée**, dotée d’instruments d’observation et de calcul

Corpus et discours

- Des exigences supplémentaires
  - Textes entiers
  - Mise en forme respectée
- Des outils aux fonctionnalités élargies
  - XML et famille
  - Outils de visualisation
- Importance des corpus annotés

Des corpus annotés discursivement

- Penn Discourse TreeBank: 1808 articles étiquetés du Wall Street Journal extraits du Penn TreeBank
- RST TreeBank: 385 articles, même origine, modèle du discours différent

Annotation de relations du discours dans le PDTB

Les connecteurs et leurs arguments:

- **Connecteur explicite**: connecteur explicitement présent dans le texte, qui établit une relation de discours entre deux arguments.
- **Connecteur implicite**: établit une relation qui n'est pas réalisée explicitement dans le texte, et qui doit être inférée par le lecteur.
  - Dans ce cas, adjonction au nœud correspondant au connecteur d'un trait contenant le connecteur qui semble le mieux exprimer la relation + sa classe sémantique.
- Argument d'un connecteur = expression d'un objet abstrait en relation (implicite ou explicite) avec une autre expression du même type.
  - ➔ cas les plus simples: argument = proposition.
  - ➔ étiquetage des arguments: Arg1 et Arg2.

Annotation de relations de discours dans le PDTB (connecteurs explicites et implicites)

She hasn't played any music **since the earthquake hit.**

Third-quarter sales in Europe were exceptionally strong, boosted by promotional programs and new products – **although weaker foreign currencies reduced the company's earnings.**

The small, wiry Mr. Morishita comes across as an outspoken man of the world. **IMPLICIT=WHEN IMPLICIT=FOR EXAMPLE** (Stretching his arms in his silky white shirt and squeaking his black shoes) **he lectures a visitor about the way to sell American real estate and boasts about his friendship with Margaret Thatcher's son.**

Annotation de l'attribution dans le PDTB

Les attributions

- Qui dit quoi?: « … relation d'appartenance entre des objets abstraits et des individus ou agents ».
  - ➔ Attribuer des croyances/assertions exprimées dans le texte à l'agent (aux agents) qui les exprime(nt).
  - ➔ But: classification de la factualité des objets abstraits associés aux arguments des relations de discours et à la relation elle-même.

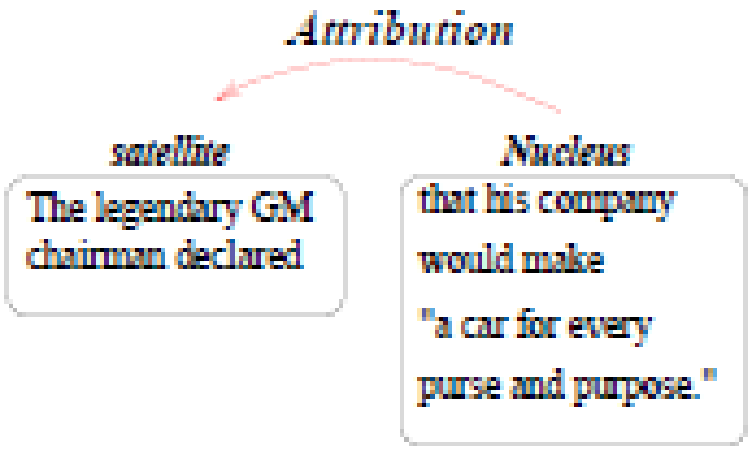
- Les traits de l'attribution
  - Source: rédacteur (**Writer**) vs. autre (**Other**)
  - Factualité: fait (**Fact**) vs. non-fait (**Non-fact**)
  - Polarité: positive (**Positive**) vs. négative (**Negative**)

PDTB: Exemple d'attribution

When **Mr. Green won a $240,000 verdict in a land condemnation case against the State in June 1983 [ARG2]**, he says *Judge O'Kicki unexpectedlyawarded him an additional $100,000 [ARG1].*

	REL	Arg1	Arg2
[Source]	Writer	Other	Inherited

RST Treebank



RST Treebank

- Annotations selon Rhetorical Structure Theory (Mann & Thompson 1988):
  - Relations asymétriques noyau-satellite
  - 2 types de relations
    - Subject matter (cause, condition, élaboration,…)
    - Presentation (antithèse, arrière-plan, motivation,…)
  - Annotation fondée sur marqueurs de surface

