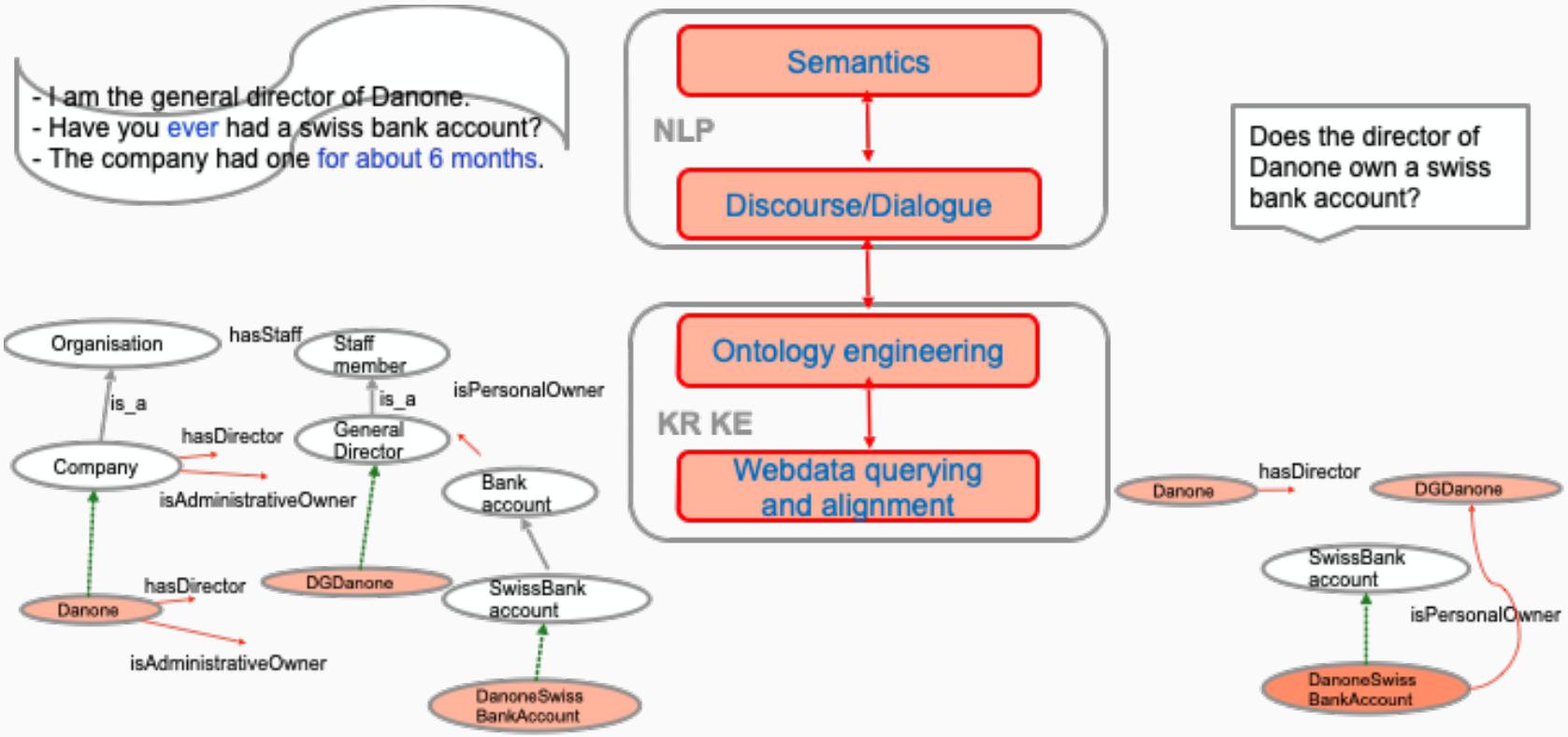
Deep learning pour le TAL et l'ingénierie des connaissances

