Implication textuelle et Logiques de description

BEDARIDE Paul

Master Informatique
Spécialité Traitmement Automatique des Langues

30/06/2006

Encadrants : Carlos Areces
Claire Gardent

L'implication textuelle (IT)

T1 : *"Le diplomate n'a pas réussi à quitter Bagdad"*

T2 : *"Le diplomate est à Bagdad maintenant"*

? ? ? $\quad T1 \Rightarrow_T T2 \quad$ **? ? ?**

Motivation

Pourquoi l'implication textuelle ?

- intérêt théorique : tâche linguistique de base
- intérêt applicatif : nécessaire pour de nombreuses applications (systèmes de Q/R, résumé automatique, extraction d'information, ...)

Le Challenge RTE

Initié en 2005
But : comparer des systèmes à un standard étalon
Ressources : corpus de paires de textes annotés pour l'IT

Méthode d'évaluation :

- paires de petits extraits de textes (Texte-Hypothèse)
- annotations manuelles des paires : T $\Rightarrow_T$ H
- annotations automatiques des paires par les systèmes
- comparaison des résultats trouvés par les systèmes avec les annotations manuelles

Résultats :

- entre 52 et 75% de réponses justes
- meilleurs résultats pour les méthodes symboliques que pour les méthodes statistiques

Approche adoptée

Approche symbolique

Utilisation des logiques de description pour :

- approximer le sens de la langue naturelle
- détecter l'implication textuelle

Les logiques de description

Les logiques de description servent à représenter des connaissances

Définies par une signature contenant :

- des individus : `pierre`, `marie`, `rex`,...
- des concepts : `CHIEN`, `HUMAIN`, ...
- des rôles : `Acheteur`, `Vendeur`, ...

Les logiques de description (LDs)

Ainsi que par la syntaxe et la sémantique suivante :

Constructeur	**Syntaxe**	**Sémantique**
nom de concept	C	$C^{\mathcal{I}}$
top	$\top$	$\Delta^{\mathcal{I}}$
négation	$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
conjonction	$C_1 \sqcap C_2$	$C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$
disjonction	$C_1 \sqcup C_2$	$C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}$
quantificateur universel	$\forall R.C$	$\{d_1 \mid \forall d_2 \in \Delta^{\mathcal{I}}.(R^{\mathcal{I}}(d_1, d_2) \rightarrow d_2 \in C^{\mathcal{I}})\}$
quantificateur existentiel	$\exists R.C$	$\{d_1 \mid \exists d_2 \in \Delta^{\mathcal{I}}.(R^{\mathcal{I}}(d_1, d_2) \wedge d_2 \in C^{\mathcal{I}})\}$
nom de rôle	R	$R^{\mathcal{I}}$
rôles inverses	R^{-1}	$\{(d_1, d_2) \mid R^{\mathcal{I}}(d_2, d_1)\}$

Les logiques de description (LDs)

Séparation de la connaissance en deux parties :

- La A-Box : contenant des assertions de la forme :

$$\begin{aligned} \texttt{pierre} &: \texttt{HOMME} \\ \texttt{marie} &: \texttt{FEMME} \\ (\texttt{pierre}, \texttt{marie}) &: \texttt{Enfant} \end{aligned}$$

- La T-Box : contenant des axiomes terminologiques de la forme :

$$\begin{aligned} \texttt{HOMME} \sqcup \texttt{FEMME} &\sqsubseteq \texttt{HUMAIN} \\ \texttt{PÈRE} &\doteq \texttt{HOMME} \sqcap \exists \texttt{Enfant}.\texttt{HUMAIN} \end{aligned}$$

Raisonnement sur les LDs

Soit $\langle T, A \rangle$ une base de connaissances, C_1, C_2 des concepts, R un rôle et a et b des individus, on a :

- Subsomption : $\langle T, A \rangle \models C_1 \sqsubseteq C_2$
- Vérification d'instance : $\langle T, A \rangle \models a : C_1$
- Vérification de rôle : $\langle T, A \rangle \models (a, b) : R$
- Cohérence de la base de connaissances : $\langle T, A \rangle \not\models \bot$

Soit $\langle T, A \rangle$ une base de connaissances, A est saturée si pour chaque individu a, concept C et rôle R on a :

- $a : C$ si et seulement si $\langle T, A \rangle \models a : C$
- $(a, b) : R$ si et seulement si $\langle T, A \rangle \models (a, b) : R$

Logiques de description & Langue naturelle

Représentation des connaissances : T-Box

Représentation des textes : A-Box

Représentation des connaissances

Sémantique lexicale :

- synonymie, antonymie, hyperonymie, ...
- sens des verbes

Utilisation de deux bases de données linguistiques :

- WordNet
- FrameNet

WordNet

Base de données lexicale, traitant des relations comme :

- la synonymie : *chat*, *matou*
- l'antonymie : *présent*, *absent*
- l'hyperonymie : *animal*, *chat*
- la méronymie : *bras*, *corps*

Large couverture :