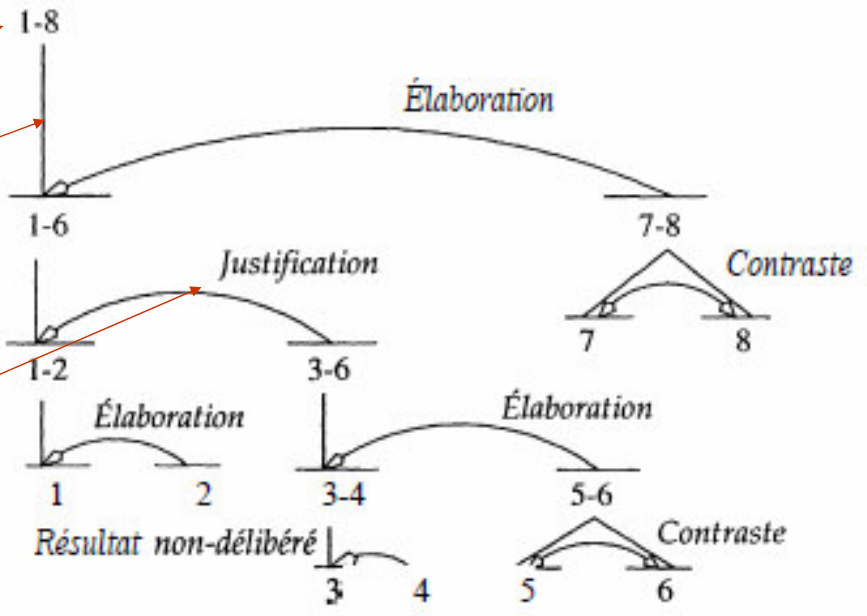
Voir http://www.isi.edu/~marcu/discourse/

Représentation RST d'un texte

Segment de texte

Noyau vs. Satellite

Relation orientée et étiquetée



Annotation RST et résumé: « le temps sur Mars »

[With its distant orbit {–50 percent farther from the sun than Earth –} and slim atmospheric blanket,1] [Mars experiences frigid weather conditions.2] [Surface temperatures typically average about –60 degrees Celsius (–76 degrees Fahrenheit) at the equator and can dip to –123 degrees C near the poles.3] [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,4] [but any liquid water formed that way would evaporate almost instantly5] [because of the low atmospheric pressure.6] [Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,7] [most Martian weather involves blowing dust or carbon dioxide.8] [Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap.9] [Yet even on the summer pole, {where the sun remains in the sky all day long,} temperatures never warm enough to melt frozen water.10]

