Un annotateur automatique d'expressions temporelles du français et son évaluation sur le TimeBank du français

André Bittar Caroline Hagège
XRCE, 6 Chemin de Maupertuis, 38240 Meylan, FRANCE
Andre.Bittar@xrce.xerox.com, Caroline.Hagege@xrce.xerox.com

RÉSUMÉ

Dans cet article, nous présentons un outil d'extraction et de normalisation d'un sous-ensemble d'expressions temporelles développé pour le français. Cet outil est mis au point et utilisé dans le cadre du projet ANR Chronolines[1] et il est appliqué sur un corpus fourni par l'AFP. Notre but final dans le cadre du projet est de construire semi-automatiquement des chronologies événementielles à partir de la base de depêches de l'AFP. L'une des étapes du traitement est l'analyse de l'information temporelle véhiculée dans les textes. Nous avons donc développé un annotateur d'expressions temporelles pour le français que nous décrivons dans cet article. Nous présenterons également les résultats de son évaluation.

ABSTRACT

An Automatic Temporal Expression Annotator and its Evaluation on the French TimeBank

In this article, we present a tool that extracts and normalises a subset of temporal expressions in French. This tool is being developed and used in the ANR (French National Research Agency) project Chronolines, applied to a corpus of provided by the Agence France Presse. The aim of the project is to semi-automatically construct event chronologies from this corpus. To do this, a detailed analysis of the temporal information conveyed by texts, is required. The system we present here is the first version of a temporal annotator that we have developed for French. We describe it in this article and present the results of an evaluation.

MOTS-CLÉS : Analyse temporelle, évaluation.

KEYWORDS: Temporal processing, evaluation.

1 Introduction

Le travail présenté ici s'insère dans un cadre plus ambitieux qui est la constitution semi-automatique de chronologies événementielles à partir d'une requête effectuée sur un grand ensemble de depêches de l'AFP pour le français et pour l'anglais. Afin de pouvoir constituer des chronologies événementielles, il est capital de pouvoir analyser dans un premier temps le contenu textuel (afin de repérer les événements) mais aussi de reconnaître et normaliser les expressions temporelles associées à ces événements. L'outil présenté ici est un annotateur d'un sous-ensemble d'expressions temporelles qui :
– repère les expressions temporelles

[1] ANR-10-CORD-010, http://www.chronolines.fr

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 463–470,
Grenoble, 4 au 8 juin 2012.

– normalise ces expresssion temporelles
– rajoute des annotations concernant la modalité des événements auxquels ces expressions se rapportent

Dans une première partie, nous présentons un bref état de l'art concernant l'analyse automatique de la temporalité en général et pour le français en particulier. Puis nous décrivons l'outil que nous avons développé pour le français. Nous présenterons enfin les résultats obtenus par l'annotateur en les comparant avec le TimeBank du français (TBF), un corpus de textes en français annotés selon la norme ISO-TimeML (Pustejovsky *et al.*, 2010).

2 Etat de l'art

L'analyse de la temporalité est un élément important pour un grand nombre de tâches et de ressources relevant du traitement informatique des textes. (Li *et al.*, 2005) montre l'importance de l'analyse de la composante temporelle pour les systèmes de Questions/Réponses. Pour le résumé multi-document, l'ajout de la composante temporelle aide à repérer les éléments textuels véhiculant une information similaire (Barzilay et Elhadad, 2002). Les grandes bases de connaissances constituées automatiquement ou semi-automatiquement grâce à l'extraction d'information textuelles s'enrichissent actuellement d'une composante temporelle (Wang *et al.*, 2010). Par ailleurs, la norme ISO-TimeML est aujourd'hui largement adoptée, tant dans le cadre de la mise en place de ressources annotées avec des informations temporelles pour diverses langues dont le français (Bittar *et al.*, 2011), mais aussi lors des compétitions TempEval (Pustejovsky et Verhagen, 2010) au cours desquelles divers outils d'annotation automatique des informations temporelles sont évalués. Plus spécifiquement pour le français, outre le TimeBank mentionné ci-dessus, plusieurs travaux visant à l'analyse automatique de la temporalité ont vu le jour. Concernant l'annotation et le typage des expressions temporelles (ET) nous pouvons citer les travaux de (Battistelli *et al.*, 2008) qui présente une représentation algébrique des expressions de type date, de (Ehrmann et Hagège, 2009) qui propose un typage des ET accompagné de critères syntaxiques et sémantiques, mais aussi dans le domaine de l'annotation automatique, l'outil décrit dans (Parent *et al.*, 2008), ainsi que celui de (Teissèdre *et al.*, 2011).