- 145000 noms
- 25000 verbes
- 30000 adjectifs
- 5000 adverbes

WordNet et Logiques de description

Utilisations de WordNet pour obtenir les relations qui existent entre les unités lexicales

Transformation de ces relations en des axiomes terminologiques (T-Box) :

WordNet	Description Logic	X	Y
X est un synonyme de Y	$\mathtt{X} \doteq \mathtt{Y}$	*"Chat"*	*"Matou"*
X est un antonyme complémentaire de Y	$\mathtt{X} \doteq \neg\mathtt{Y}$ $\mathtt{Y} \doteq \neg\mathtt{X}$	*"Présent"*	*"Absent"*
X est un antonyme scalaire de Y	$\neg\mathtt{X} \sqsubseteq \mathtt{Y}$ $\neg\mathtt{Y} \sqsubseteq \mathtt{X}$	*"Grand"*	*"Petit"*
X est un hyperonyme de Y	$\mathtt{Y} \sqsubseteq \mathtt{X}$	*"Animal"*	*"Chat"*
X est un méronyme de Y	$\mathtt{Y} \sqsubseteq \exists\mathtt{Composé\_de}.\mathtt{X}$	*"Bras"*	*"Corps"*

FrameNet

FrameNet est basé sur la sémantique des cadres et associe à chaque foncteur sémantique de la langue (e.g. verbe, noms, ...) :

- un cadre (i.e, un concept)
- un ensemble d'éléments cadres (i.e., un ensemble de rôles thématiques)

Exemple pour le cadre *transaction commerciale* :

- mots : acheter vendre payer marchander coûter dépenser
- éléments cadres : acheteur vendeur marchandise prix

Large couverture :

- 792 cadres conceptuels
- 9894 unités lexicales

Représentation des textes

Approximation du sens :

- prédicats-arguments
- modificateurs
- adjectifs
- négation

Omission :

- quantification
- modalité
- relation rhétoriques

Les dépendances prédicat-arguments

Représentation basée sur la sémantique de Davidson :

- verbe (prédicat) représenté par un concept
- lié à d'autres individus (arguments) par des rôles

Utilisation de FrameNet pour choisir les noms des concepts et des rôles associés à chaque verbe

Pour le verbe *vendre* on a donc :

- le concept associé `TRANSACTION_COMMERCIALE`
- les rôles associés `Acheteur`, `Vendeur`, `Marchandise`, `Prix`

Exemple : dépendances prédicat-arguments

"Jean achète du nutella au supermarché pour 2 euros"

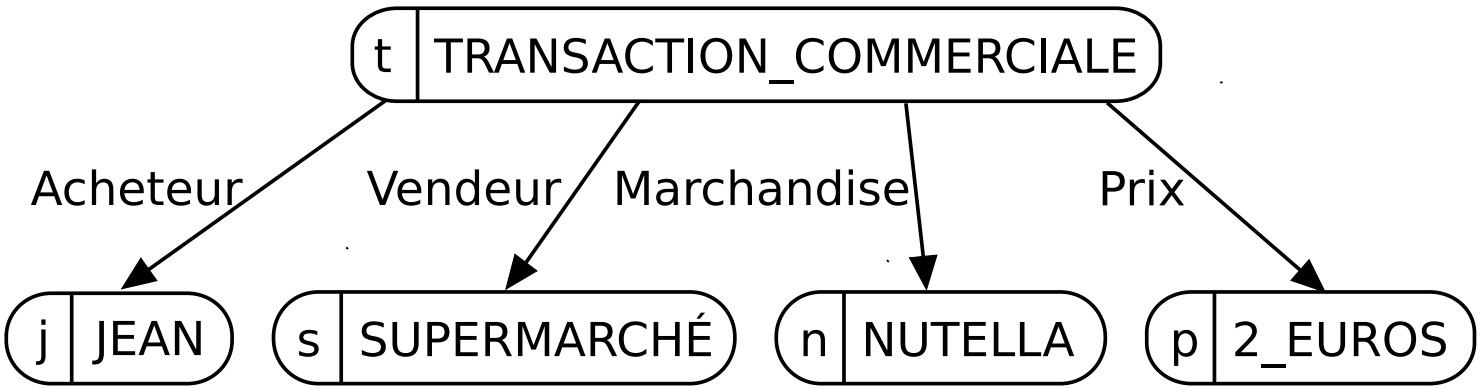
t : TRANSACTION_COMMERCIALE

j : JEAN — (t, j) : Acheteur

s : SUPERMARCHÉ — (t, s) : Vendeur

n : NUTELLA — (t, n) : Marchandise

p : 2_EUROS — (t, p) : Prix



Les modificateurs

Altération du sens d'un verbe par des modificateurs de lieu, de temps ou de manière.

