Annotation déterministe un système à base de règles en deux étapes

Sommaire

Résumé

Ce chapitre présente le système d'annotation déterministe utilisé pour produire de manière semi-automatique les annotations de référence sur les données du corpus MEDIA. *Il comporte 2 étapes distinctes : une instanciation des frames et frame-éléments par dètection de motifs, puis une inférence à base de règles logiques des composants manquants et des relations entre les frames.*

6.1 Introduction

L'application de modèles stochastiques à la composition sémantique impose l'utilisation de données d'apprentissage pour construire les tables de probabilités conditionnelles supportant ces modèles.

L'annotation manuelle en frames de l'intégralité du corpus n'était pas envisageable, tant pour des raisons de coûts que de délai de disponibilité. Un processus d'annotation en deux étapes à base de règles a donc été développé pour produire les annotations en frames des données d'apprentissage.

La première étape du processus, décrite en 6.2, utilise les modèles définissant les frames pour déclencher l'instanciation de frames et de leurs FE selon que LU et CU sont rencontrés dans les données à annoter. La seconde étape, décrite en 6.3, compose les frames et FE proposés durant l'étape précédente grâce à l'application d'une série de règles logiques. Ce processus est progressivement enrichi pour améliorer ses performances.

6.2 Reconnaissance de modèles

Les modèles définissant les frames et leurs FE, présentés en 5.3, sont composés d'unités conceptuelles (CU) et lexicales (LU). La présence de ces CU et/ou LU dans les données à annoter déclenche l'instanciation des frames et FE auxquels ils sont associés.

Aux paires concept-valeur annotées dans le corpus MEDIA peuvent être attachés un mode (affirmatif, négatif, interrogatif ou optionnel) et un spécifieur (définissant les relations entre concepts). Ces informations ne sont pas reprises dans la définition des frames et FE pour préserver leur généricité. Seuls les unités lexicales composant le message et les concepts de base annotés dans le corpus MEDIA servent de support à la définition des objets sémantiques frames et FE.

L'algorithme de *pattern-matching* développé pour décider l'instanciation des objets sémantiques intègre plusieurs options :

- la prise en compte des LU peut être liée ou non à la présence des CU associés à l'objet sémantique à instancier ;
- les FE instanciés peuvent ou non être automatiquement reliés aux frames mères candidates ;
- un segment lexical associé à un CU peut être autorisé ou non à déclencher l'instanciation de plusieurs objets sémantiques.

Les frames et FE produits lors de cette phase d'instanciation peuvent être vus comme des fragments isolés de représentation sémantique du message du locuteur. Les seuls liens relationnels établis entre les frames et FE sont les liens d'appartenance d'un FE à une frame. La majorité des frames et FE instanciés lors de cette étape est composée d'objets sémantiques concrets, déclenchés par la présence de LU ou de CU identifiés

dans leurs modèles. Certaines frames abstraites de haut niveau, essentiellement sous-tendues par la présence d'autres frames et/ou FE, sont rarement instanciées lors de cette étape.

Le message *"réserver un hôtel"*, fréquemment rencontré dans le corpus MEDIA, est ainsi annoté lors de cette première phase à l'aide des deux seules frames `RESERVE` et `HOTEL`, non reliées entre elles (figure 6.1).

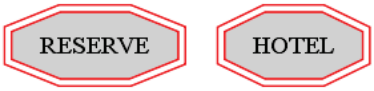

FIGURE 6.1 – *Annotation initiale par reconnaissance de modèles du message "réserver un hôtel"*

La frame `LODGING`, définissant la notion globale d'hébergement, va permettre de lier ces frames pour obtenir une représentation sémantique consistante du message sous la forme donnée figure 6.2.

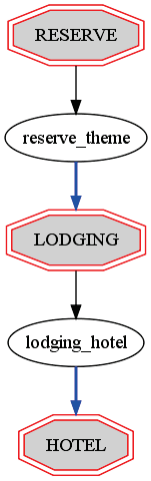

FIGURE 6.2 – *Annotation complète du message "réserver un hôtel"*

Cette représentation par arbre sémantique est obtenue lors de la deuxième étape du processus, grâce à l'application de règles d'inférences. Cette étape est décrite dans la section 6.3.