3 Description de l'annotateur

L'annotateur que nous avons développé est intégré à un analyseur linguistique (Aït-Mokhtar *et al.*, 2002) qui produit une analyse syntaxique en dépendances à partir d'un texte d'entrée (texte brut ou XML). L'annotateur a été développé pour les besoins du projet Chronolines qui vise à traiter le corpus de dépêches (couvrant les années allant de 2004 à 2011) mis à disposition par l'Agence France-Presse. Ce corpus est constitué d'environ 1 million de documents (chaque document correspondant à une dépêche) comprenant environ 9,4 millions d'expressions temporelles de tout type. Ces dépêches sont disponibles dans un format XML (NewsML).

Le module spécifique pour la reconnaissance des expressions temporelles est constitué de plusieurs éléments qui seront détaillés plus bas :
– Ajout d'information lexicale permettant de typer les adverbes de temps.
– Règles locales permettant de délimiter les ET et si possible de les typer. Ces règles sont intégrées

à la grammaire générale de l'analyseur.
– Règles utilisant les dépendances syntaxiques permettant de procéder à un typage plus fin de ces ET. Ces règles sont également intégrées à la grammaire générale.
– Programme Java externe à la grammaire qui utilise les informations linguistiques pour désambiguiser certaines expressions et pour procéder à la normalisation des ET sélectionnées.

3.1 Information lexicale

L'information lexicale spécifique à l'analyse temporelle consiste essentiellement en l'ajout de traits sémantiques sur des éléments lexicaux qui vont rentrer dans la composition d'une ET. Par exemple, les noms de jours (e.g. *lundi*) se voient attribuer un trait spécifique *[day :+]*. De même, les noms de mois, de fêtes, des adverbes de temps seront marqués dans le lexique.

3.2 Règles locales

Les règles locales vont permettre d'assembler des éléments lexicaux quand ils peuvent potentiellement correspondre à une expression temporelle incluant plusieurs constituants de base. Par exemple, une expression comme *mi-mars 2012* est segmentée originellement par l'analyseur morphologique en 4 segments *mi* + - + *mars* + *2012*. Une règle locale regroupe ces quatre segments afin de constituer une seule expression temporelle. Lors de l'application de ces règles locales, un premier typage des expressions temporelles est effectué dans le cas où celles-ci ne sont pas ambiguës. Par exemple, une expression comme *mi-mars 2012* peut être typée comme une date absolue[2] sans avoir recours au contexte .

3.3 Règles de dépendances raffinant le typage

Certaines expressions temporelles reconnues par une analyse lexicale ou par l'application de règles locales ne peuvent cependant pas être typées *a priori*. En effet, l'analyse de la seule expression ne permet de déterminer de quel type d'expression il s'agit et seul un contexte plus large permet de désambiguïser. Parfois, une même expression mettant en jeu des unités linguistiques constitutives d'une ET peut s'avérer, en contexte, ne pas être une ET. Par exemple, une expression telle que *trois ans*, n'est pas une ET dans (1) mais correspond à une durée dans (2). L'expression *en avril* dans (3) correspond à une date relative alors que dans (4), elle correspond à une date récurrente (qui peut être paraphrasée par *tous les mois d'avril*).

1. Jean fêtera bientôt ses **trois ans**.
2. Il est resté **trois ans** sans la voir.
3. Il était malade **en avril**.
4. **En Avril**, il fait le grand nettoyage de printemps.

Grâce à l'analyse syntaxique sous-jacente, des restrictions utilisant à la fois des informations syntaxiques (fonction syntaxique de l'ET potentielle), et des informations sémantiques lexicales

[2] nous détaillons à la section 3.5 les différents types de dates que nous extrayons et la terminologie adoptée pour les distinguer.

(verbe *rester* est un verbe de permanence), permettent de filtrer et de mieux typer des ET extraites lors des étapes précédentes de traitement (analyse lexicale et règles locales).

3.4 Programme externe

Une API java de l'analyseur permet d'étendre les traitements linguistiques et d'utiliser les résultats des analyses dans du code extérieur à l'analyseur. La normalisation[3] des ET extraites est effectuée par ce biais. Une fois de plus, à ce stade, l'analyse des expressions temporelles est encore raffinée afin de pouvoir procéder correctement à la normalisation. Par exemple, une expression comme *lundi* est une expression de type date relative par rapport au moment de l'énonciation (ME) mais elle peut être antérieure ou postérieur au ME selon les contextes, ainsi qu'en témoignent les exemples suivants.

