L'élaboration et l'étude du corpus

1) Introduction

L'importance du RTE dans le TALN a poussé les chercheurs à s'investir dans ce domaine et à explorer différents chemins pour parvenir à détecter et à classifier différents types d'inférences.

Dans les chapitres précedents, nous avons d'abord étudié les groupes travaillant sur la reconaissance de l'inférence textuelle et nous avons remarqué qu'aucun groupe n'utilisait l'inférence temporelle dans son système. Dans le chapitre précedent nous avons étudié le temps dans la langue et nous avons remarqué que les groupes travaillant sur l'inférence temporelle se base sur l'amélioration des détéctions des relations temporelles existantes entre évènements et expressions temporelles mais ils n'essayaient en aucun cas d'intégrer leurs travaux a un systéme d'inférence textuelle.

Afin de répondre au manque de l'inférence temporelle dans le RTE, notre objectif est d'intégrer le système de détéction d'inférence temporelle dans un systéme d'inférence textuelle. Pour cela, nous avons l'obligation d'étudier les relations temporelles qui peuvent exister entre deux ségments de textes à travers un corpus que nous avons élaboré. Ceci nous a permis de distinguer différents types d'inférences.
Nous allons montrer tout au long de ce chapitre comment nous avons concrétisé ces différents objectifs.

2) L'élaboration du corpus

La première étape à entreprendre consiste à créer le corpus constitué de paires de textes et hypothèses (T-H) qui correspond à des informations collectées à travers le web dans des domaines différents. Nous avons choisi d'établir notre corpus en langue anglaise car jusqu'à nos jours les recherches les plus abouties sur l'inférence temporelle et aussi sur le RTE sont en langue anglaise.

Pour cela, nous avons choisi d'utiliser le corpus de questions élaborées pour le test par la compagne d'évaluation des systèmes de recherches d'informations (clef[9]) pour l'année 2006.

[9]Le lien du challenge clef : http://www.elda.org/article225.html

Le challenge CLEF est crée en 2000 pour fournir une infrastructure visant à soutenir le développement, d'essai et d'évaluation des systèmes de cross-langue de recherche d'information dans plusieurs langues européennes (Français, Italien, Allemand).

Pour pouvoir développer et évaluer notre système, nous avons sélectionné des questions portant sur des événements temporels et nous avons soumis ces questions au système de question-réponse answerbus [10] disponible sur le web. Nous avons récupéré les réponses correspondantes et nous les avons modifiées pour obtenir l'inférence souhaitée. Nous avons aussi transformé les questions à l'affirmatif.

Nous illustrons ces démarches par l'exemple montré ci-dessous :

La question numéro 13 du corpus de test de challenge clef 2006:

In what year did the catastrophe in Chernobyl happen?

La requête va être mise dans le système de question réponse Answerbus. Le résultat est montré ci-dessous :