Journée Deep Learning — Action CDIA, 1er juille 2021

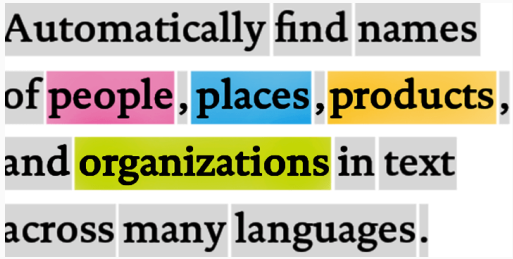
Philippe Muller

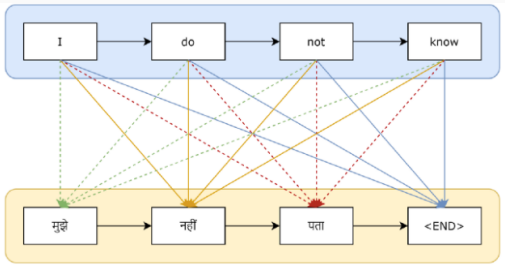
Thématiques de recherche Melodi: TAL IC

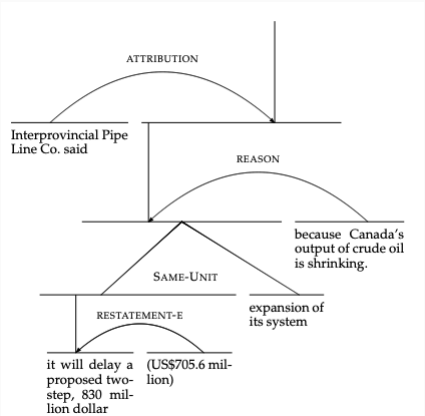


Problèmes classiques de TAL

- classification de texte
  sentiment, opinion,
  modération de contenus, ...
- extraction d'information
- génération de texte
  traduction, résumé, dialogue
- production de structures:
  syntaxe, structure de
  document,
  graphe de connaissances,







Spécificités du TAL

- Des données intrinsèquement séquentielles
  - $\neq$ ”bag of words”
  - dépendances pas uniquement locales
  - $\rightarrow$ structure sous-jacente
  - deux niveaux (au moins):
    - phrase / document
    - énoncé / conversation

Spécificités du TAL

- Des données intrinsèquement séquentielles
  - $\neq$ "bag of words"
  - dépendances pas uniquement locales
  - $\rightarrow$ structure sous-jacente
  - deux niveaux (au moins):
    - phrase / document
    - énoncé / conversation
- des entrées discrètes
  - unités de base: mot et/ou caractère
  - $\neq$ son, image, etc

Spécificités du TAL

- Des données intrinsèquement séquentielles
  - $\neq$ "bag of words"
  - dépendances pas uniquement locales
  - $\rightarrow$ structure sous-jacente
  - deux niveaux (au moins):
    - phrase / document
    - énoncé / conversation
- des entrées discrètes
  - unités de base: mot et/ou caractère
  - $\neq$ son, image, etc
- des entrées éparses et inter-dépendantes
  - vocabulaire large, fréquence en power law
  - liens sémantiques entre les mots

Des entrées discrètes et rares

- entrée sparse : beaucoup de données pour faire un modèle représentatif
- les symboles (mots) ont une sémantique / sont interdépendants

- *Apple a acheté Shazam en 2018*
  – *L'acquisition de Shazam par Apple a eu lieu il y a 3 ans.*
- relations de synonymie, antonymie, implication,
- importance du contexte

Sémantique distributionnelle

caractériser exhaustivement les liens de sens : impossible

la déduire de corpus = sémantique distributionnelle

The functional interplay of philosophy and	should, as a minimum, guarantee..
The rapid advance in	today suggests...
...calculus, which are more popular in	-oriented schools.

dans les années 2000 -> méthode non supervisées à partir de matrices de co-occurrences

petite révolution avec modèles à réseaux de neurones : 2008 puis 2013

Sémantique distributionnelle

caractériser exhaustivement les liens de sens : impossible

la déduire de corpus = sémantique distributionnelle

The functional interplay of philosophy and	science	should, as a minimum, guarantee..
The rapid advance in	science	today suggests...
...calculus, which are more popular in	science	-oriented schools.

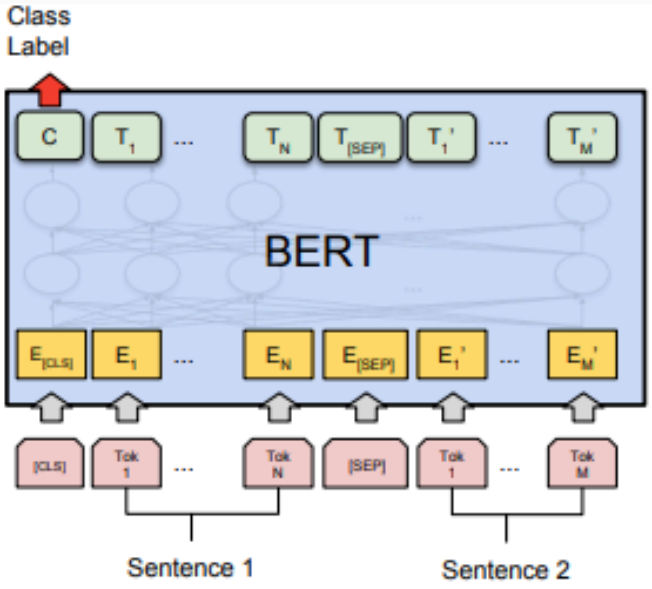
dans les années 2000 -> méthode non supervisées à partir de matrices de co-occurrences

