

# Une analyse linguistique des modèles de reconnaissance automatique de relations causales\*

6<sup>e</sup> Journées de linguistique suisse

Cécile GRIVAZ

Directeurs : Jacques MOESCHLER<sup>1</sup> et Martin RAJMAN<sup>2</sup>

<sup>1</sup>Département de Linguistique  
Université de Genève

<sup>2</sup>Laboratoire d'Intelligence Artificielle  
École Polytechnique Fédérale de Lausanne

Vendredi 10 Décembre 2010

---

\*. Ce travail est financé par le projet Fonds National (n° 100012-113382)

# Lignes directrices

Introduction

Théories linguistiques : l'exemple de la SDRT

Approche informatique

# Lignes directrices

Introduction

Théories linguistiques : l'exemple de la SDRT

Approche informatique

# Systèmes de détection de relations causales.

- Systèmes automatiques.
- Relations causales **implicites** entre propositions.
- Langue naturelle.

## Exemple

Nous rentrons des courses. J'ai acheté des aubergines. **Jean est content, il a acheté un grille-pain.**

# Utilité des systèmes automatiques

- Résumé automatique.
- Systèmes de question réponse.
- Systèmes de dialogue.

# L'écart entre les théories linguistiques et les systèmes automatiques

- Plusieurs théories linguistiques ou philosophiques expliquent comment on peut détecter les relations causales implicites.
- Elles sont souvent **difficiles** à implémenter en pratique.
- Les systèmes automatiques existants simplifient ou ignorent les théories existantes.

# Lignes directrices

Introduction

Théories linguistiques : l'exemple de la SDRT

Approche informatique

# Comment peut-on reconnaître la causalité implicite ?

Comment un humain peut-il distinguer ces deux énoncés ?  
[Lascarides and Asher(1993)]

1. Max est tombé. Jean l'a poussé.
2. Max s'est levé. Jean l'a salué.

Deux grandes théories :

- Segmented discourse representation theory.
- Théorie de la pertinence.

# SDRT

- La SDRT est une théorie générale sur les relations de discours.
- Se base sur des règles utilisées par la personne qui interprète le discours.
- 6 règles sont nécessaires dans ce cas :

## SDRT suite

Max est tombé. Jean l'a poussé.  
Max s'est levé. Jean l'a salué.

- 1 La relation par défaut est la narration.
- 2 Dans la narration, les évènements sont dans un ordre chronologique.
- 3 S'il s'agit de narration, les deux unités de discours doivent avoir un sujet commun.

## SDRT fin

Max est tombé. Jean l'a poussé.

Max s'est levé. Jean l'a salué.

- 4 Si deux évènements sont liés par une relation de discours et l'un décrit un TOMBE(X) et l'autre un POUSSE(X,Y), les deux évènements sont normalement liés causalement et le POUSSE cause le TOMBE.
- 5 Une cause précède un effet temporellement.
- 6 Si deux évènements doivent être liés et le second est la cause du premier, alors la relation est *explication*.

## Difficulté d'implémentation

- 1 *La relation par défaut est la narration :*  
Facile.
- 2 *Dans la narration, les évènements sont dans un ordre chronologique :*  
Facile tant qu'il ne faut pas tester.
- 3 *S'il s'agit de narration, les deux unités de discours doivent avoir un sujet commun :*  
Facile tant qu'il ne faut pas tester.

## difficulté d'implémentation : suite

- 4 *Si deux évènements sont liés par une relation de discours et l'un décrit un TOMBE(X) et l'autre un POUSSE(X,Y), les deux évènements sont normalement liés causalement et le POUSSE cause le TOMBE :*
- Difficulté pour produire une telle liste.
  - Difficulté de reconnaissance des rôles thématiques.
  - Difficulté de reconnaissance de la nature des évènements en cas de paraphrase.

## difficulté d'implémentation : fin

- 5 *Une cause précède un effet temporellement :*  
Facile tant que ça ne doit pas être vérifié.
- 7 *Si deux évènements doivent être liés et le second est la cause du premier, alors la relation est explication :*  
Facile, si la causalité a été reconnue.

*En plus :*

Relative difficulté à implémenter le système de logique qui fait fonctionner les règles les unes avec les autres.