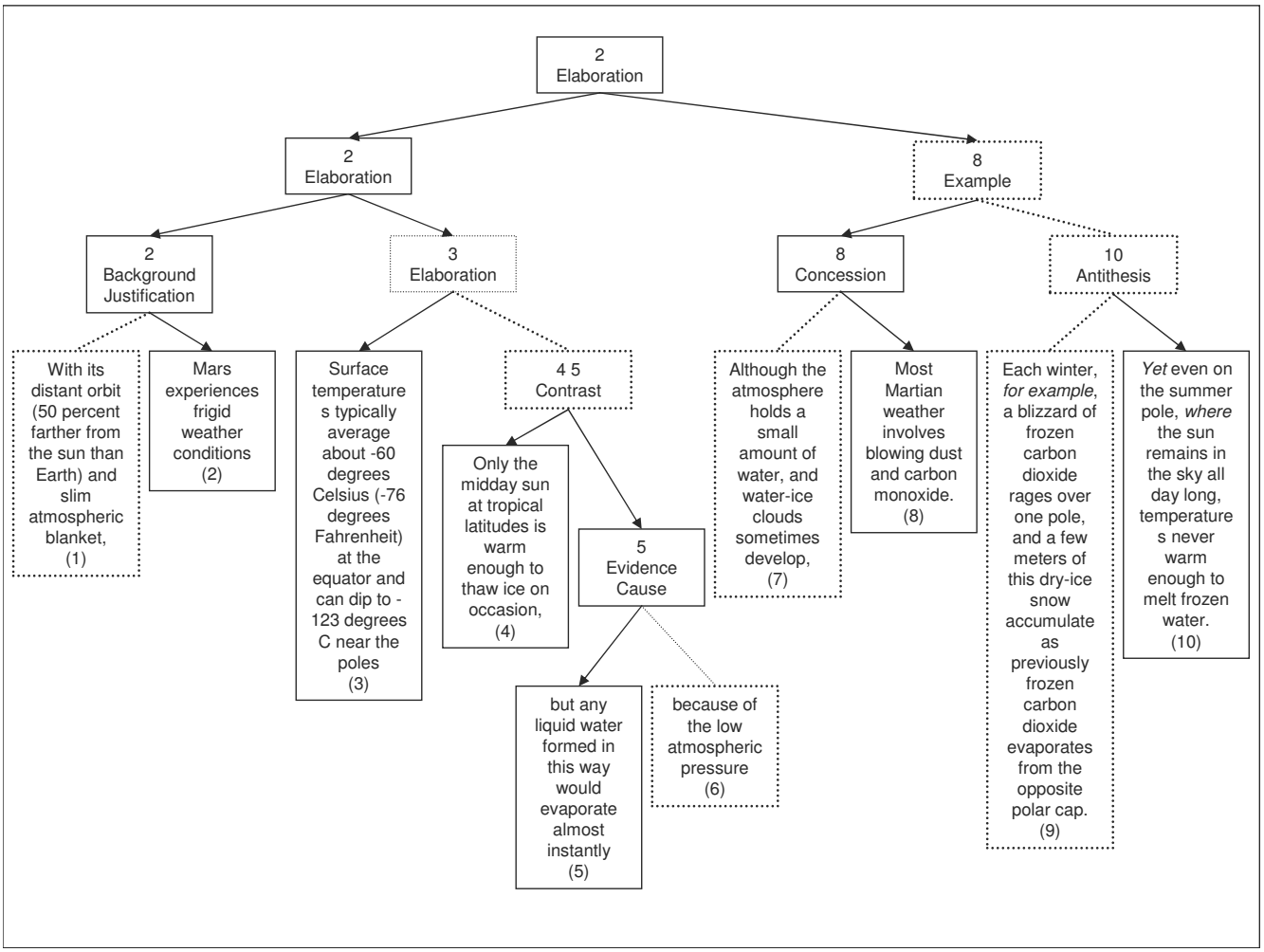
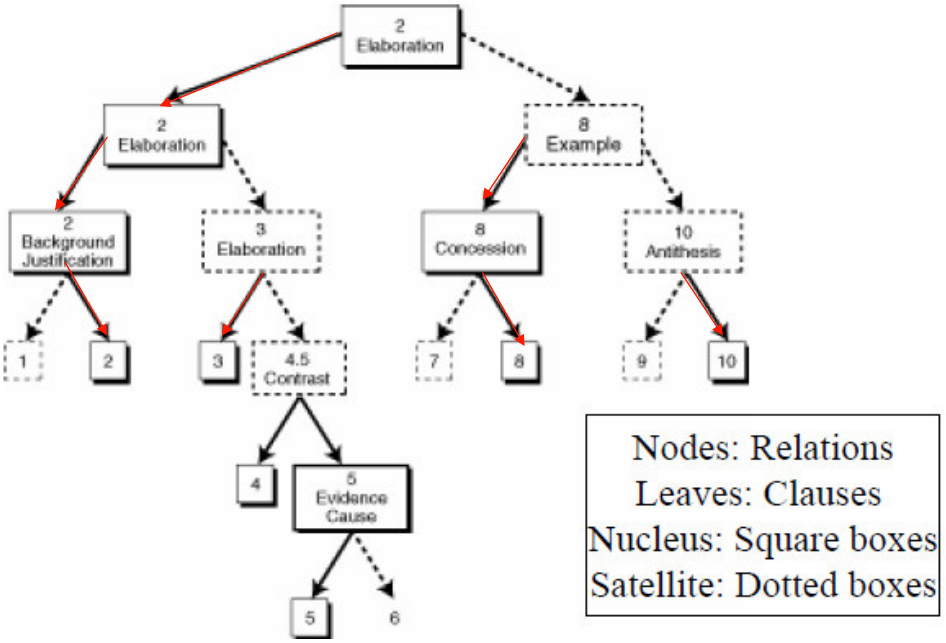


Illustration of Node Promotion (Marcu)