petite révolution avec modèles à réseaux de neurones : 2008 puis 2013

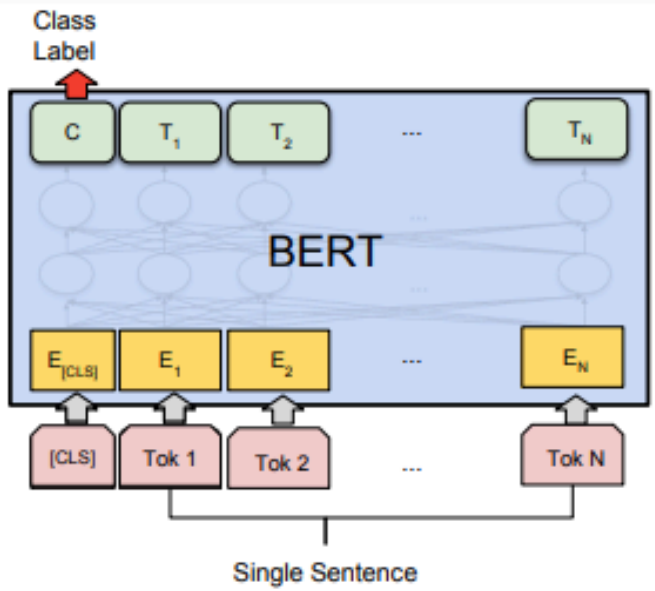
Apport du Deep Learning: embeddings prédictifs

- word2vec = modèle prédictif -> plus facile à contraindre
- mais ? hors contexte
- modèles contextuels: elmo, bert et co
  $\rightarrow$ sémantique dans le cadre de la phrase

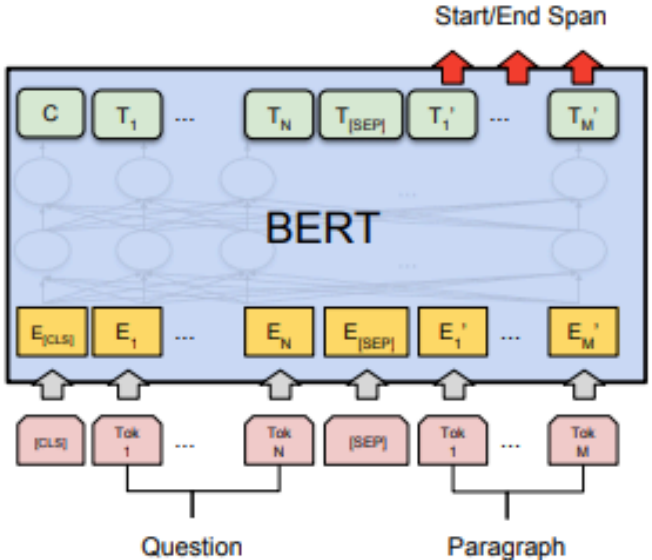
Apport du Deep Learning: préentrainement et transfert