Différences des modificateurs par rapport aux arguments :

- ils sont indépendants du verbe auquel ils s'appliquent
- un verbe peut avoir plusieurs modificateurs du même type

On représente les modificateurs par des concepts ajoutés à l'individu représentant le verbe

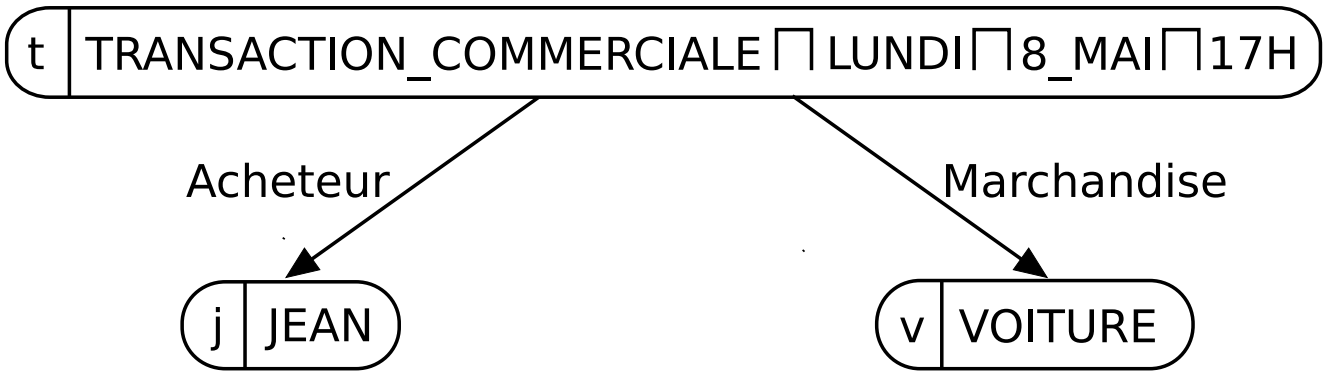
Exemple : modificateurs

"Jean a acheté une voiture le Lundi 8 mai à 17h"

$t : \texttt{TRANSACTION\_COMMERCIALE} \sqcap \texttt{LUNDI} \sqcap \texttt{8\_MAI} \sqcap \texttt{17H}$

$j : \texttt{JEAN} \qquad (t,j) : \texttt{Acheteur}$

$v : \texttt{VOITURE} \quad (t,v) : \texttt{Marchandise}$



Les adjectifs

Approche Montague et Davidson :

adjectifs représentés comme prédicats à un argument

Problème :

il faut que *"Le chat rose"* $\Leftrightarrow_T$ *"Le chat est rose"*

On représente donc les deux phrases par :

$$c : \texttt{CHAT} \sqcap \texttt{ROSE}$$

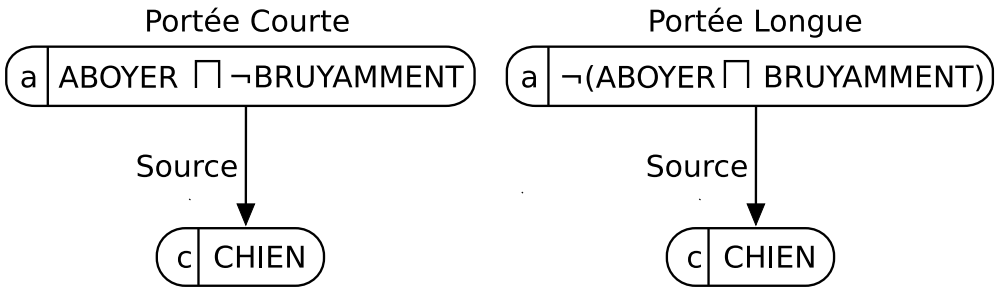
Le verbe *être* est considéré comme un verbe à part

La négation

La négation peut avoir plusieurs portées
Pour la phrase *"Le chien n'aboie pas bruyamment"*, on a les interprétations :

- portée courte : le chien aboie mais il n'aboie pas bruyamment
- portée longue : on ne sait pas si le chien aboie mais si il aboie il ne le fait pas bruyamment

On choisit arbitrairement la portée longue en espérant qu'elle soit l'interprétation voulue dans la majorité des cas



L'algorithme de détection

Une A-Box peut être représentée par un ou plusieurs graphes orientés

Pour qu'un texte T implique un texte H il faut que le graphe de H soit un sous-graphe du graphe de T

Pour que cela soit vrai, il faut que :

- pour chaque individu de H ait une correspondance dans T
- chaque rôle de H existe dans T via la correspondance établie précédemment

L'algorithme de détection

Propriété

Soit Tn l'ensemble des noeuds de T et Hn l'ensemble de ceux de H. Une condition nécessaire pour que H soit un sous-graphe de T, est qu'il faut qu'il existe une fonction f qui à chaque élément x de Hn fait correspondre un élément y de Tn, tel que l'ensemble des concepts associés à x soit un sous-ensemble de ceux associés à y.