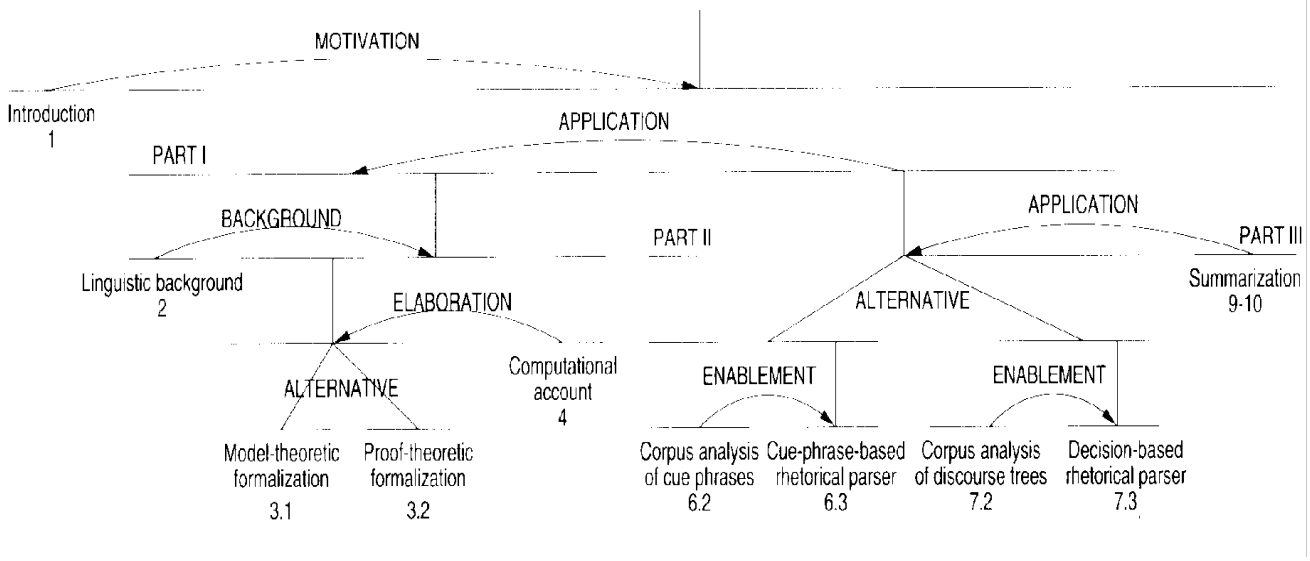
Résumé RST de « le temps sur Mars »:

2 > 8 > {3,10} > {1,4,5,7,9}

[With its distant orbit {–50 percent farther from the sun than Earth –} and slim atmospheric blanket,1] **[Mars experiences frigid weather conditions.2]** **[Surface temperatures typically average about –60 degrees Celsius (–76 degrees Fahrenheit) at the equator and can dip to –123 degrees C near the poles.3]** [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,4] [but any liquid water formed that way would evaporate almost instantly5] [because of the low atmospheric pressure.6] [Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,7] **[most Martian weather involves blowing dust or carbon dioxide.8]** [Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap.9] **[Yet even on the summer pole, {where the sun remains in the sky all day long,} temperatures never warm enough to melt frozen water.10]**

Marcu (2000): *The Theory and Practice of Discourse Parsing and Summarization*

A rhetorical map of the book



Annoter pour quoi faire?

- Entraînement d'analyseurs
- Construction de grammaires
- Matériau de base pour la construction de lexiques
- Evaluation de divers traitements, notamment analyse syntaxique
- Base de données pour la recherche linguistique

Annoter *discursivement* pour quoi faire?

