Informatique et Industrie de la langue
Annotation

Pascal Amsili

Université Sorbonne Nouvelle

Février 2020

Ce diaporama reprend en grande partie des transparents utilisés par Isabelle Tellier au cours des années scolaires précédentes.

Plan

1. Annoter pour quoi faire

Qu'est-ce qu'annoter ? (point de vue d'un(e) informaticien(ne))

- c'est une tâche relevant de la fouille de textes
- les données de départ sont constituées d'items (éléments) reliés entre eux
  - texte = séquence linéaire de mots (1 relation d'ordre)
  - arbre = structure arborée de balises (2 relations d'ordre)
- les items appartiennent à un vocabulaire fini
- la sortie : chaque item de la donnée de départ est associée à une étiquette
- les étiquettes appartiennent à un autre vocabulaire fini
- les données et les annotations ont la même structure

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Etiquetage POS ("part of speech")

- item = "mot"
  donnée = séquence de mots (=phrase)
  annotation = séquence des catégories morpho-syntaxiques des mots dans la phrase
- plusieurs niveaux d'annotations possibles :

Le	*petit*	*chat*	*est*	*mort*
Det	Adj	NC	V	Adj
DetMSDef	AdjMS	NCMS	VIP3S	AdjMS

- difficulté : plusieurs étiquettes possibles pour chaque mot, un dictionnaire ne suffit pas !
- format textuel possible :
  phrase annotée = Le/Det petit/Adj chat/NC est/V mort/Adj

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Segmentation d'un texte en "chunks"

- chunk = constituant non récursif
- analyse syntaxique superficielle
- équivaut à un parenthésage total non récursif typé ou non
- peut être codé par une annotation B/I (Begin/In)

Il	$voit$	sa	$voisine$	$avec$	des	$jumelles$
(Il)	$(voit)$	$(sa$	$voisine)$	$(avec$	des	$jumelles)$
B	B	B	I	B	I	I
$(Il)_{GN}$	$(voit)_{GV}$	$(sa$	$voisine)_{GN}$	$(avec$	des	$jumelles)_{GP}$
GN-B	GV-B	GN-B	GN-I	GP-B	GP-I	GP-I

- format textuel possible :

  phrase annotée = <GN>Il</GN> <GV>voit</GV> <GN>sa voisine</GN> <GP>avec des jumelles</GP>

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Reconnaissance des entités nommées (extraction d'information)

- entité nommée = nom propre (personne/lieu/organisation), date, valeur numérique
- porteur de l'information factuelle d'un texte
- peut être codé par une annotation BIO (Begin/In/out)

En	2016	*les*	*Jeux*	*Olympiques*	*auront*	*lieu*	*a*	*Rio*	*de*	*Janeiro*
	date		evt	evt				lieu	lieu	lieu
O	D-B	O	E-B	E-I	O	O	O	L-B	L-I	L-I

- format textuel possible :

  phrase annotée = En <Date>2016</Date>, les <Evt>Jeux Olympiques</Evt> auront lieu à <Lieu>Rio de Janeiro</Lieu>.

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Reconnaissance des entités nommées (extraction d'information)

- format "CoNLL"

unité	étiquette
En	O
2016	D-B
,	O
les	O
Jeux	E-B
Olympiques	E-I
...	

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Alignement de phrases bilingues

- nécessite des phrases alignées, traductions l'une de l'autre :

	J'	aime	le	chocolat
I	X			
like		X		
chocolate				X

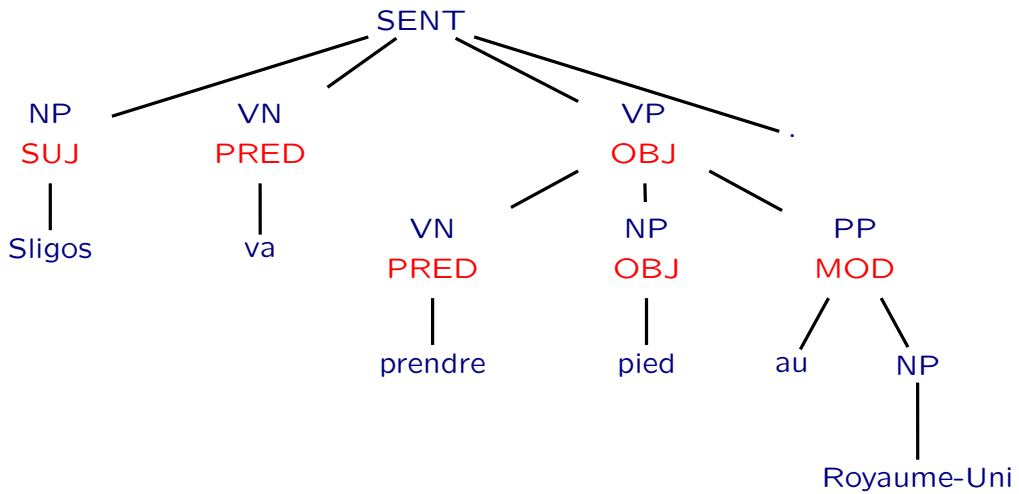
- on code les correspondances entre mots par des annotations