Propriété

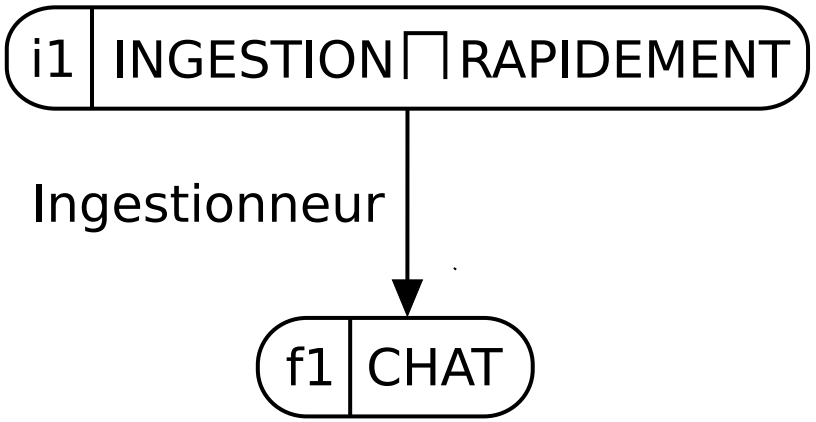
Soit Ta l'ensemble des arcs de T, et Ha l'ensemble des arcs de H. Les arcs sont représentés par des triplets (noeud source, noeud cible, nom de l'arc). On a alors : H est un sous graphe de T, si $(x, y, R) \in Ha$ *implique que* $(f(x), f(y), R) \in Ta$.

Exemple simple

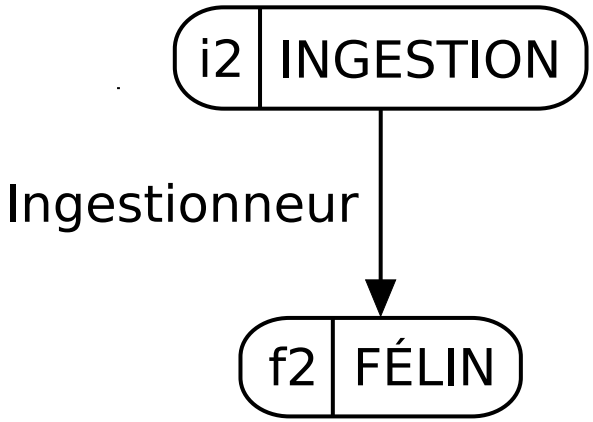
En utilisant le T-Box :

$$\texttt{CHAT} \sqsubseteq \texttt{FÉLIN}$$

T : "le chat mange rapidement"



H : "le félin mange"

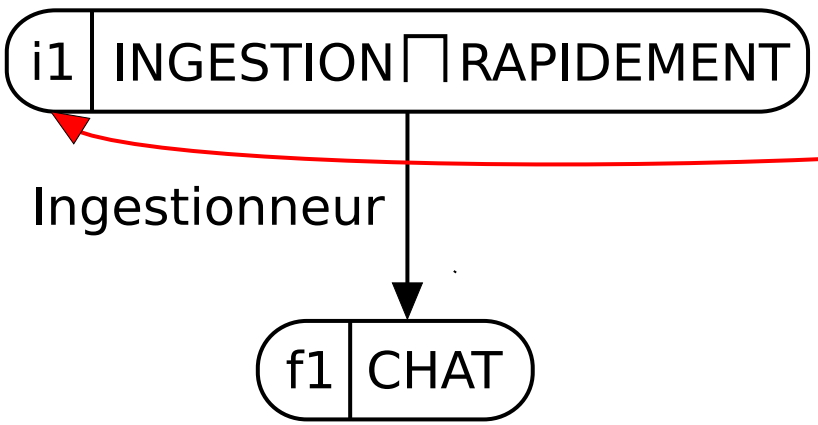


Exemple simple

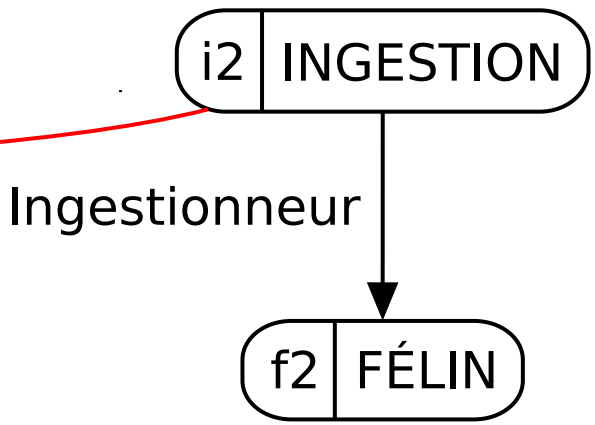
En utilisant le T-Box :

$$\text{CHAT} \sqsubseteq \text{FÉLIN}$$

T : "le chat mange rapidement"