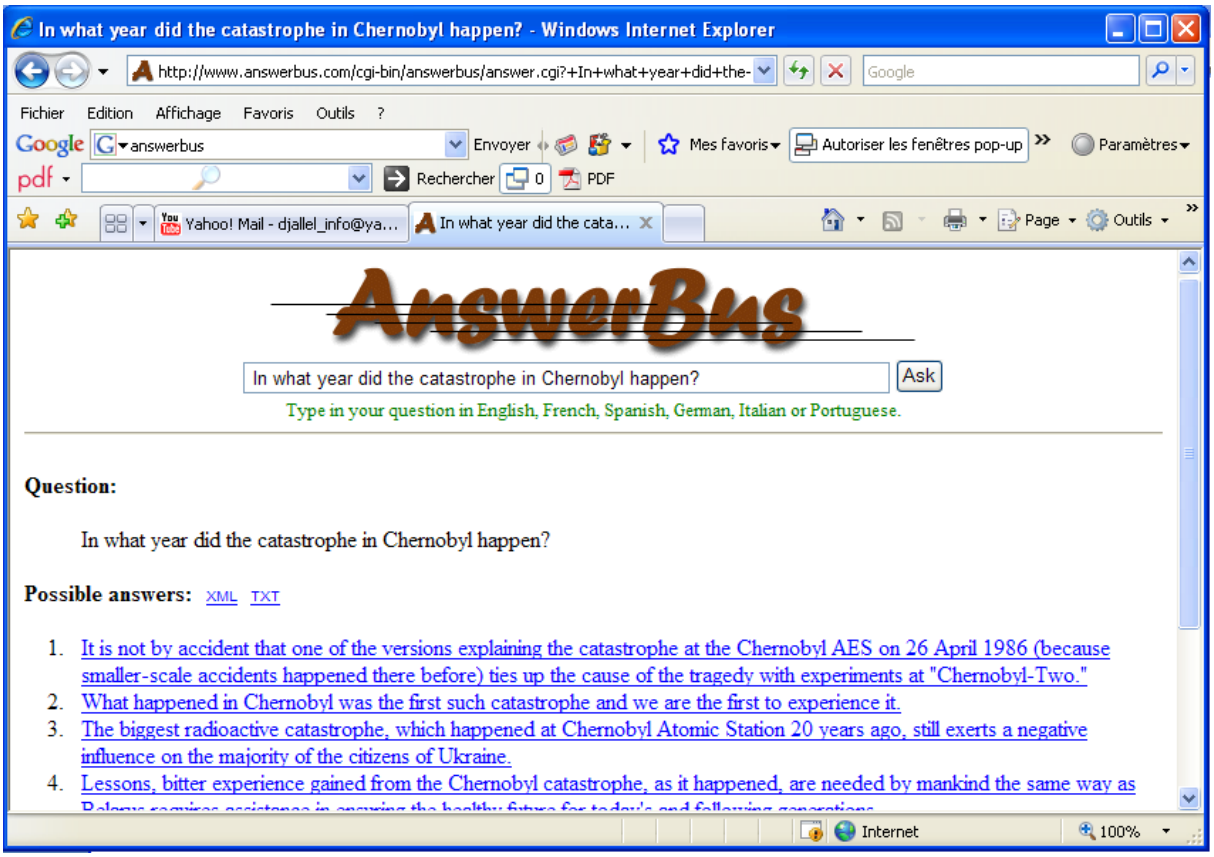

Figure 3.1 : Représente la réponse du système AnswerBus

Nous choisissons la première réponse donnée par le système qui est :

[10] http://www.answerbus.com

H: It is not by accident that one of the versions explaining the catastrophe at the Chernobyl AES on 26 April 1986 (because smaller-scale accidents happened there before) ties up the cause of the tragedy with experiments at "Chernobyl-Two" .

Aussi, nous transformons la question en affirmatif en répondant à la question. Comme résultat nous avons la réponse suivante :

T: the catastrophe of Chernobyl happens in 1987.

Finalement nous avons une paire de texte de la forme :

T: the catastrophe of Chernobyl happens in 1987.
H: It is not by accident that one of the versions explaining the catastrophe at the Chernobyl AES on 26 April 1986 (because smaller-scale accidents happened there before) ties up the cause of the tragedy with experiments at "Chernobyl-Two" .

Comme dans le challenge RTE, les exemples sont divisés en deux types de corpus (corpus de développement et corpus de test).

Les deux corpus sont constitués de 30 paires de textes et chaque portion du corpus doit inclure 50% d'exemples avec une inférence vrai 50% d'exemples avec une inférence fausse. Pour cela, chaque exemple (T-H) paire est jugé par un annotateur pour voir s'il y a une inférence textuelle dans la paire de texte entre (T-H) ou pas.