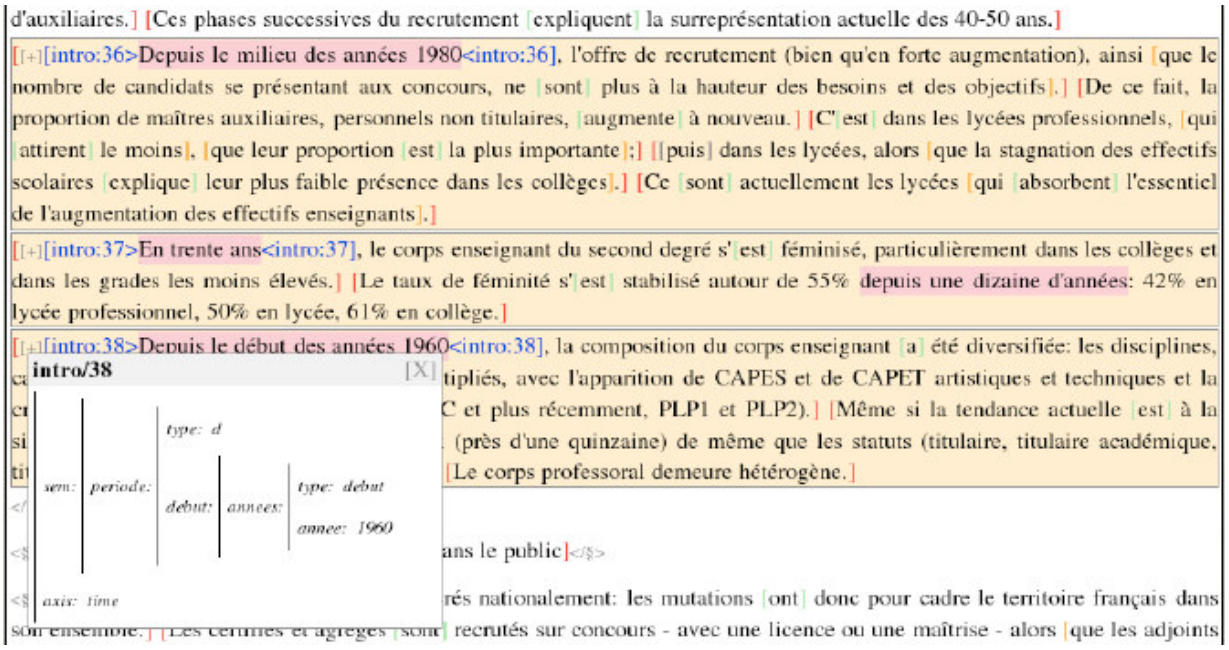
- Indexation en lien avec structures discursives pour EI, QR
- Identification de segments homogènes par rapport à un critère donné (navigation)
- Hiérarchisation de segments; repérage de zones fonctionnelles pour la synthèse automatique
- …

- Annoter pour expérimenter

Segments homogènes par rapport à un critère: cadres de discours

De 1965 à 1985, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible dans le Sud-Ouest et le Massif central, modérée en Bretagne et à Paris, l'augmentation a été considérable dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne où les effectifs ont souvent plus que doublé. Les variations de la population et les baisses plus ou moins fortes du nombre des naissances selon les régions **ne suffisent pas** à expliquer ces différences d'accroissement des effectifs du secondaire. Intervient aussi l'allongement des scolarités, qui a été plus marqué dans les départements où, **au milieu des années 1960**, la poursuite des études après l'école primaire était loin d'être la règle. [...]

Géosem et Linguastream: identification, annotation, exploitation de tels segments



Geosem et LinguaStream (suite)

- Un objectif applicatif final: l'accès au contenu de documents géographiques par le biais de critères spatiaux et temporels
- Un objectif de recherche « en route »: plateforme d'expérimentation et mise en place d'instruments intégrés pour l'analyse des structures discursives : Linguastream (F. Bilhaut) http://www.linguastream.org/
  - exploiter les procédures de TAL éprouvées pour les niveaux de grains inférieurs (notion *d'enrichissement incrémental* des vues sur le corpus, Widlöcher & Bilhaut 2005)
  - répondre aux besoins spécifiques pour ce niveau d'analyse en termes d'annotation et de visualisation

Annoter pour expérimenter (1)

Ho-Dac (2007)

- Grand corpus différencié (23235 phrases, 3 groupes)
- Annotation automatique sur corpus « syntexisé » (ana. synt. par Syntex) et « xml-isé »
- Expérimenter = faire « jouer » les marqueurs annotés (analyse quantitative) pour en observer les interactions
- Approche « data-driven » → au-delà de la validation d'hypothèses, des découvertes
- Annotations: titres, cadres, reprises, expressions référentielles (degré d'accessibilité), structure du document (sections, paragraphes, niveaux de titres)

Annoter pour expérimenter (2)

Titre de niveau 0

Titre de niveau 1

Introducteur de cadre en début de section

Reprise du titre de niveau 1

http://w3.univ-tlse2.fr/erss/textes/pagespersos/hodac/index.html

La lutte contre le terrorisme : essai de bilan institutionnel

Face à l'urgence : les premières décisions de l'administration Bush

Dans le mois qui a suivi l'attentat du 11 septembre, l'administration a procédé à un certain nombre d'initiatives spectaculaires à plus d'un titre, notamment par l'intrusion massive des autorités fédérales dans différents domaines où, jusqu'alors, l'interventionnisme fédéral n'était pas de mise.