J'_1	$aime_2$	le_3	$chocolat_4$	I_1	$like_2$	$chocolate_3$
1	2	-	3	1	2	4

- chaque annotation réfère aux mots de l'autre phrase
- étape préliminaire des systèmes de traduction automatique statistique

1. Annoter pour quoi faire
Exemple d'annotations sur des arbres

étiquetage fonctionnel d'arbres syntaxiques



- étiquetage en rôles thématiques/sémantiques d'arbres syntaxiques : idem mais avec annotation agent, patient, etc.
- extraction d'information sur le Web ou les documents XML

1. Annoter pour quoi faire
Exemples d'annotations

Autres types d'annotations de séquences

- annotation de mots en phonèmes pour la synthèse vocale
- segmentation de textes en unités lexicales
- segmentation de phrases en clauses (propositions indépendantes)
- segmentation de dialogues en tours de paroles
- annotation de phrases successives d'une dépêche d'agence :
  item = phrase
  annotation = classe (evt présent / evt passé / commentaire)
- annotation d'une page HTML :
  item = segment de page
  annotation = classe (menu / publicité / titre / contenu)

1. Annoter pour quoi faire

Synthèse

- annoter une donnée = l'enrichir en préservant sa structure (mais sans la créer)
- permet de traiter de nombreuses tâches
- chaque tâche requiert de spécifier :
  - la nature des items (découpage initial)
  - les relations entre items : séquence, ordres dans un arbre...
  - la nature des annotations et leur interprétation
  - les relations entre annotations
  - les relations entre les items et leur annotation annotation possible (dictionnaires utilisables pour cela)
- pré-traitements et post-traitements souvent nécessaires pour coder/décoder les annotations
- différents formats d'annotation possibles, plus ou moins riches
- on peut faire des annotations successives de données

Utilisation des annotations

Les annotations servent :

- aux linguistes (phénomènes non surfaçiques) :
  - pour trouver des exemples
  - pour compter des occurrences

  dans un but de :
  - évaluation d'un modèle théorique
  - observation d'un phénomène linguistique
- aux informaticiens :
  - pour entraîner des systèmes de TAL
  - pour évaluer des systèmes de TAL

\+ comme « patrimoine »

Plan

Protocole d'une annotation manuelle multiple

Pour annoter à la main :

1. faire annoter une même portion des données par plusieurs annotateurs
2. observer les annotations qui convergent
3. discuter des annotations qui divergent jusqu'au consensus
4. mettre à jour le *guide d'annotation*
5. ré-appliquer les étapes de 1 à 4 sur les autres portions des données
6. produire un corpus annoté de référence
   (accord inter-annotateurs égal à 100%)

Qualité d'une annotation manuelle

Une bonne annotation est une annotation reproductible, donc généralisable

Qualité d'une annotation manuelle

Une bonne annotation est une annotation reproductible, donc généralisable

- Un seul annotateur : aucune possibilité de mesurer la qualité

Qualité d'une annotation manuelle

Une bonne annotation est une annotation reproductible, donc généralisable

- Un seul annotateur : aucune possibilité de mesurer la qualité
- Plusieurs annotateurs : aucune possibilité de mesurer la qualité après adjudication (=arbitrage)

Qualité d'une annotation manuelle

Une bonne annotation est une annotation reproductible, donc généralisable

- Un seul annotateur : aucune possibilité de mesurer la qualité
- Plusieurs annotateurs : aucune possibilité de mesurer la qualité après adjudication (=arbitrage)
- Avant adjudication : mesure d'un **accord inter-annotateur**

Mesures d'accord inter-annotateurs

- on peut comparer les annotateurs entre eux
- on peut comparer les annotateurs avec la référence

Méthodes :

- Matrice de contingence

Sam \ Sim	OUI	NON
OUI	20	5
NON	10	15

- Kappa (κ) de Cohen (ou de Fleiss)

Mesures d'accord inter-annotateurs

- on peut comparer les annotateurs entre eux
- on peut comparer les annotateurs avec la référence

Méthodes :

- Matrice de contingence

Sam \ Sim	OUI	NON
OUI	20	5
NON	10	15

- Kappa (κ) de Cohen (ou de Fleiss)

accord = 0,7
κ = 0,4 ("accord modéré")

Annoter à la main...

- c'est long donc coûteux
- ça requiert des compétences linguistiques
- c'est à recommencer pour toute nouvelle donnée
- c'est la seule façon de garantir une annotation de grande qualité
- (à condition d'utiliser plusieurs annotateurs pour mesurer la qualité)

Annoter à la main...