La figure suivante montre un exemple du corpus après annotation :

```
<pair id="754" value="TRUE" >
     <t> the catastrophe of Chernobyl happens in 1987</t>
     <h> It is not by accident that one of the versions explaining the catastrophe at the
     Chernobyl AES on 26 April 1986 (because smaller-scale accidents happened there
     before) ties up the cause of the tragedy with experiments at "Chernobyl-Two" . </h>
</pair>
```

Id : représente le numéro de la pair.

Value : représente la décision de l'annotateur (vrai ou faux).

Figure 3.2 : Exemple du corpus annoté

L'exemple est évalué par un second juge qui évalue les paires de textes et d'hypothèses, sans avoir pris conscience de leurs contextes.

Les annotateurs étaient d'accord avec le jugement dans 86,66 % des exemples, ce qui correspond à 0.6 Kappa qui est une mesure statistique pour calculer a quel point deux personnes A et B sont d'accord pour classer N éléments dans K catégories mutuellement exclusives, les 13,33% du corpus où il n'y a pas eu d'accord ont été supprimés. Le reste du corpus est considéré comme un «gold standard» ou « BASELINE » pour l'évaluation.

3) Classification de l'inférence temporelle

Apres avoir conçu notre corpus, nous avons annoté manuellement les événements, les dates et les différents types d'inférences (lexicales, syntaxiques et temporelles) existant entre les segments de textes. Cela nous a permis de détecter les différents types d'inférences temporelles entre les segments de textes.
Nous détaillons dans ce qui suit les différentes classes que nous avons distingué:

3.1) Les inférences entre expressions temporelles

L'inférence permet d'établir des relations temporelles liant date, heure et durée entre elles. Dans le même contexte, nous avons distingué trois types d'inférences temporelles liant des expressions temporelles.

Cette figure représente le nombre de paires de textes pour chaque sous classe d'inférence dans notre corpus de développement.

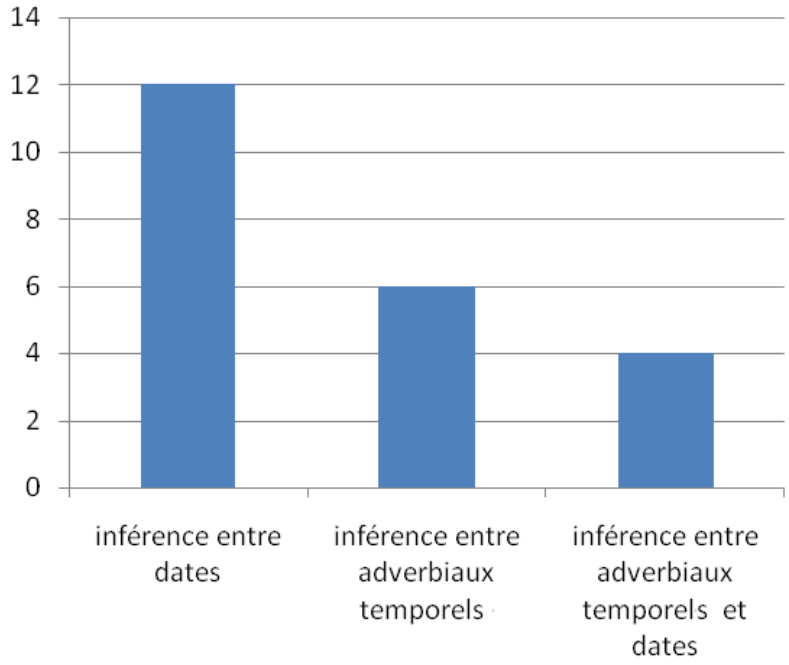

Figure 3.3 : Pourcentage de paires par types d'inférences

Dans ce qui suit, nous présentons les trois types d'inférences :

3.1.1) Les inférences entre dates

C'est la relation temporelle entre qui peut y avoir entre les dates du texte T et les dates du texte H.

L'exemple suivant permet de montrer la relation qui peut exister entre les dates.

