Méthodes Symboliques en Traitement Automatique des Langues

Journée des instituts INS2I & INSHS du CNRS sur le Traitement Automatique des Langues en France (organisée par Isabelle Tellier)

Nicholas Asher Christian Retoré

(IRIT, Toulouse)

LORIA, Nancy, 15 janvier 2013

Plan

Motivation pour les méthodes symboliques

État de l'art et perspectives en

- syntaxe
- sémantique
- pragmatique
- discours
- dialogue

Que serait un traitement symbolique des langues ?

Modélisation (souvent logique) et étude de la complexité du domaine choisi

- Implémentation directe d'une axiomatisation
- Moins présent qu'il y a 20 ans dans les colloques majeurs de TAL mais néanmoins utile.
- Guide pour l'annotation des données et pour les traits en vue de l'apprentissage sur les données.
- Apprentissage guidé par des connaissances linguistiques préalables.

Quels objectifs requièrent des méthodes formelles ?

- Recherche d'informations précises / *question answering*
- Dialogue homme machine

Analyse profonde de peu de données

La langue permet d'affirmer, de réfuter, de raisonner
la LOGIQUE joue un rôle central
($\neq$ images, signaux, autres types d'information)

Analyse profonde de peu de données

Connecteurs logiques, coréférence des pronoms :

Ex. de tweets sur les catastrophes naturelles (LIRMM) :

Il y a du vent, mais ce n'est pas un ouragan

L'immeuble tremble, mais la nuit il n'y a pas de métro

Ex. de réponse à des questions :

Quelle était la femme de Geach ?

Etait-il l'élève de Wittgenstein ?

En 1941, IL épousa la philosophe Elizabeth Anscombe, grâce à LAQUELLE IL entra en contact avec Ludwig Wittgenstein. Bien qu'IL N'ait JAMAIS suivi l'enseignement académique de CE DERNIER, cependant IL EN éprouva fortement l'influence (Article sur Geach, sur Wikipedia)

Analyse profonde de peu de données

Reconstruction d'itinéraires dans des récits de voyage du XIXe au voisinage d'une localité, une fois retrouvés les passages pertinents.

Nous coupons ici un sentier qui vient du port de Barroude (...)
Plus loin, de nobles hêtres montent sur le versant (...)
cette route qui monte sans cesse pendant deux lieues
Le chemin pavé de calcaire et de pierres luisantes (...)
serpente à travers fourrés de buis et de noisetiers
Puis, cinq minutes nous conduisent à un petit pont (...) qui nous porte sur la rive droite.

La syntaxe en symbolique : panorama

- Classes de langages (légèrement contextuels)
- Description unifiée par des métagrammaires (HPSG, LFG, TAG)
- Équivalences de certaines théories syntaxiques via les ACGs (TAGs, CGs, MCFG,..)
- Deux descriptions des arbres d'analyse syntaxique
  1. **Arbres satisfaisant un ensemble de contraintes logiques.**
  2. **Arbres obtenus récursivement par des règles.**

La syntaxe en symbolique : acquisition

Tâche en cours :

Ecriture de la grammaire "à la main"
tâche ardue, de quelle description de la langue partir ?
(grande grammaire du français d'Anne Abeillé)

Extraction de grammaires sur corpus
(entre 1 et 500 règles par mot, 100 en moyenne)
–> super tagging, analyse des n meilleures séquences de TAGs (sinon 20^{100} analyses ! !).
Info sémantiques difficiles à acquérir.

Syntaxe et TAL : exemple de travaux en cours

Création de grammaires HPSG multilingues à partir de paramètres testés en analyse et en génération.

Génération automatique à partir de représentations sémantiques d'exercice pour les personnes apprenant une langue étrangère.

La sémantique

- Bonne maitrise théorique des équivalences entre théories sémantiques (dynamiques et statiques)
- Interface syntaxe-sémantique des théories syntaxiques (TAG, HPSG, LFG), peu satisfaisante, peu de couverture—exceptions, Boxer (Bos), Grail (Moot) avec les CG
- Passage de Montague à la DRT présentée de manière compositionnelle (Muskens, de Groote)
- Des projets en cours (par ex. ANR Polymnie)
- On regrette le GDR *Sémantique et modélisation* (2002–2008, Corblin)

La sémantique lexicale (dans un cadre compositionnel)

- une attention accrue sur la polysémie (Cruse 1986, Pustejovsky 1995)
  1. **The lunch was delicious but took forever (copredication avec des prédicats avec des restrictions de sélection incompatibles : nourriture, événement)**
  2. **John studies piano (coercition)**
  3. **John studies a piano. (pas de coercition)**
  4. **The omelet is getting restless (coercition)**
- la polysémie est très fréquente dans la langue (aspect, verbe + préposition) et varie de langue en langue

Theories correspondantes (lexiques compositionnels)

- Cadre formel : dans la lignée de la sémantique lexicale de Dowty (1979) elle-même issue de la sémantique de Montague.
- La polysémie est elle un phénomène sémantique ou pragmatique ?
- beaucoup d'arguments pour que ce doit être prise en compte par la sémantique lexicale et compositionnelle (Asher 2011)
- Modèles formels sémantiques de la polysémie en théorie des types qui convergent :
  Système F (Bassac, Moot, Retoré), Type Composition Logic (Asher), Modern Type Theories (Luo, Fernando, Cooper)

Questions pour l'avenir sur le lexique

- Nous avons maintenant des très gros corpus pour l'analyse distributionnelle, aussi des corpus digitalisés de dictionnaires ainsi que des corpus associatifs Est-ce qu'on peut utiliser ces données ensemble ?
- Si oui, comment ?
- Comment est-ce que ces données interagissent avec les approches symboliques ? Par exemple des données statistiques sur la polysémie.