Annoter pour expérimenter (3): variabilité des marqueurs discursifs

- Importance de la position d'un marqueur dans le texte: p. ex. un introducteur de cadre temporel signale un nouveau segment ssi en début de section
  → un marqueur discursif est en fait une configuration [expr. lexico-synt A (+ expr. lexico-synt B) + position textuelle]
- fonctionnements différents des marqueurs selon type de texte
  → les configurations doivent inclure des paramètres liés au type de texte

Plan de l'exposé

Préambule

1. « Niveau discours »: un certain regard
2. Une linguistique du discours en corpus
3. **TAL et discours**

TAL et discours

A. Nazarenko (2005) Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel?

> « Au-delà de la recherche d'information qui se contente de sélectionner des documents dans une base documentaire, on voit apparaître depuis une décennie des outils d'accès au contenu même des textes et documents. Ces outils reposent sur des méthodes variées, depuis le simple surlignage de textes jusqu'à l'extraction d'information et aux systèmes de question-réponse. Dans cet article, nous nous interrogeons sur la nature de l'analyse sémantique qu'ils mettent en œuvre. » (p.211)

Applications envisagées: EI, Q-R, navigation, résumé

Constat : « Une analyse par îlots de texte »

> « On pourrait reprocher aux méthodes d'accès au contenu de considérer le texte comme une succession de syntagmes nominaux complexes. C'est mieux que les « sacs de mots » des moteurs de recherche mais est-ce suffisant? » (p.223)

Deux exemples

- FilText, NaviTexte: filtrage sémantique (http://www.lalic.paris4.sorbonne.fr/~minel/fichiers/presentation/Filtext/sld001.htm)
- Argumentative zoning: caractérisation de zones fonctionnelles (Teufel & Moens 2002)

FilText : extraction de phrases

- **Définition :**

« *Vapeur d'eau, gaz carbonique, monoxyde de carbone, méthane, oxyde d'azote, et ozone* ***sont ce que l'on appelle*** *communément des gaz à effet de serre* »

- **Récapitulation, conclusion thématique :**

« *Notre* ***deuxième conclusion*** *est que, à cause de l'effet de serre, l'intérêt de développer l'électronucléaire est devenu évident à un certain nombre d'hommes politiques* »

- **Annonce thématique :**

« ***Cet article*** *brosse un portrait du paysage électrique chinois actuel* »

- **Relation entre concepts :**

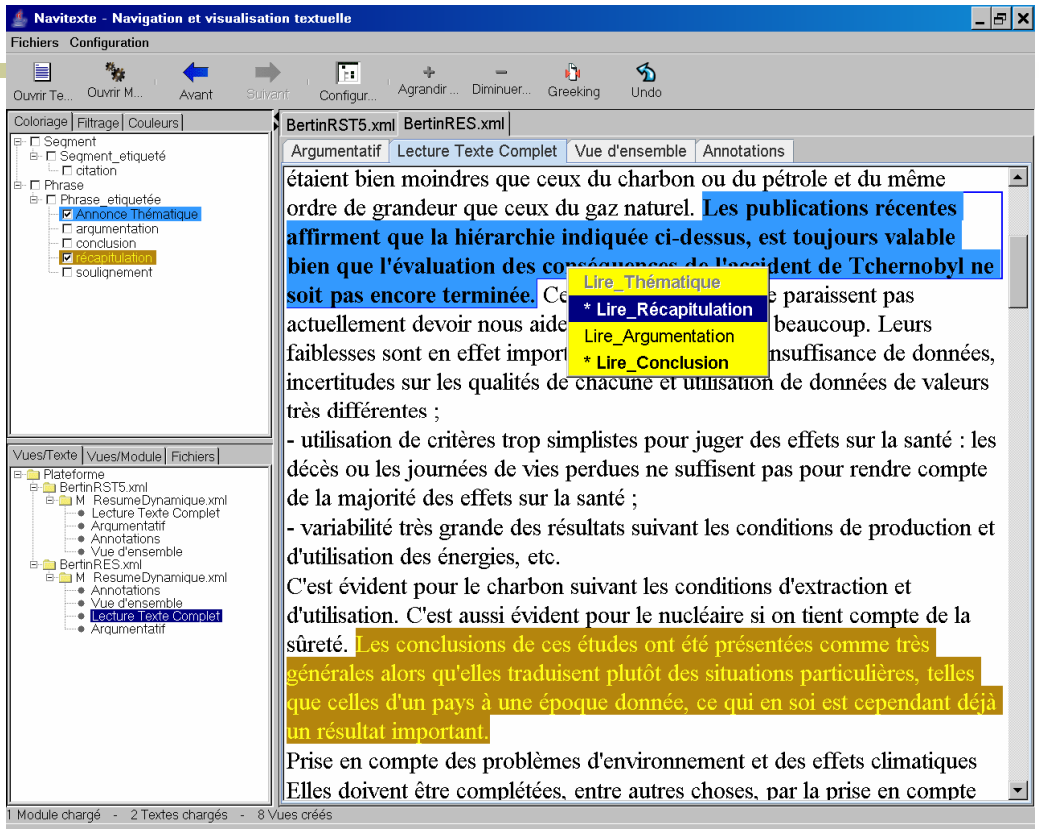
« *le livre* ***est composé*** *de plusieurs chapitres* »