1. Elle est partie **lundi**.
2. Elle partira **lundi**.

3.5 Type de dates extraites reconnues

Bien que nous délimitions tout type d'ET, nous avons mis l'accent pour une première utilisation de l'annotateur sur la normalisation d'un sous-ensemble d'expressions temporelles qui est le suivant :

Nous considérons les dates absolues, et les dates relatives au moment de l'énonciation qui correspondent à des intervalles bornés ou à des points. Nous avons pour ce faire défini des critères de segmentation et de typage décrits dans (Bittar *et al.*, 2012). Toutes ces dates que nous extrayons doivent être normalisées.

Parmi ces expressions nous avons des ET comme *En 2003*, *En janvier 2003*, *le 24 juin 2010*, *le mois dernier*, *lundi*, *dans quatre mois*, etc. Les trois premiers exemples sont des dates absolues, les exemples suivants sont des dates relatives au ME dans la mesure où pour procéder à la normalisation de ces dates, il est nécessaire de connaître la date correspondant à l'assertion durant laquelle cette date est mentionnée.

3.6 Extraction d'information sur la modalité

Dans le cadre plus large du projet, nous avons souhaité distinguer les événements datés factuels (c'est à dire les événements considérés comme avérés par l'auteur de la dépêche) des événements datés hypothétiques ou relevant du discours rapporté. Dans la mesure où nous utilisons un analyseur linguistique, nous disposons des liens syntaxiques entre l'ET extraite et le prédicat nominal ou verbal que cette ET modifie. La prise en compte de la modalité dans le traitement de la temporalité est un vaste et riche domaine (voir (Battistelli, 2009) pour une présentation détaillée) que nous n'avons pas considéré dans son ensemble. Nous avons cependant distingué

[3]Nous entendons ici par normalisation le fait d'attribuer une valeur correspondante au calendrier (éventuellement sous-spécifiée) à une ET.

<DCT value="20040101"/>
L' <EN TYPE="LOCORG">Irlande</EN> s'apprête à prendre <EC TYPE="DATE" SUBTYPE="REL" REF="ST" value ="20040101">jeudi</EC> la présidence tournante de l' <EN TYPE="ORG">Union européenne</EN> qui doit vivre <EC TYPE="DATE" SUBTYPE="REL" REF="ST" value ="20040501" FACTUAL="MODAL">le 1er mai</EC> un élargissement historique...

FIG. 1 – Exemple de sortie de l'annotateur.

les cas suivants, qui se produisent de manière assez fréquente sur les corpus que nous traitons et qui nous semblent pertinents pour la finalité applicative que nous avons dans le cadre du projet.

– L'expression temporelle se trouve dans une proposition dont le verbe principal est à une forme modale ou future.
– L'expression temporelle se trouve dans une enchâssée qui relève du discours rapporté.
– L'expression temporelle est associée à un verbe *dicendi* introduisant un discours rapporté.

Dans le premier cas, l'ET est soit le modifieur d'un verbe utilisé à une forme modale ou au futur (information dont nous disposons grâce à l'analyse morpho-syntaxique), soit le modifieur d'un nom argument d'un verbe utilisé à une forme modale ou future. L'exemple ci-dessous marque l'ET *en 2006* par un attribut *FACTUAL="MODAL"* dans la mesure où le prédicat auquel elle se rapporte (*s'achever*) est employé à une forme modale.

Exemple :
Il a réaffirmé la primauté de son mandat de cinq ans , qui doit s'achever <EC TYPE="DATE" SUBTYPE="ABS" value ="2006XXXX" FACTUAL="MODAL">*en 2006*</EC>

Dans le deuxième cas, l'ET est modifieur d'un verbe d'une enchâssée dépendante d'un verbe *dicendi*. Dans ce cas, l'annotation de l'ET est enrichie par l'attribut *REPORTED="YES"*.

Exemple :
Il a annoncé que des élections se dérouleraient <EC TYPE="DATE" SUBTYPE="ABS" value ="2004XXXX" REPORTED="YES" FACTUAL="MODAL">*en 2004*</EC>

Enfin, dans le dernier cas, l'ET est modifieur d'un verbe marqué comme étant un verbe introducteur d'un discours rapporté. Dans ce cas, l'annotation de l'ET comporte l'attribut *DECLARATION="YES"*.

Exemple :
La Libye avait annoncé <EC TYPE="DATE" SUBTYPE="REL" REF="ST" DECLARATION="YES" value ="20031219">*le 19 décembre*</EC> *sa décision de renoncer aux armes de destruction massive.*

3.7 Exemple de sortie

L'annotateur prend en entrée du texte brut ou un texte au format XML. Il produit en sortie un texte au format XML qui correspond au texte initial enrichi par les annotations des ET que nous avons décrites. Les entités nommées sont également marquées. Un exemple de sortie est illustré dans la Figure 1.