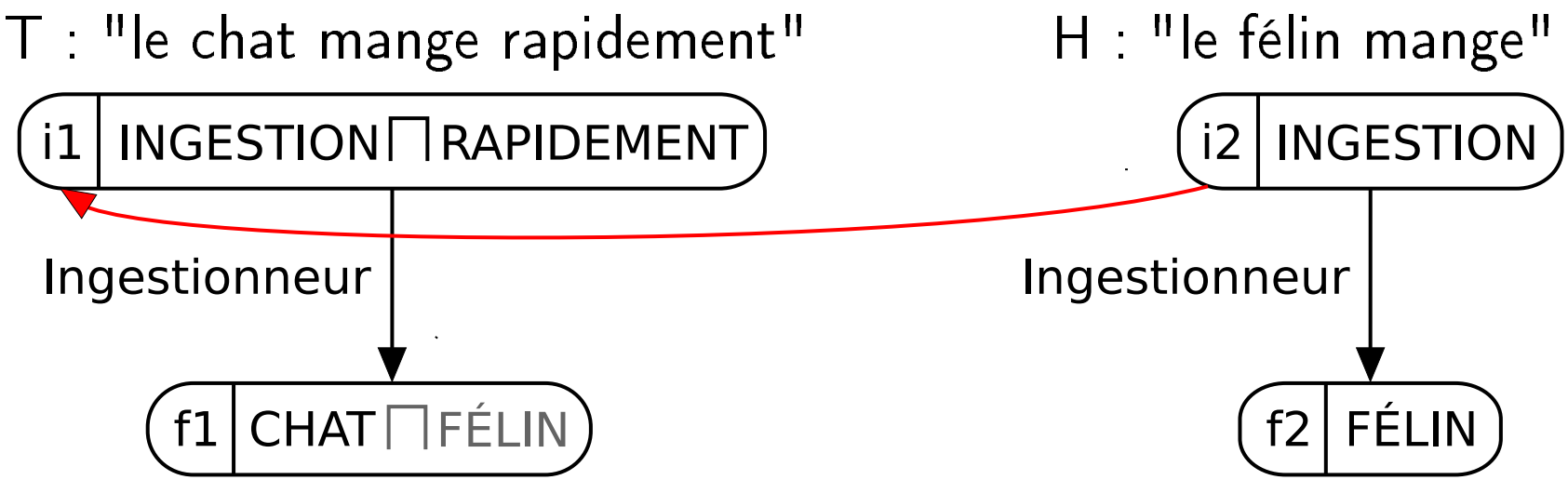
H : "le félin mange"



Exemple simple

En utilisant le T-Box :

$$\text{CHAT} \sqsubseteq \text{FÉLIN}$$



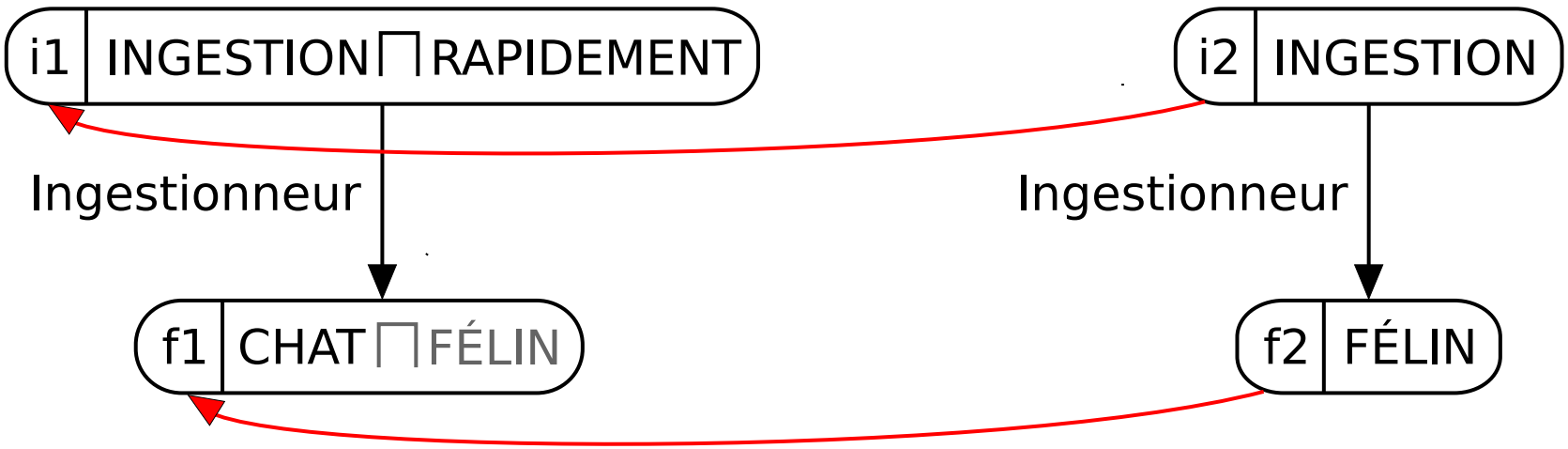
Exemple simple

En utilisant le T-Box :

$$\text{CHAT} \sqsubseteq \text{FÉLIN}$$

T : "le chat mange rapidement"

H : "le félin mange"



