Panel de discussion sur le thème “Variation et diversité linguistique”

LIFT 2023 : Journées Scientifiques du Groupement de Recherche « Linguistique Informatique, Formelle et de Terrain »

Questions proposées

Question	Likes	Dislikes
Comment traiter les données de manière efficace et en même temps ne pas laisser de côté la variation linguistique dans les langues ?	2	0
Quelle est la portée de la génération de variations artificielles pour surmonter le problème des langues à faibles ressources ?	1	0
Ouvrir la boîte noire des Transformers pour comprendre les corpus d’apprentissage automatique : comment rendre accessibles les LMs ou les LLMs ?	0	4
Les approches formalistes peuvent-elles perdurer dans une époque où la variabilité linguistique est de plus en plus étudiée ?	10	0
Comment évaluer la représentativité d’un corpus ?	10	2
Quelles métadonnées pour décrire la variation potentielle dans les données?	10	3
Dans le cas des langues peu documentées, faut-il prendre en compte la variation dès le début des traitements automatiques ou faut-il privilégier une normalisation des données pour les exploiter ?	13	1
dans quelle mesure doit-on expurger son corpus des possibles insultes, propos sexistes/ homophobes ? Même si ces propos ne sont pas souhaitables, ils font partie de l’usage réel de la langue.	4	1
comment traiter la variation dans une perspective du TALN quand les frontières ou distinctions entre zones linguistiques et groupes de population ne sont parfois si évidents?	6	0
Quelle place donner à la norme langagière ?	2	1
Quelles dimensions de la variation posent le plus de difficultés pour l’analyse linguistique ?	11	0
Comment traiter les langues typologiquement rares automatiquement quand les modèles sont entraînés à partir des langues les plus documentées ?	2	0
Comment faire évoluer les représentations et les technologies avec la variation diachronique en cours ?	3	0

Questions sélectionnées et notes de discussion

Dans le cas des langues peu documentées, faut-il prendre en compte la variation dès le début des traitements automatiques ou faut-il privilégier une normalisation des données pour les exploiter ?

Prise de notes par Carole Werner

Exemple : corpus immense contenant variation diachronique. Annotateurs différents qui ont transcrit de la variation ; variation pose problème quand on veut entraîner des modèles de keyword spotting ; ce qui est le plus intéressant pour nous c'est de pouvoir exploiter les données et de faire face à la variation qui est une partie inhérente de la langue. Pour entraîner un outil qui marche bien, la variation pose problème. Trouver une solution d'abord ou trouver une solution après ?

- Quand on annote : les transcriptions ne sont pas régulières (ex. dans le cas des créoles, à la française, à la portugaise ou à la créole)
- Variation ling doit être représentée d'une façon ou d'une autre :
- Façon dont on choisit de la représenter : engendre des biais
- De quel type de variation parle-t-on ?
  - Tout dépend du contexte de la langue, du contexte
  - Difficile de donner une réponse globale, car corpus-dépendante
- Dans le cas d'un travail sur le français diachronique :
  - Variation dans l'analyse des données d'entraînement
  - Taille des phrases
  - Modèle entraîné sur la syntaxe d'une période peut-il analyser la syntaxe d'une autre période ? période proche oui, période éloignée, non.
- Diachronie et synchronie ?
- Cas de l'oralité : essayer de passer l'étape de la transcription et de chercher directement dans le signal
- Pour l'écrit contemporain : choix politique : qu'est-ce qu'on décide qui est la norme ?
- En diachronie : pas de normalisation souhaitable, car altération de la donnée brute
- Variation donne des résultats faussés dans certains cas

Quelles dimensions de la variation posent le plus de difficultés pour l'analyse linguistique ?

Prise de notes par Marina Seghier

2 questions se dégagent :

1. quelles dimensions sont difficiles à décrire ?
2. en quoi elles sont un obstacle à l'analyse linguistique ?

quelles variations ? variations à l'intérieur d'une langue, entre les langues…

4 dimensions : diachronie (temps), diaphasique (situations de communication), diatopique (lieux), diastratique (classes sociales)

difficultés :

- quand il n'y a pas de conventions orthographiques
- registre va être différent selon genre (homme/femme)
- différence entre discours préparés et spontanés

corpus interactifs —> unité sur plusieurs tours de parole ?

- modèles pour l'acquisition du langage —> les adultes ne peuvent pas forcément

interpréter ce que produisent les enfants
    - catégorisation des unités chez un enfant (ex : une petite fille en voyant la mer pour la première fois dit “à boire beau” —> à boire, soit l’eau, considéré comme un groupe nominal)
- variation au sein d’une même langue selon plusieurs “genres”, catégories de textes (ex : encyclopédie, informations, poésie, prose, documents officiels…), mettent en difficulté les outils du TAL pour plusieurs tâches (étiquetage en parties du discours, morphosyntaxique, REN…)
- productions de locuteurs atypiques (raisons physiques, motrices, etc…) = accès aux données, imprévisibilité…
- idem pour la langue des signes, il y a des signeurs atypiques

Les approches formalistes peuvent-elles perdurer dans une époque où la variabilité linguistique est de plus en plus étudiée ?

Prise de notes par Julien Bezançon

Cadre

- période d’opposition entre approches quantitatives statistiques et approches qualitatives
- aussi période d’opposition entre approches interprétables et non-interprétables
- variabilité de plus en plus étudiée, puisque les systèmes sont de plus en plus performants (on étudie ce qui bloque)

Première piste : définir “Formalisme” ?

- de manière générale, c’est donner des règles, catégoriser
- en langue des signes, ça correspond à une non-ambiguïté (exemple : signes lisibles et compréhensibles pour une machine)
- accord sur le fait que la notion de formalisme n’est pas forcément la même d’un domaine à l’autre
- approches formalistes vont avoir du mal avec la notion de variabilité (= avec les éléments qui ne correspondent pas à une catégorie formelle / qui vont à l’encontre des règles établies dans un cadre formel)

Formalisme vs autres approches

- première comparaison avec les approches statistiques, avec notamment les LLMs : qu’apprennent les LLMs ? Sur quoi ? Comment ? Peuvent-ils remplacer les approches formelles ?
- opposition majeure entre approches formalistes et statistiques : les approches formalistes sont créées par des humains de manière à interpréter, à isoler des phénomènes tandis que les LLMs sont entrainés sur de grosses quantités de données textuelles, sans réelle possibilité d’interprétation
- on fait néanmoins des progrès dans l’interprétation des LLMs : pourrons-nous un jour comprendre exactement comment ils fonctionnent ?
- conclusion de ce point en considérant que les approches formalistes et statistiques peuvent être également complémentaires

Autres points proposés

- les sorties des systèmes reposant sur des approches formalistes ne sont-elles pas justement trop “formelles” ?
- le formalisme face à l’évolution constante des langues -> problématique ?

Quelles métadonnées pour décrire la variation potentielle dans les données?

Prise de notes par Pablo Ruiz Fabo

- comment définir métadonnées

- infos avant de regarder le texte et que l’on peut extraire du texte
  - les deux sont métadonnées
- - données biographiques des locuteur·trice·s
    - scolarisation des locuteurs, bilingue ou multilingue
    - lieux du locuteur
    - âge / date de production

- les choix de métadonnées vont orienter l’analyse
- choix ou contraintes imposés sur le texte
  - pas corpus oral mais p-ê oralité représentée
  - choix éditoriaux

- données avant regarder texte
  - être conscient d’apriori possibles