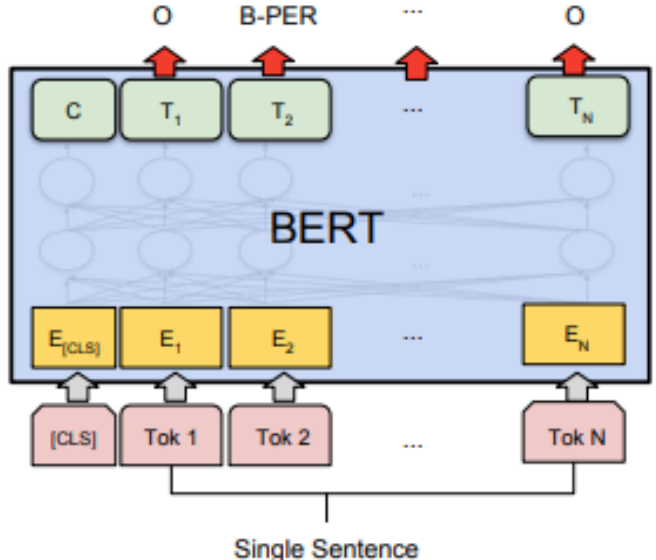


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

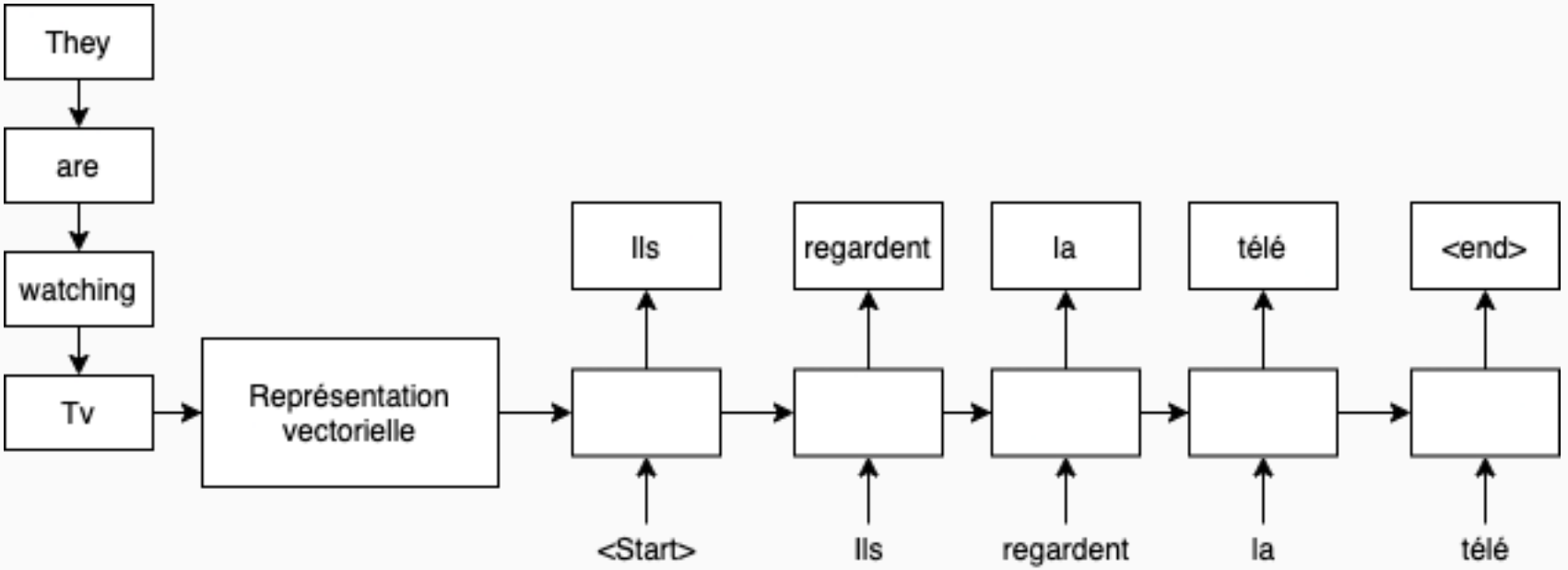


(b) Single Sentence Classification Tasks: SST-2, CoLA





Apport du Deep Learning: encodeur-décodeur



Problèmes habituels du DL

- explicabilité / transparence
- robustesse / biais
- manque de données

encore pire en TAL !

Problèmes habituels du DL

- explicabilité / transparence
- robustesse / biais
- manque de données

encore pire en TAL !

- parties de l'entrée interdépendantes; embeddings compliquent le lien avec l'entrée

Problèmes habituels du DL

- explicabilité / transparence
- robustesse / biais
- manque de données

encore pire en TAL !

- parties de l'entrée interdépendantes; embeddings compliquent le lien avec l'entrée
- aspect discret: instabilité

Problèmes habituels du DL

- explicabilité / transparence
- robustesse / biais
- manque de données

encore pire en TAL !

- parties de l'entrée interdépendantes; embeddings compliquent le lien avec l'entrée
- aspect discret: instabilité
- annotations complexes et coûteuses; datasets récents de mauvaise qualité

Contributions MELODI

- **classification de texte** : détection de l'ironie, langage sexiste, monitoring de crise, positionnement politique
  $\rightarrow$ problématiques éthiques et sociétales
- **extraction** / étiquettage de séquences:
  extraction d'information, segmentation de texte
- **génération**:
  génération de questions pour améliorer des systèmes de question/réponse
  résumé automatique
- **prédiction de structure**: analyse de structure de discours / de conversation
- **apprentissage de représentations** sémantiques: apprentissage intermédiaire pour transfert/fine-tuning en plusieurs étapes

Contributions MELODI

Questions transversales

- **intégrer connaissances** dans les modèles statistiques
- aspects **sociétaux** / éthiques
- **explicabilité** pour comprendre les biais : questions de robustesse, de fairness
- **supervision faible** : contrer la rareté de données annotées pour certains problèmes en combinant connaissances/données bruitées/apprentissage
- **apprentissage multi-tâches** : faire des ponts entre niveaux d'analyse; approches multi-lingues