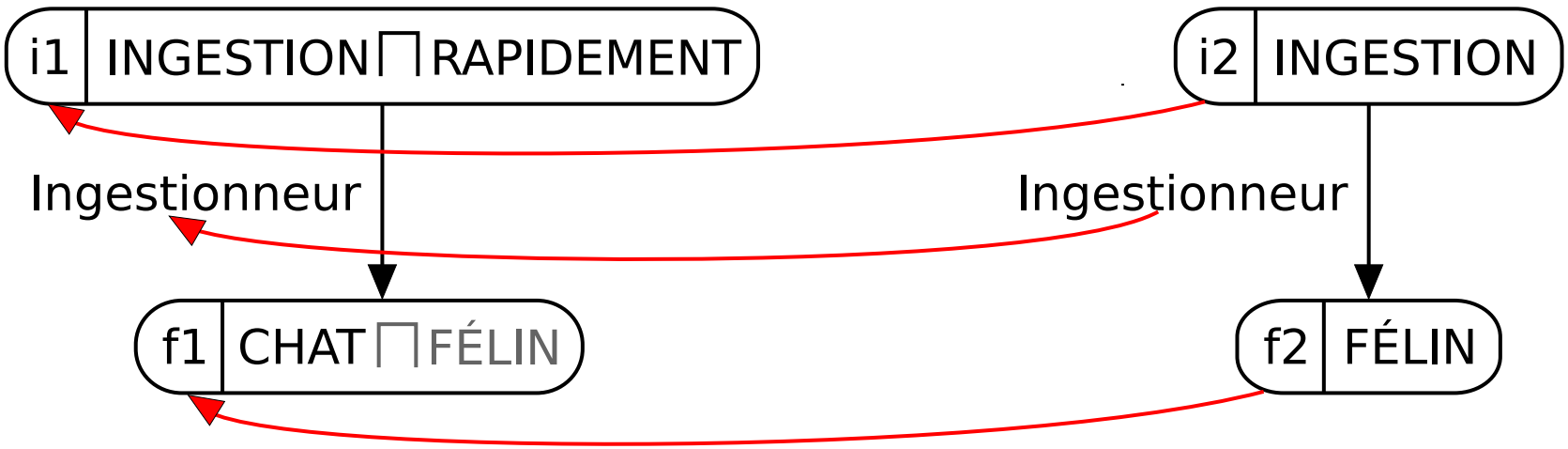
Exemple simple

En utilisant le T-Box :

$$\text{CHAT} \sqsubseteq \text{FÉLIN}$$

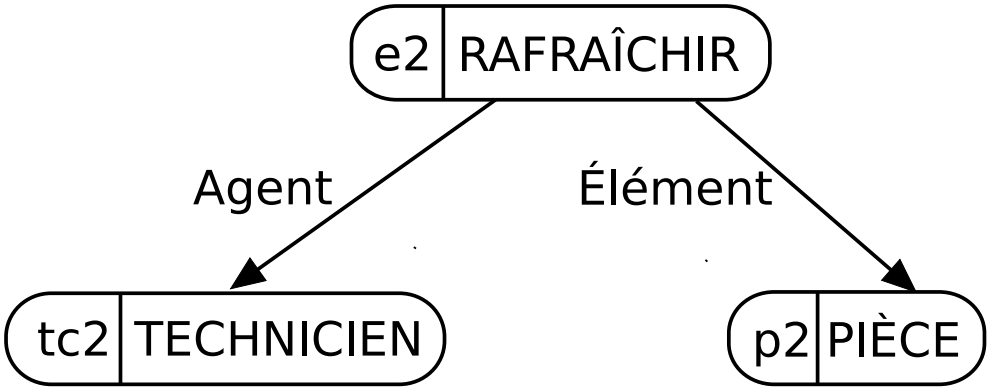
T : "le chat mange rapidement"

H : "le félin mange"

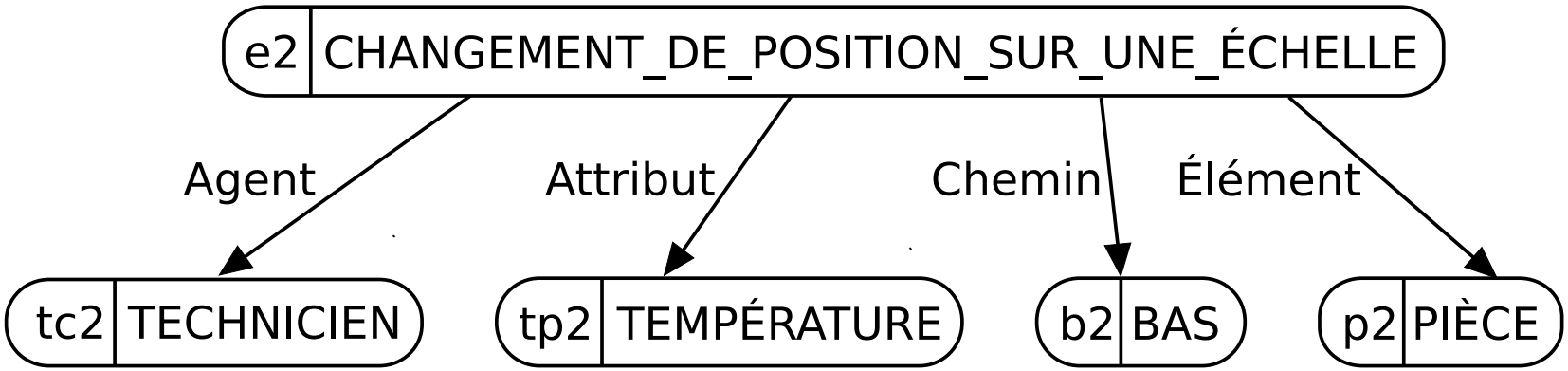


Exemple avec entité implicite

"Le technicien rafraîchit la pièce"