4 Evaluation sur TimeBank

Afin de pouvoir évaluer notre outil, nous avons comparé les annotations produites avec le corpus TimeBank du français (TBF) annoté selon la norme ISO-TimeML. Cette évaluation a nécessité quelques adaptations afin de faire correspondre les sorties de l'outil d'analyse aux données de TimeBank. Ces adaptations sont décrites ci-dessous. Par ailleurs, l'information concernant la modalité et le discours rapporté n'a pas été évaluée.

4.1 Adaptations

Les adaptations ont été nécessaires pour trois raisons principales : la délimitation des ET selon notre approche n'est pas tout à fait semblable à celle adoptée par la norme ISO-TimeML, nous ne considérons qu'un sous-ensemble des ET envisagées dans TBF et enfin, le format de normalisation que nous adoptons est différent de celui de ISO-TimeML.

Dans TBF, les balises qui marquent les ET (<TIMEX3>) n'incluent pas les éventuels marqueurs de relation (ex. les prépositions temporelles telles que *avant, après, etc.*). En effet, la norme ISO-TimeML préconise de les annoter séparément avec une autre balise (<SIGNAL>). Selon notre schéma d'annotation (Bittar *et al.*, 2012) cependant, les marqueurs de relations sont annotés à l'intérieur d'une seule balise délimitant l'expression temporelle. Afin de résoudre cette différence, nous avons converti le TimeBank selon notre format par application d'un simple transducteur qui place le contenu textuel de la balise <SIGNAL> à l'intérieur de la balise <TIMEX3> qui la suit directement. Par ailleurs, notre annotateur a été adapté pour fournir une sortie contenant la balise (<SIGNAL>). Nous obtenons donc pour ce type d'expression la représentation finale(1).

1. <TIMEX3><SIGNAL>depuis< /SIGNAL>mars 2003< /TIMEX3>

Notre annotateur ne traite pour l'instant que des dates absolues et des dates relatives au moment de l'énonciation correspondant à des intervalles bornés et dont la granularité n'est pas inférieure au jour. Nous avons donc retiré du corpus TBF toutes les annotations des expressions ne correspondant à ces catégories. Concrètement, toute balise <TIMEX3> obéissant à l'un des quatre critères mentionnés ci-dessous a été supprimée du corpus de référence.

– l'attribut `temporalFunction="true"` (qui indique qu'il s'agit d'une date relative) est présent et l'attribut `anchorTimeID` est différent de `t1` (la DCT). Ce critère permet de retirer du corpus de référence toutes les dates relatives à un référent textuel.
– l'attribut `type` a une des valeurs `DURATION`, `SET` ou `TIME`. Ce critère permet de retirer du corpus de référence toutes les durées, les heures ou les agrégats temporels.
– l'attribut `value` est `PAST_REF`, `PRESENT_REF` ou `FUTURE_REF`. Ce critère permet de retirer du corpus de référence toutes les expressions de dates floues.
– la balise a l'attribut `MOD` indiquant que la date a un modifieur (de début, de fin, d'approximation, etc.). Ce critère permet également de retirer du corpus de référence des dates floues.

Enfin, nous avons fait converger les formats de la valeur normalisée des ET afin d'obtenir une représentation comparable entre TBF et les sorties de l'annotateur. Le corpus adapté pour notre évaluation contient 299 expressions temporelles annotées (sur les 608 du corpus original).

4.2 Resultats pour le français

Les performances, en termes de rappel, précision et F-mesure, ainsi que la mesure kappa (Cohen, 1960), figurent dans le Tableau 1. Pour la détection des expressions temporelles, les performances sont satisfaisantes, mais peuvent encore être améliorées. Les erreurs sont dues essentiellement à des manques de couverture dans la grammaire, qui est encore en cours de développement. Aucune erreur de typage des expressions n'a été commise. De très bons résultats ont été obtenus pour la normalisation des expressions. La principale source d'erreurs pour la normalisation provient des cas où une ET apparaît dans un contexte où elle n'est pas reliée par une dépendance à un prédicat verbal. Ceci est parfois le cas lorsque la clause où apparaît l'ET ne contient effectivement pas de verbe, mais cela peut également se produire suite à une erreur de l'analyseur syntaxique. Dans un de ces cas, le temps verbal est donc indisponible pour le calcul de la valeur correcte normalisant l'ET. Enfin, lors de l'évaluation, un certain nombre de désaccords entre la sortie du système et la référence ont révélé des erreurs du TBF. Ces erreurs n'ont pas été prises en compte pour l'évaluation et elles ont été transmises au gestionnaire du corpus.