Exemple 1:

<pair id="8" value="TRUE" >
T: the football world cup finished on t1: **july 12 th 2006**.
H: the football world cup finished in t2: **july 2006**.

Dans cet exemple, nous remarquons que l'inclusion entre les deux dates t1 et t2 a permis d'avoir l'inférence temporelle.

Exemple 2:

1) <pair id="1" value="TRUE" >
T: the second world war finished in t1: **1945**.
H: the end of the second world war took part t2: **between 1940 and 1950.**

Dans cet exemple nous remarquons aussi que l'inclusion entre les deux dates t1 et t2 a permis d'avoir l'inférence temporelle.

3.1.2) Les inférences entre adverbiaux temporels

L'inférence permet d'établir une relation temporelle entre adverbiaux de référence temporelle qui exprime la localisation d'un événement dans le temps.
L'exemple suivant permet de montrer la relation qui peut exister entre deux adverbiaux temporels.

Exemple 1:

<pair id="15" value="TRUE" >
T: he has worked **during 10 days**.
H: He has worked **for many days**.

Dans cet exemple, nous pouvons remarquer que l'adverbial temporel **« During 10 days »** l'infère l'adverbial **« many days ».**

Exemple 2:

14) <pair id="14" value="TRUE" >
T: **the day before yesterday,** Paul disappeared.
H: **two days ago,** Paul disappeared.

Dans cet exemple nous remarquons que l'adverbial temporel **« the day before yesterday»** infère l'adverbial **« two days ago».**

3.1.3) Les inférences entre dates et adverbiaux temporels

L'inférence permet d'établir des relations temporelles entre dates et adverbes.
L'exemple suivant permet de montrer la relation qui peut exister entre un adverbial temporel et une date.

Exemple 1:

18) <pair id="18" value="TRUE" >
T: the building collapsed at **2 o'clock p.m**.
H: in **the afternoon** the building collapsed.

Dans l'exemple précédant nous pouvons remarquer que **« 2 o'clock p.m »** infère l'adverbial **«the afternoon ».**

Exemple 2:

19) <pair id="19" value="TRUE" >
T: Mark has arrived **on Monday**, **the day after** Celine has arrived.
H: Celine has arrived **on Tuesday**.

Dans l'exemple précédant nous pouvons remarquer que si nous ajoutons **« the day after »** à **« Monday»** nous arrivons à **«Tuesday».** Ceci implique une inférence entre ces adverbiaux temporels.

3.3.2) Les inférences entre évènements

L'inférence permet d'établir des relations temporelles entre événements. Dans ce contexte, nous avons détecté deux types d'inférences, une qui demande la relation entre événements pour détecter l'inférence, et l'autre ne demande que l'inférence lexico sémantique.
Cette figure présente le nombre de paires de textes dans chaque sous classe dans le corpus.

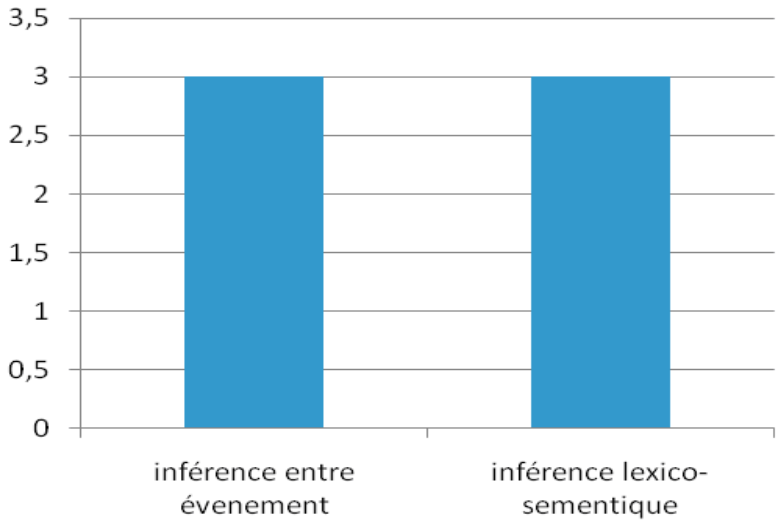

Figure 3.4 : Nombre de paires par types d'inférences

3.3.2.1) Les relations entre évènements temporels

La relation temporelle entre événements est établie par rapport aux relations qu'elle peut avoir avec d'autres événements dans le texte.