FilText: recherche de citations

Dans son discours *de clôture, adapté pour tenir compte des réactions négatives enregistrées pendant la conférence, le* **Président Bush** ***affirme*** **qu'il** *ne propose pas le développement des recherches comme un moyen de remettre à plus tard les actions indispensables* […]

Il ***souligne*** *par ailleurs, à nouveau, qu'il ne propose pas une sorte de compromis entre développement économique d'une part, et protection de l'environnement d'autre part, mais qu'au contraire* **il** ***souhaite*** **que** *ces deux objectifs soient poursuivis simultanément et de manière synergique aussi rapidement que possible*

NaviTexte: aide à la navigation



Argumentative zoning: un modèle discursif d'annotation

Table 1
Annotation scheme for rhetorical status.

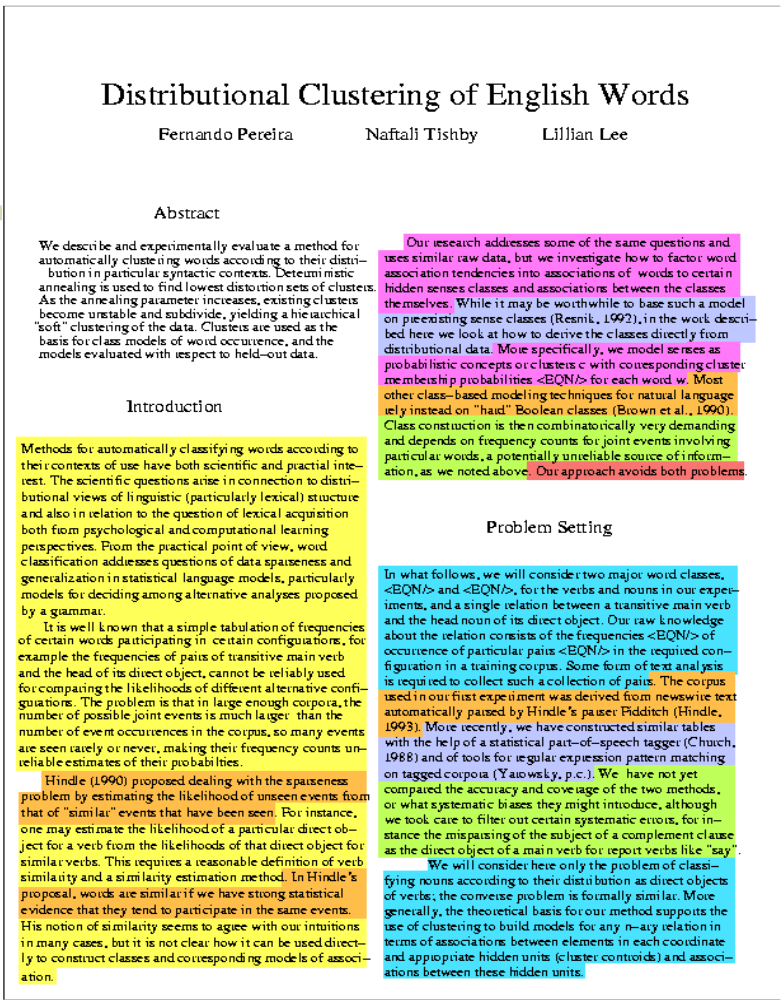
AIM	Specific research goal of the current paper
TEXTUAL	Statements about section structure
OWN	(Neutral) description of own work presented in current paper: Methodology, results, discussion
BACKGROUND	Generally accepted scientific background
CONTRAST	Statements of comparison with or contrast to other work; weaknesses of other work
BASIS	Statements of agreement with other work or continuation of other work
OTHER	(Neutral) description of other researchers' work

Argumentative zoning: exemples

AIM	10	*Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves.*
	11	*While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data.*
	162	*We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.*
BASIS	19	*The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993).*
	113	*The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering (Rose et al., 1990), in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter* EQN *following an annealing schedule.*
CONTRAST	9	*His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.*
	14	*Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above.*

Argumentative zoning et navigation

- BKG: General scientific background (yellow)
- OTH: Neutral descriptions of other people's work (orange)
- OWN: Neutral descriptions of the own, new work (blue)
- AIM: Statements of the particular aim of the current paper (pink)
- TXT: Statements of textual organization of the current paper (in chapter 1, we introduce...) (red)
- CTR: Contrastive or comparative statements about other work; explicit mention of weaknesses of other work (green)
- BAS: Statements that own work is based on other work (purple)



Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between these hidden units.