	Précision	Rappel	F-mesure	Kappa
Étendue des balises	0.90	0.84	0.87	0.71
Attribut `type`	1.0	1.0	1.0	1.0
Attribut `value`	0.94	1.0	0.96	0.92

TAB. 1 – Performances du système sur l'ensemble du corpus d'évaluation.

5 Conclusion

Nous avons développé une première version d'un outil d'analyse des ET du français à partir de textes tout-venant. Nous avons utilisé cet annotateur pour effectuer des expériences visant à extraire les dates importantes mentionnées dans de grands volumes de texte. L'annotateur a été évalué grâce au TimeBank du français. L'élargissement de cet annotateur à d'autres types d'ET est en cours (prise en compte d'autres des dates référentielles par rapport à un référent introduit dans le discours, des dates répétitives (fréquences dans la terminologie ISO-TimeML), et des dates qui ne peuvent être assimiliée à un point ou à un intervalle temporel borné).

Remerciements

Remerciement à l'ANR qui a financé une partie de ce travail, ainsi qu'à X. Tannier, R. Kessler et V. Moriceau qui ont utilisé et commenté les sorties de l'annotateur. Nous remercions également D. Teyssou de l'AFP qui nous a donné accès au corpus de dépêches.

Références

Aït-Mokhtar, S., Chanod, J.-P. et Roux, C. (2002). Robustness beyond Shallowness : Incremental Deep Parsing. *Natural Language Engineering*, 8:121–144.

Barzilay, R. et Elhadad, N. (2002). Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Battistelli, D. (2009). *La temporalité linguistique : circonscrire un objet d'analyse ainsi que des finalités à cette analyse*. Université Paris-Ouest Nanterre La Défense (Paris 10).

Battistelli, D., Couto, J., Minel, J.-L. et Schwer, S. (2008). Représentation algébrique des expressions calendaires et vue calendaire d'un texte. *In* (Bechet *et al.*, 2008).

Bechet, F., Bellot, P., Bonastre, J.-F. et Jimenez, T., éditeurs (2008). *Actes de TALN 2008 (Traitement automatique des langues naturelles)*, Avignon. ATALA, LIA.

Bittar, A., Amsili, P., Denis, P. et Danlos, L. (2011). French TimeBank : An ISO-TimeML Annotated Reference Corpus. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers*, volume 2, Portland. Association for Computational Linguistics.

Bittar, A., Hagège, C., Tannier, X., Moriceau, V. et Teissèdre, C. (2012). Temporal Annotation : A Proposal for Guidelines and an Experiment with Inter-annotator Agreement. *In Proceedings of LREC 2012 - to appear*, Istanbul. ELRA.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 43(6):551–558.

Ehrmann, M. et Hagège, C. (2009). Proposition de caractérisation et de typage des expressions temporelles en contexte. *In* (Nazarenko et Poibeau, 2009).

Li, W., Li, W., Lu, Q. et Wong, K.-F. (2005). A Preliminary Work on Classifying Time Granularities of Temporal Questions. *In Proceedings of Second international joint conference in NLP (IJCNLP 2005)*, Jeju Island, Korea.

Nazarenko, A. et Poibeau, T., éditeurs (2009). *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis. ATALA, LIPN.

Parent, G., Gagnon, M. et Muller, P. (2008). Annotation d'expressions temporelles et d'événements en français. *In* (Bechet *et al.*, 2008).

Pustejovsky, J., Lee, K., Bunt, H. et Romary, L. (2010). ISO-TimeML : An international standard for semantic annotation. *In* Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M. et Tapias, D., éditeurs : *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Pustejovsky, J. et Verhagen, M. (2010). SemEval-2010 Task 13 : Evaluating Events, Time Expressions, and Temporal Relations (TempEval-2).

Teissèdre, C., Battistelli, D. et Minel, J.-L. (2011). Recherche d'information et temps linguistique : une heuristique pour calculer la pertinence des expressions calendaires. *In Actes de TALN 2011 (Traitement automatique des langues naturelles)*, Montréal. ATALA.

Wang, Y., Zhu, M., Qu, L., Spaniol, M. et Weikum, G. (2010). Timely YAGO : Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia. *In Proceedings of the* 13^{th} *International Conference on Extending Database Technology (EDBT), Lausanne, Switzerland, March 22-26*, pages 697–700.