6.3 Règles d'inférences

La seconde étape soumet les frames et FE instanciés au cours de l'étape précédente à l'application d'une série de règles logiques. Frames et FE déterminent les valeurs de vérités des prédicats de ces règles. Selon ces valeurs de vérité, des frames et des FE peuvent être créés, supprimés, modifiés ou reliés.

La création de frames et FE concerne essentiellement les frames abstraites comme illustré par l'exemple de la frame `LODGING` présenté en 6.2. La suppression et la modification de frames et FE sont motivées par la présence d'objets redondants, instanciés par la présence de redondances dans de nombreux messages oraux du type *"je veux un hotel... euh... un hotel proche de la mer."*.

Enfin, la dernière fonction de cette étape d'annotation est la création de liens entre les frames et les FE. La création de ces liens est justifiée par la hiérarchie présente au sein des frames, traduite par l'aptitude de certains FE à prendre des frames pour valeurs. Les fragments sémantiques obtenus à l'issue de l'étape de reconnaissance des modèles sont composés. Les liens instanciés par les règles logiques permettent d'obtenir un arbre sémantique représentant le sens du message du locuteur.

Le langage de programmation logique `Prolog` (Colmerauer et Roussel, 1996), basé sur le calculs des prédicats du premier ordre, a été utilisé pour réaliser toutes les inférences logiques. Un programme `Prolog` se compose d'une base de faits et de règles logiques décrivant les relations entre des faits potentiels. Cette base rassemble les *connaissances* du programme. Les faits sont représentés par des prédicats affirmatifs et les règles logiques s'expriment simplement sous la forme `conclusion SI condition`.

L'exécution d'un programme `Prolog` consiste à soumettre une requête à l'interpréteur. `Prolog` cherche à prouver que la requête est vraie en analysant chaque règle et prouvant qu'elle est vérifiée.

L'implémentation des règles de composition sémantique utilisées dans cette étape d'annotation est réalisée sous `SWI-Prolog` (Wielemaker, 2003).

Une des règles `Prolog` s'écrit sous la forme :

```
do_link(LODH,H) :-
  is_fe(lodging_hotel,LODH),
  is_concept_of(hotel,LODH),
  is_fr(HOTEL,H).
```

où le symbole `:-` signifie "`SI`" [1].

Dans cet exemple, la règle crée un lien entre le FE `lodging_hotel` et la frame `HOTEL` sous la condition que le FE `lodging_hotel` ait été déclenché par la présence du concept `hotel` associé au message. Le FE `lodging_hotel` prend la frame `HOTEL` pour valeur, construisant ainsi une branche de l'arbre sémantique représentant le message.

Environ 100 règles sont actuellement appliquées. Agissant sur les frames et les FE, elle peuvent prendre en compte la présence de mots et de concepts. L'ordre dans lequel les frames et les FE ont été instanciés avant d'être soumis au programme Prolog n'influence pas les inférences réalisées.

L'inférence logique est appliquée itérativement, chaque sortie fournissant les faits soumis à l'étape de résolution suivante. L'itération peut être, au choix, poursuivie jusqu'à

1. Expression du modus-ponens.

ce qu'aucune modification ne soit plus inferée ou un nombre pré-défini de fois.

6.4 Évaluation

Pour évaluer les performances des système de composition sémantique développés dans ce travail, la préparation d'un ensemble de données de test a été nécessaire. Les 3005 tours de parole utilisateur du lot de test MEDIA, manuellement transcrits et annotés en concepts de base, ont été automatiquement annotés en frames et FE par le système à base de règles puis corrigés manuellement par un linguiste expert.

Etant données l'ampleur de la tâche et la disponibilité d'un unique linguiste expert, il n'a pas été possible d'obtenir d'IAg sur les annotations en frames et FE. Les 3005 tours de parole utilisateur manuellement transcrits, annotés en concepts de base et en frames et FE composent l'ensemble de test nommé "REF". Les annotations produites par le système d'annotation en deux étapes à base de règles sur les 3005 tours de parole du lot de test ont été évaluées par comparaison à cet ensemble "REF".

Les performances de ce système ont été mesurées en termes de précision, rappel et F-mesure. La précision est le nombre de frames, FE ou liens corrects proposés par le système rapporté au nombre total de frame, FE ou liens proposés par le système. Le rappel est le nombre de frames, FE ou liens corrects proposés par le système divisé par le nombre total de frames, FE ou liens contenus dans l'annotation de référence. La F-mesure est la moyenne harmonique standard de la précision et du rappel.