L'exemple suivant permet de montrer la relation qui peut exister entre un adverbial temporel et une date.

Exemple 1:

22) <pair id="22" value="TRUE" >
T: since **the death of Turing**, the scientific community **gives** the Turing prize to researchers who found out discoveries in computer science.
H: The Turing prize was not **given** before **the death of Turing**.

Dans l'exemple précédent, nous pouvons apercevoir que les deux événements « **given** » apparaissant dans les deux segments dépendent d'autres événements « **the death of turing** » pour se situer dans le temps.

Exemple 2:

23) <pair id="23" value="TRUE" >
T: Algeria has become **independent**.
H: before its **independence** Algeria was colonized.

Dans l'exemple précédent, nous pouvons apercevoir que l'événement « **independent** » apparaissant dans le segment H, dépendent de l'événement « **was colonized** » pour se situer dans le temps.

3.3.2.2) Les inférences lexico sémantiques

La relation temporelle entre événements est établie par rapport aux relations sémantiques qui peuvent exister entre eux.
L'exemple suivant permet de montrer la relation lexico-sémantique existante entre deux évènements.

Exemple 1:

26) <pair id="26" value="TRUE" >
T: France has **won** the match against Brasil.
H: France has **played** the match against Brasil.

Dans l'exemple précédent, nous pouvons constater que l'évènement « **won** » se produit après l'évènement « **played** ».

Exemple 2:

27) <pair id="27" value="TRUE" >
T : Amine **was dreaming** .
H : Amine **was sleeping deeply**.
Dans l'exemple précédant nous pouvons constater que l'évènement « **was dreaming**» se produit durant l'évènement « **was sleeping deeply**».

3.3.4) Les inférences entre évènements et expressions temporelles

L'inférence permet d'établir des relations temporelles entre événements et expressions temporelles.
Cette figure représente le nombre de paires de textes où existe ce type d'inférence.

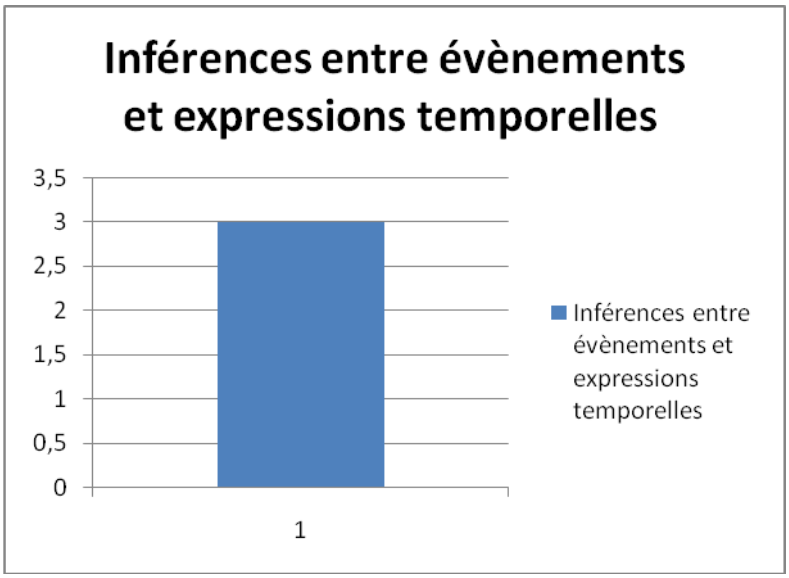

Figure 3.5 : Nombre de paires par types d'inférences

L'exemple suivant permet de montrer la relation temporelle existante entre évènements et expressions temporelles.
Exemple 1:

29) <pair id="29" value="TRUE" >
T: Japan gave weapons back after **the explosion of the first atomic bomb**.
H: Japan gave weapons back in **1945**.