"Le technicien baisse la température de la pièce"



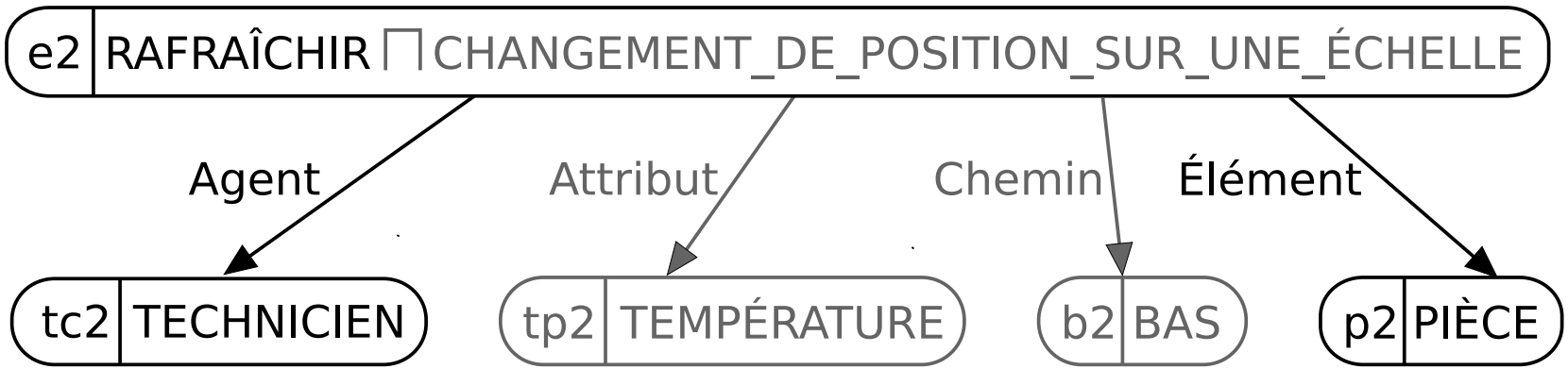
Exemple avec entité implicite

En utilisant la T-Box suivante :

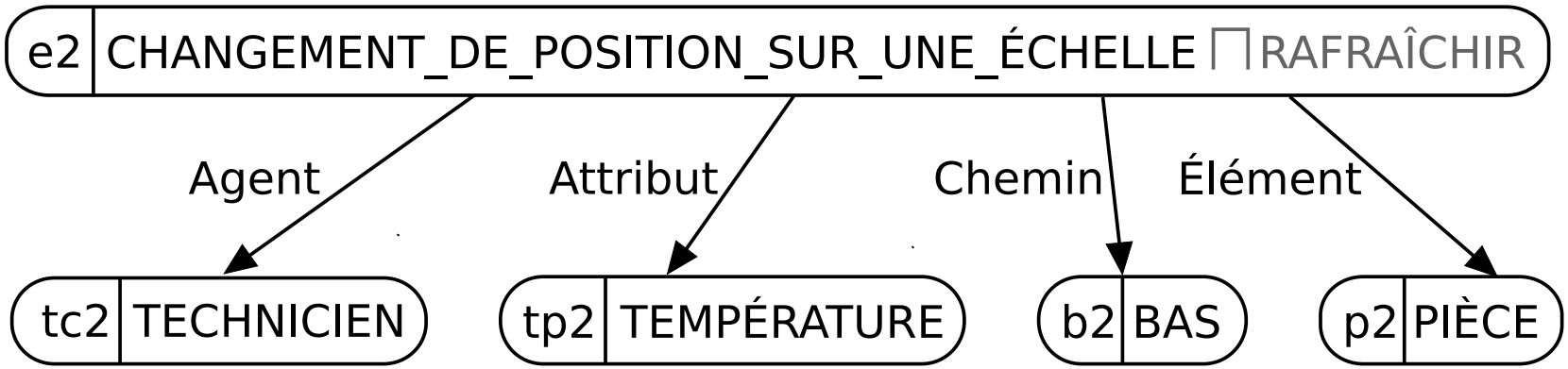
$$\begin{aligned} \texttt{RAFRAÎCHIR} \quad \doteq \quad & \texttt{CHANGEMENT\_DE\_POSITION\_SUR\_UNE\_ECHELLE} \sqcap \\ & \exists \texttt{Attribut.TEMPÉRATURE} \sqcap \\ & \exists \texttt{Chemin.BAS} \end{aligned}$$

Exemple avec entité implicite

"Le technicien rafraîchit la pièce"



"Le technicien baisse la température de la pièce"