Le tableau 6.1 présente les résultats obtenus par le système d'annotation à base de règles sur les 3005 tours de parole de référence.

Les différents niveaux d'évaluation sont :

- **Frames** : les hypothèses de frames sont considérées comme correctes dès lors que les frames correspondantes sont présentes dans la référence (sans prise en compte des FE qui les composent).
- **FE** : les hypothèses de FE sont considérées comme correctes dès lors que les FE correspondants sont présents dans la référence.
- **FE{Frames}** : seules les hypothèses de FE appartenant à des hypothèses de frames correctes sont examinées. L'ensemble de référence est restreint aux FE appartenant aux frames correspondantes dans la référence.
- **Liens** : les hypothèses de liens sont considérées comme correctes dès lors que les liens correspondants sont présents dans la référence.
- **Liens{Frames}** : seules les hypothèses de liens reliant des hypothèses de frames et FE correctes sont examinées. L'ensemble de référence est restreint aux liens reliant les frames et FE correspondants dans la référence.

Les résultats obtenus par le système d'annotation à base de règles, avec des F-mesures toutes supérieures à 90%, confirme sa fiabilité et sa capacité à produire sur l'ensemble du corpus MEDIA des données d'apprentissage consistantes.

Le nombre total de frames, FE et liens présents sur les tours de parole de l'ensemble REF ansi que sur l'ensemble de test MEDIA et les dialogues d'entraînement (train) annotés grâce au système à base de règles sont indiqués dans le tableau 6.2.

		Frames	FE	FE{Frames}	Liens	Liens{Frames}
Système	p	0.98	0.97	1.00	0.95	1.00
à base de règles	r	0.99	0.94	0.95	0.86	0.88
	F-m	**0.98**	**0.96**	**0.97**	**0.90**	**0.94**
	$\bar{p}$	0.99	0.99	1.00	0.99	1.00
	$\bar{r}$	0.99	0.97	0.97	0.94	0.95
	$\overline{F\text{-}m}$	**0.99**	**0.97**	**0.98**	**0.95**	**0.96**

TABLE 6.1 – *Précision ($\bar{p}$), Rappel ($\bar{r}$) et F-mesure ($\bar{F}$-m), précision moyenne ($\bar{p}$), rappel moyen (r) et F-mesure moyenne ($\overline{F\text{-}m}$) obtenus par le système d'annotation à base de règles sur les 3005 tours de parole de l'ensemble de test* MEDIA.

Ensemble	Annotation	Frames	FE	Liens
TRAIN MEDIA	à base de règles	33923	35101	15828
	nb moyen par tour de parole	2,83	2,93	1,32
TEST MEDIA	à base de règles	8315	8680	3845
	nb moyen par tour de parole	2,77	2,89	1,28
TEST MEDIA	manuelle	8241	9020	4251
	nb moyen par tour de parole	2,74	3,00	1,41

TABLE 6.2 – *Nombre de frames, FE et liens présents dans les ensembles d'apprentissage et de test* MEDIA*, annotés grâce au système à base de règles et après correction manuelle.*

On remarque que les arbres associés aux tours de parole sont d'ordre peu élevé avec une moyenne de moins de 3 frames et 3 FE par tour. Leur taille est également restreinte avec moins de 1,5 lien par tour de parole. La comparaison entre les résultats obtenus sur le test MEDIA annoté manuellement et ceux obtenus sur la version annotée par le système à base de règles indique une tendance du système déterministe à insérer des frames et à omettre des FE et des liens.

6.5 Conclusion

La mise en place de modèles stochastiques dans un système applicatif nécessite l'emploi des tables de probabilités conditionnelles qui leur sont associées. Les valeurs numériques rassemblées dans ces tables de probabilités conditionnelles doivent donc être apprises sur des ensembles de données comportant les informations que l'on souhaite étudier.

Le corpus MEDIA n'étant pas annoté en frames et FE, un système à base de règles en deux étapes a été développé pour permettre l'annotation des données d'apprentissage.

Ce système crée tout d'abord frames et FE par reconnaissance de modèles puis associe ces objets sémantiques lors d'une étape d'inférence logique.

Évalué sur les données de test du corpus MEDIA, les données annotées automatiquement s'avèrent suffisamment fiables pour être utilisées comme données d'apprentissage par les systèmes stochastiques.

Génération des fragments sémantiques