Dans l'exemple précédent, nous pouvons remarquer que l'évènement « **the explosion of the first atomic bomb** » est ancré temporellement avec l'expression temporelle «**1945** ».

Exemple 2:

30) <pair id="30" value="TRUE" >
T: Germany has become unified since **the fall down of the Berlin wall.**
H: Germany unified **19 years ago**.

Dans l'exemple précédent, nous pouvons remarquer que l'évènement «**the fall down of the Berlin wall**» est ancré temporellement avec l'expression temporelle «**19 years ago**».

4) Le bilan de l'étude du corpus

Dans notre élaboration du corpus, nous nous sommes limités à des segments de textes relativement brefs et concrets. Nous retrouvons dans ce corpus des inférences temporelles sous des formes variées.
Le tableau suivant représente le pourcentage de paires du corpus de développement par type d'inférence temporelle existante, sachant qu'il existe 30 paires dans notre corpus.

Types d'inférences temporelles	Nombres de paires
Inférences entre expressions temporelles	21/30
Inférences entre évènements	6/30
Inférences entre évènements et expressions temporelles	3/30

Tableau 3.1 : Nombre de paire dans le corpus

Cette figure représente le pourcentage de paires de chaque type d'inférence dans le corpus de développement :

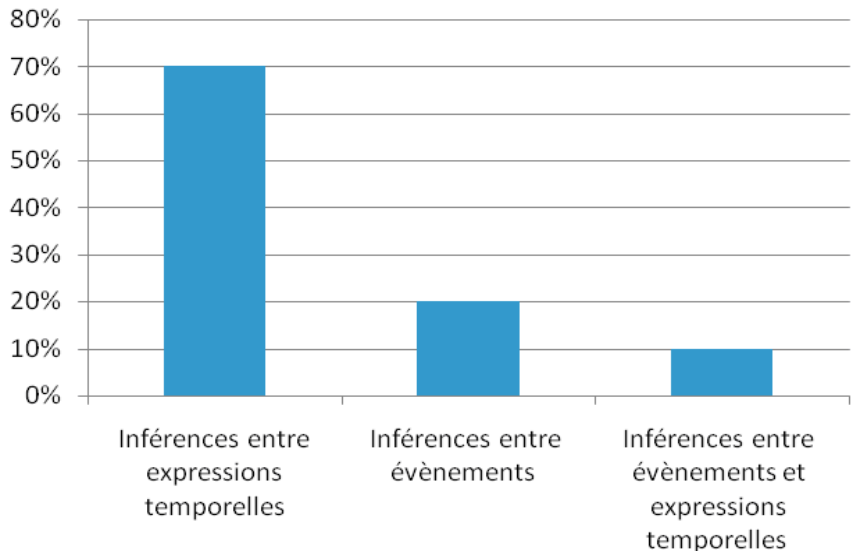

Figure 3.6 : Pourcentage de paires par types d'inférences

Nous constatons que notre corpus de développement a un pourcentage élevé de paires contenant une inférence temporelle entre expression temporelle et cela est dû à une forte présence de questions d'ordres temporelles extraites du corpus de test du challenge clé.
Les détails des corpus de test et de développement sont disponibles en annexe.

5) Conclusion

Dans ce chapitre nous avons expliqué, comment nous avons élaboré un corpus contenant des paires de segments de textes integrant des relations temporelles, ensuite nous avons fait une classification des différents types d'inférences temporelles existantes dans le corpus.
La suite logique de ce travail consiste à déduire des régles d'inférences temporelles et à les intégrer à un systéme d'inférence texuelle. Ces démarches sont l'objet du chapitre suivant que nous allons exposer.