L'interface entre sémantique et pragmatique

- Relié à la sémantique formelle (par ex. via Grice)
- Aspect théoriques encore balbutiant, mais en bonne voie.
- Présupposition et implicature et structuration discursive.
- Formalisation et implémentation en logique non-monotone

Discours

- Plusieurs théories (RST, DLTAG, SDRT, Graphbank), plusieurs corpus.
- Implémentation symbolique actuellement hors d'atteinte sauf pour des fragments très restreints.
- Le versant empirique est très important, et différents corpus sont disponibles. La modélisation symbolique aide à déterminer les structures adéquates, à évaluer les annotations et à extraire les traits.
- Interactions entre modèles symboliques et méthodes statistiques : un défi attrayant.

Théories du discours

- Structure discursive—choix entre arbres ou graphes, avec ou sans récursivité...
- Liens entre la structure discursive et les effets sémantiques et pragmatiques

Données pour discours

- Annotation de structures complètes (RST treebank, ANNODIS, DISCOR, Graphbank) ou partiel (PDTB)
- Compromis entre taille du corpus et couverture de l'annotation
- Annotation complexe, donc très coûteuse et souvent bruitée
- Faut-il un unique standard de référence ou en considérer plusieurs ?

Interactions entre modèles symboliques et methodes statistiques

- Les structures temporelles sont très contraintes (consistance, transitivité), elle peuvent guider des modèles statistiques
- Il y a aussi des contraintes sur les structures discursives mais elles sont moins connues.
- contraintes globales sur l'apprentissage local, par exemple, d'attachements (Muller et al, Coling, 2012)
- Une définition sémantique des relations discursives peut guider le choix des traits pour l'apprentissage des relations et donner des contraintes globales sur les modèles d'apprentissage.

Dialogue

- Moins de théories que pour le discours (SDRT, modèle QUD)
- Moins de données "réels" annotées (Verbmobil, exception CID Aix, mais pas d'annotation sémantique)—surtout sur la conversation stratégique (Corpora de Traum et celui de STAC en cours)
- Les modèles actuels étendent les modèles discursifs, les perspectives différentes des interlocuteurs différents leur échappent.
- Nécessité d'intégrer des modèles issus de la théorie des jeux.
- la cooperativité gricéenne est problématique dans la conversation strateégique.
- Jeux de signalisation et problèmes de crédibilité, de ce qui a été dit ou communiqué (Asher & Lascarides, 2013).
- Jeux infinis

Conversations stratégiques

Misdirection

(1a) **Prosecutor : Do you have any bank accounts in Swiss banks, Mr. Bronston ?.**

(1b) **Bronston : No, sir.**

(1c) **P : Have you ever ?**

(1d) **B : The company had an account there for about six months, in Zurich.**

Les étapes d'analyse

- **Analyse lexicale et compositionelle.**
- **Analyse discursive**
- **Évaluation stratégique de ce qui a été dit (plausible deniability)**
- **Évaluation de la crédibilité.**
- **Enchaînement des réponses et poursuite de la conversation.**

Les jeux infinis

- Pourquoi : une conversation n'a pas de fin conclusive.
- Une riche typologie de jeux infinis
  - Banach Mazur
  - Gale Stewart
- une conversation est une suite infinie de coups de dialogue
- qui forment une espace topologique métrisable
- On peut la caractériser en utilisant la hiérarchie de Borel des conditions de gagner dans ces jeux pour caractériser la complexité des stratégies conversationnelles (accessibilité, stabilité, Muller)

Un exemple issu de Stac

Speaker	Id	Turn	Dom. function	Rhet. function
Euan	47	[And I alt tab back from the tutorial.].1	other	
		[What's up ?].2	other	Result*(47.1,47.2)
Joel	48	[do you want to trade ?]	offer ⟨Joel, ?, ?,Euan⟩	Q-elab(47.2, 48)
Card.	49	[joel fancies a bit of your clay]	strat.-comment	Expl*(48, 49)
Joel	50	[yes]	other	Ackn(49, 50)
Joel	51	[!]	other	Comment(50, 51)
Euan	52	[Whatcha got ?]	counteroffer ⟨Euan, ?, ?,Joel⟩	Q-elab([48-50], 52)
Joel	53	[wheat]	has-resources ⟨Joel,wheat⟩	QAP(52, 53)
Euan	54	[I can wheat for clay.]	counteroffer ⟨Euan,wheat,clay,Joel⟩	Elab([52,53], 54)
Joel	55	[awesome]	accept(54)	Ackn(54, 55)

En guise de conclusion : quelques questions classiques et utiles de sémantique/discours

- Descriptions définies, déterminants et quantifications (qui fait quoi)
- Chaînes anaphoriques (structuration du discours)
- Négation, raisonnement dans la langue (pertinents en recherche d'information)
- Sémantique du temps et de l'espace. (par ex. analyse d'itinéraires)

Pour toutes ces questions de nature LOGIQUE, un défi actuel est la gestion CONTEXTE discursif.