# Lignes directrices

Introduction

Théories linguistiques : l'exemple de la SDRT

Approche informatique

## Approche informatique : la classification supervisée

- Système **statistique** plus facile à implémenter et plus difficile à interpréter qu'un système à base de règles.
- Se base sur des **exemples** et construit des règles à partir d'eux.
- Construit les règles à partir d'**indices de classification**.
- Ces indices doivent être déterminés à l'avance.
- Système le plus populaire pour la reconnaissance de relations de discours et de causalité.

# Indices de classification et leurs interprétations linguistiques : exemples de [Pitler et al.(2009)Pitler, Louis, and Nenkova]

Quelques indices faciles à interpréter

- Paires de mots : connaissance du monde.
- Étiquettes syntaxiques des verbes : temporalité.
- Classe de Levin des verbes :
  - D'après Pitler et al. les verbes de même classe seraient indices d'un développement.
  - Mais [Moeschler(2003)] indique que la causalité serait plus facile avec des alternances d'évènements et d'états.

# Indices de classification et leurs interprétations linguistiques

## Indices plus difficiles à interpréter

- Trois premiers et dernier mot de la phrase : contiennent souvent des connecteurs ou des expressions apparentées.
- Contexte : les expériences de classification automatique indiquent que les relations n'apparaissent pas dans n'importe quel ordre. (Contingences implicites après comparaisons explicites.)

## Interpréter les résultats d'un classificateur

- Le fait qu'un indice augmente la qualité des résultat prouve qu'il varie avec les classes, mais il est parfois difficile de savoir comment.
- Les résultats d'un classificateur peuvent être **difficiles à interpréter**.
- Il est souvent possible de les rendre plus clairs en utilisant un algorithme moins puissant.

## Exemple d'artefact de classification : le cas des mots à catégorie fermée

- Idée pour obtenir un corpus d'entraînement d'implicites : prendre des explicites et **enlever** le connecteur. [Marcu and Echihabi(2001)], [Sporleder and Lascarides(2007)]
- Modélisation de la connaissance du monde : paires de mots.
- Enlever les mots à catégorie fermée des indices de classification rend les résultats moins bons.
- La présence de connecteur indique une **autre** relation de discours. [Pitler et al.(2009)Pitler, Louis, and Nenkova]

*Jean n'aime pas les haricots, mais il mange beaucoup de biscuits, parce qu'il est très gourmand.*

**cause** Jean n'aime pas les haricots, **mais** il mange beaucoup de biscuits, ~~parce qu'il~~ est très gourmand.

**contraste** Jean n'aime pas les haricots, ~~mais~~ il mange beaucoup de biscuits, **parce qu'il** est très gourmand.

# Conclusions

- Difficulté d'implémenter les théories linguistiques dans un système informatique.
- Difficulté d'analyser linguistiquement les approches informatiques statistiques.
- **Utilité** des travaux de chaque discipline pour l'autre.

## Bibliographie I



A. Lascarides and N. Asher.

Temporal interpretation, discourse relations and commonsense entailment.

[Linguistics and Philosophy](#), 16(5), 1993.



D Marcu and A. Echihabi.

An unsupervised approach to recognizing discourse relations.

In [ACL '02 : Proceedings of the 40th Annual Meeting on Association for Computational Linguistics](#), pages 368–375, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

doi : <http://dx.doi.org/10.3115/1073083.1073145>.



J. Moeschler.

Causality, lexicon, and discourse meaning.

[Rivista di Linguistica](#), 15.2, pages 343–369, 2003.

## Bibliographie II



E. Pitler, A. Louis, and A. Nenkova.

Automatic sense prediction for implicit discourse relations in text.

In ACL-IJCNLP '09 : Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2, pages 683–691, Morristown, NJ, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-46-6.



C. Sporleder and A. Lascarides.

Exploiting Linguistic Cues to Classify Rhetorical Relations.

In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, Recent Advances in Natural Language Processing IV : Selected Papers from RANLP 2005, volume



## Bibliographie III

292 of Current Issues in Linguistic Theory, pages 157–166.  
John Benjamins, Amsterdam & Philadelphia, 2007.