- c'est long donc coûteux
- ça requiert des compétences linguistiques
- c'est à recommencer pour toute nouvelle donnée
- c'est la seule façon de garantir une annotation de grande qualité
- (à condition d'utiliser plusieurs annotateurs pour mesurer la qualité)

Alternative : annoter par programme !

2. Annoter par programme

Programmes disponibles en ligne

- segmenteur/étiqueteur POS de l'université de Copenhague (surtout anglais et danois) http://cst.dk/online/pos_tagger/uk/
- étiqueteur de Lancaster (anglais) : http://ucrel.lancs.ac.uk/claws/trial.html
- un autre (appris sur le Penn Treebank) : http://nlpdotnet.com/services/Tagger.aspx
- le prix de la plus jolie interface : http://parts-of-speech.info
- POS + chunks : http://www.infogistics.com/posdemo.htm
- étiqueteur POS et entités nommées de Chicago : http://cogcomp.cs.illinois.edu/page/demo_view/POS http://cogcomp.cs.illinois.edu/page/demo_view/NER
- segmenteur/étiqueteur multilingue (dont français !) : https://open.xerox.com/Services/fst-nlp-tools

2. Annoter par programme

Programmes existants gratuits en français

- pour l'annotation POS : TreeTagger, Melt, SEM
- pour le chunking : SEM
- pour la reconnaissance des entités nommées : CascEN, SEM

Problèmes

- ils font des erreurs (parfois aberrantes)
- pourquoi ? ça dépend comment ils ont été construits !
- combien ? il existe des mesures pour les quantifier

2. Annoter par programme
Comment sont construits les programmes ?

1ère approche : un programme écrit "à la main"

- fondé sur des ressources (dictionnaires, listes...)
- et sur des règles (si "le" précède un verbe : PRO, sinon : DET)
- reflète une expertise linguistique
- est spécifique d'une langue, d'un genre de textes...
- est utilisable sur des données bien normées
- cette écriture est longue, laborieuse, à recommencer sur des données avec des propriétés différentes...
- un des débouchés possibles des linguistes !
- outil possible : Unitex

2. Annoter par programme
Comment sont construits les programmes ?

2ème approche : un programme appris automatiquement

- Apprentissage automatique (AA) : branche de l'IA
  AA = art de transformer des données (exemples) en programmes !
- provient de la recherche en informatique, devenu accessible
- l'AA s'adapte à la langue des exemples mais nécessite :
- de disposer (de beaucoup) d'exemples de données déjà annotées
- d'où : encore besoin des linguistes pour annoter (à la main) des exemples !
- et proposer des indices (attributs) permettant de choisir une étiquette

2. Annoter par programme
Annoter par apprentissage automatique

Application à une tâche d'annotation

unité	Maj ?	Chiffre ?	Ponct ?	étiquette
En	1	0	0	O
2016	0	1	0	D-B
,	0	0	1	O
les	0	0	0	O
Jeux	1	0	0	E-B
Olympiques	1	0	0	E-I
...				

- il existe de nombreuses méthodes pour prédire la dernière colonne : Weka

Apprentissage automatique

- Supervisé “classique” : on fournit les exemples et on recherche les indices (= traits, = variables indépendantes)
- Supervisé neuronal : on fournit les exemples, et le programme d’apprentissage choisit la bonne représentation (ex : images, plongements lexicaux)
- Non supervisé : on fournit des données, mais sans les annotations (ex : clustering, modèles de langue)

2. Annoter par programme

Performances actuelles

- Qualités des programmes actuels
  - les programmes écrits à la main se comportent bien sur les cas connus mais oublient les cas particuliers
  - les programmes appris automatiquement généralisent bien mais peuvent créer des règles trop générales
- Ordre de grandeur des meilleurs programmes (obtenus par AA) :
  - annotation en POS : $97 - 98\%$ de bonnes étiquettes
  - chunking : environ 95% de bonnes étiquettes
  - analyse syntaxique (dépendances) : environ 90% de bonnes étiquettes
  - entités nommées : entre 60% (compétition récente sur des tweets) et 90% (textes journalistiques)

Plan

Conclusion

- De très nombreux traitements linguistiques sur des données s'expriment par une tâche d'annotation
- Pour annoter, on peut :
  - annoter à la main (long, laborieux...)
  - chercher un programme existant (rarement parfaitement adapté au besoin) : il faut comprendre/analyser ses erreurs
  - construire soi-même un programme !
    - en programmant "à la main"
    - en annotant un échantillon et en utilisant de l'apprentissage automatique
- Dans tous les cas : des compétences linguistiques sont utiles, à un moment ou un autre...
- ... et des compétences informatiques !