L'implémentation

Utilisation du prouveur RACER utilisant la logique de description $\mathcal{ALCQHI}_{\mathcal{R}^+}(\mathcal{D}^-)$

RACER implémente une méthode des tableaux optimisée

RACER est l'un des prouveurs de logique de description les plus avancés

Utilisation de python, car c'est un langage de prototypage

Utilisation de XML pour les fichiers d'entrée

Les tests

Système testé sur un mini-corpus

Les représentations sont faites manuellement, mais pourraient être automatisées

Exemples choisis pour illustrer la capacité du système à traiter :

- des phénomènes linguistiques
- des phénomènes théoriques liés à l'utilisation des LDs

Exemples qui fonctionnent

Phénomènes linguistiques :

- La synonymie :
  "Jean a un vélo" $\Leftrightarrow_T$ "Jean a une bicyclette"
- L'antonymie :
  "Le chat est grand" $\Rightarrow_T$ "Le chat n'est pas petit"
- Les modificateurs :
  "Jean court rapidement" $\Rightarrow_T$ "Jean court"
- La négation :
  "Jean ne court pas " $\Rightarrow_T$ "Jean ne court pas rapidement"
- Le sens des verbes :
  "Le technicien rafraîchit la pièce" $\Leftrightarrow_T$ "Le technicien baisse la température de la pièce"

Exemples qui fonctionnent

Phénomènes théoriques :

- individus ayant plusieurs correspondances possibles :
  "Un chat monte sur un pommier et un autre chat monte sur un oranger" $\Rightarrow_T$ "Un animal monte sur un pommier et un autre animal monte sur un oranger"
- individus implicites :
  "Jean est père" $\Leftrightarrow_T$ "Jean a un enfant"

Les limites

Pas d'incohérence linguistique par rapport à notre représentation

Les entités implicites qui ne sont pas directement liées à une entité explicite

Les phrases que l'on ne peux pas représenter pour le moment :

- quantificateurs
- modalité
- ...

Les quantificateurs

Ne peuvent pas être représentés par une A-Box

$\Rightarrow$ donc représentation par une T-Box

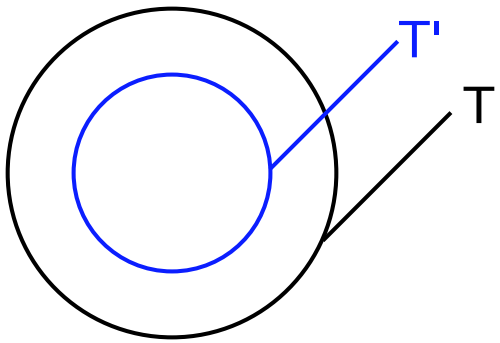
Plusieurs problèmes appararaissent :

- mélange connaissances de base et informations ajoutées par la phrase :
  "Tous les chats sont gris" $\not\Rightarrow_T$ *"Un chat est un animal"*
- impossible de séparer connaissances de base et informations ajoutées par la phrase :
  "Tous les félins mangent" $\Rightarrow_T$ *"Tous les chats mangent"*

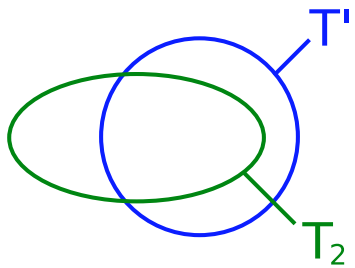
Modèlisation de la pertinance

Solution pour que $T_1 \Rightarrow T_2$, avec la T-Box T :

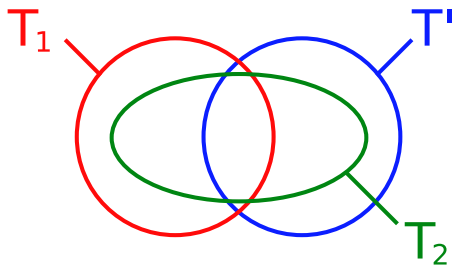
$T' \subseteq T$



$T' \not\models T_2$



$T' + T_1 \models T_2$



Une question persiste, comment trouver T' de manière efficace ?

Modèlisation de la pertinance

Propriété

On a $e_1 \Rightarrow_T e_2$, avec e_1 et e_2 des représentations sémantiques de deux textes, et $\Rightarrow_T$ qui représente l'implication textuelle et T la T-Box représentant notre connaissance de base, si et seulement si il existe T' tel que $T' \subseteq T$, et que $T' \not\models e_2$ et $T' + e_1 \models e_2$ (et donc que $T' \not\models e_1$ qui peut être déduit des deux formules précédentes).

Contributions

Une représentation des textes en logique de description

Représentation des connaissances lexicales de WordNet et FrameNet en logique de description

Un algorithme de détection de l'implication textuelle fondé sur les logiques de description

Un prototypage et une évaluation sur un ensembles de textes (60 cas différents)

Perspectives :

- calcul automatique des représentations sémantiques
- exploration de logiques plus expressives :
  - A-Box Boolennes
  - logiques hybrides
- passage à l'échelle :
  - conversion automatique de FrameNet et WordNet