TAL et discours

Chantiers nombreux :

- Annoter des corpus
- Elaborer des plate-formes TAL associant traitements, outils d'extraction et de visualisation
- Identifier la faisabilité des traitements discursifs (passage à l'échelle)
- Identifier des zones de plus-value potentielle de traitements discursifs pour les applications du TAL

Références (1)

- Bilhaut, F., Ho-Dac, M., Borillo, A., Charnois, T., Enjalbert, P., Le Draoulec, A., Mathet, Y., Miguet, H., Péry-Woodley, M.-P., & Sarda, L. (2003). Indexation discursive pour la navigation intradocumentaire: cadres temporels et spatiaux dans l'information géographique. *TALN'03,* Batz-sur-Mer..pp 315-320.
- Carlson, L., Marcu, D., & Okurowski, M. E. (2002). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In J. van Kuppevelt & R. Smith (Eds.), *Current Directions in Discourse and Dialogue* (pp. 85-109). Dordrecht: Kluwer Academic Publishers.
- Charolles, M. (1997). *L'encadrement du discours : Univers, champs, domaines et espaces.* Cahier de Recherche Linguistique 6, LANDISCO, URA-CNRS 1035 Université Nancy 2. 1-73.
- Cole, R. A., Mariani, J., Uszkoreit, H., & Varile, G. B. (Eds.). (1998). *Survey of the State of the Art in Human Language Technology*. Pisa: Giardini.
- Cornish, F. (1998). Les "chaînes topicales" : leur rôle dans la gestion et la structuration du discours. *Cahiers de Grammaire(23*), 19-40.
- Couto, J., & Minel, J.-L. (2004,). Interfaces dynamiques de fouilles textuelles. *RIAO 2004*, Avignon. pp.420-430
- Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Gap/Paris: Ophrys.
- Halliday, M. A. K., "Text as semantic choice in social contexts", in J. Webster (Ed.), *The Collected Works of M.A.K. Halliday (Volume 2): Linguistic Studies of Text and Discourse,* London, Continuum, 2003, p. 23-81, (reprinted from van Dijk, T., Petöfi, J.S. (Eds.), *Grammars and Descriptions*, Berlin, Walter de Gruyter, 1977, p. 176-226).
- Ho-Dac, L.-M. (2007). *Exploration en corpus de la position initiale dans l'organisation du discours.* Thèse de doctorat de Sciences du Langage, Université de Toulouse 2, Toulouse.

Références (2)

- Luc, C., & Virbel, J. (2001). Le modèle d'architecture textuelle : fondements et expérimentation. Verbum, 23(1), 103-123.
- *Mani & Maybury (2001)* http://www.mitre.org/about/technical_centers/itc/maybury/manimayburysummarization.pdf
- Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3), 243-281.
- Marcu. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: MIT Press.
- Minel, J.-L. (2003). *Filtrage sémantique. Du résumé automatique à la fouille de textes*. Paris: Hermès-Lavoisier.
- Nazarenko, A. (2005). Méthodes automatiques d'accès au contenu. In A. Condamines (Ed.), *Sémantique et Corpus* Paris: Lavoisier. (pp. 211-244).
- Penn Discourse TreeBank http://www.seas.upenn.edu/~pdtb/
- Péry-Woodley, M.-P., & Scott, D. (2006) (eds.). Computational Approaches to Discourse and Document Processing. *TAL*, 47(2), Introduction 7-19.
- Péry-Woodley, M.-P. (2005). Discours, corpus, traitements automatiques. In A. Condamines (Ed.), *Sémantique et Corpus* (pp. 177-210). Paris: Hermès.
- Teufel, S., & Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics, 28(4)*, 409-445.
- Widlöcher, A., & Bilhaut, F. (2005,). La plate-forme Linguastream: un outil d'exploration linguistique sur corpus. *TALN 2005*, Dourdan, France, pp.517-522, http://taln.limsi.fr/site/talnRecital05/actes-articles.htm#tome1