Détection d'inférences textuelles (RTE)

Séminaire PluriTAL « Fouille de Textes »
P. Amsili

septembre 2019

Plan

Motivation

Inférence

Opération (logique) par laquelle on admet une proposition en vertu de sa liaison avec d'autres propositions déjà tenues pour vraies.

(Petit Robert, 1990)

- notion très générale (surtout en anglais) :
  implication, présupposition, implicature, sous-entendu, raisonnement par défaut, stéréotypes...

RTE

Version orientée « langage » :

Textual Entailment concluding the truth of a *textual statement* (*Hypothesis*) based on (the truth of) another given piece of text (*Text*).

⇒ Normalisé en 2004 (Ido Dagan et collaborateurs)
PASCAL Recognizing Textual Entailment Challenges

Exemple

QA — Who painted ‘the Scream’ ?

src *“Norway’s most famous painting, ‘The Scream’, by Edvard Munch...*

→ Edvard Munch painted ‘The Scream’

A — Edvard Munch

Contexte

- besoin d'inférence dans beaucoup d'applications
- outils génériques existants :
  - ressources lexico-sémantiques manuelles (WordNet, FrameNet) ou semi-automatique (thesauri distributionnels, inference-rules...)
  - modules TAL (NER, SRL, WSD...)

$\Rightarrow$ développement d'un mécanisme d'inférence : assemblage ad hoc de composants génériques et spécifiques

$\Rightarrow$ manque d'un cadre général (indépendant d'une application) Comparer avec la situation du parsing

Attentes

- Algorithmes d'inférence
- Composants
- Ressources
- moteurs d'inférence NLP
  $\neq$ moteurs d'inférence logiques

Plan

Définition

On définit l'iplication textuelle (TE) comme une relation **orientée** entre deux expressions textuelles, appellées T (Texte) et H (Hypothèse).
On dit que T implique H si un locuteur humain, en lisant T ferait typiquement l'inférence que H est vraisemblablement vraie.

- Jugement humain $\longrightarrow$ gold standard
- indépendant d'une application
- flou volontaire dans la définition
  $\rightarrow$ cas limites à prévoir

Hypothèses

- On suppose le Texte cohérent (discursivement)
- On suppose que H est une expression linguistique bien formée, autonome (typiquement une phrase courte)

  Rq : beaucoup de phrases sont attributives, ce qui réduit l'intérêt pour les inférences générales, et rapproche la tâche d'une tâche de recherche d'information (IR).

Exemple (Dagan *et al.* , 2013)

Table 1.1: Examples of Text-Hypothesis pairs, taken from the RTE-2 development set, along with the application (Task) from which they were derived and the human annotation of whether the pair satisfies the entailment relation or not (Judgment). SUM=summarization; IR=information retrieval; IE=information extraction; QA=question answering. See Section 1.4 for the data-set generation methodology

ID	Text	Hypothesis	Task	Judgment
77	Google and NASA announced a working agreement, Wednesday, that could result in the Internet giant building a complex of up to 1 million square feet on NASA-owned property, adjacent to Moffett Field, near Mountain View.	Google may build a campus on NASA property.	SUM	YES
110	Drew Walker, NHS Tayside's public health director, said: "It is important to stress that this is not a confirmed case of rabies."	A case of rabies was confirmed.	IR	NO
294	Meanwhile, in an exclusive interview with a TIME journalist, the first one-on-one session given to a Western print publication since his election as president of Iran earlier this year, Ahmadinejad attacked the "threat" to bring the issue of Iran's nuclear activity to the UN Security Council by the US, France, Britain and Germany.	Ahmadinejad is a citizen of Iran.	IE	YES
387	About two weeks before the trial started, I was in Shapiro's office in Century City.	Shapiro works in Century City.	QA	YES
415	The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them.	Alzheimer's disease is treated using drugs.	IR	YES
691	Arabic, for example, is used densely across North Africa and from the Eastern Mediterranean to the Philippines, as the key language of the Arab world and the primary vehicle of Islam.	Arabic is the primary language of the Philippines.	QA	NO

Exemple

- tirés de cas typiques d'application
  SUM, IR, IE, QA
- dérivations d'informations nouvelles à partir des prémisses (sur la base d'un raisonnement) 294, 387
- généralisation 415
- (quasi-)équivalence/paraphrase
  ex. *buy* vs. *purchase*

RTE modélise la **variabilité des expressions linguistiques**

Remarques

Comparer

- Similarité textuelle
- Paraphrase
- Inférence textuelle

Modes :

- Reconnaissance (in : T+H, out : Y/N)
- Recherche (in : T + Corpus, out : T + $\{H_1, H_2... \in$ Corpus$\}$)
- Génération (in : T, out : $\{H_1, H_2... \}$)

Background Knowledge

Connaissances nécessaires :

- linguistiques
- extra-linguistiques

$\rightarrow$ problème d'acquisition de connaissances

Mais :
Dans la tâche RTE, la véracité de H ne peut pas provenir de la WK sans prise en considération de T

$\Rightarrow$ variante définitionnelle :

$$\text{TE} : \exists\ K \text{ t.q. } K \cup T \models H \ \&\ K \not\models H$$

RTE & logique

RTE couvre plus que la logique :
TE(T,H) si

- $T (\cup K) \models H$, ou
- H est hautement plausible étant donné T

De plus, si H est un tautologie, alors $T \models H$, mais on ne veut pas TE(T,H)

- RTE est plus proche de la logique probabiliste (bayesienne ou pas)

Extension : contradiction

- RTE classique : classification binaire
- RTE augmentée : Entailment, Contradiction, Unknown de Marneffe *et al.* (2009)
- RTE « diminuée » : « Text relatedness » = Entailment ∪ Contradiction

Trace

Entailment / Contradiction / Unknown?

- Text:

  The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

- Hyp 1: BMI acquired an American company.

Trace (2)

Entailment / Contradiction / Unknown?

- Text:

  The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

- Hyp 2: BMI bought employee-owned LexCorp for $3.4Bn.

Trace (3)

Entailment / Contradiction / Unknown?

- Text:

  The purchase of Houston-based LexCorp by BMI for $2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

- Hyp 3: BMI is an employee-owned concern.

13

Trace

Tâches existantes :

- NER
- Coref
- SRL
- lexical inference-rules (purchase $\rightarrow$ acquire)

Tâches non existante

- Monotonie
- bridging
- Relations Cause-effet
- ...

Applications

- QA : test d'une relation RTE entre le passage et la relation extraite : accuracy 30.6% $\rightarrow$ 42.7% *(Harabagiu & Hickl, 2006)*
- RE : (relation extraction $\in$ IE) : augmentation de la performance d'un système non supervisé :
  - search mode
  - template hypothese *(Romano* et al. *, 2006)*
- SUM : entailment-based summary selection : meilleur résumé parmi 6 dans 86% des cas *(Harabagiu* et al. *, 2007)*
- etc.

en général, RTE est utilisé :

- pour valider/filtrer des candidats
- pour identifier des réponses possibles/variables pertinentes avec template hypotheses

Plan

Challenges

RTE1-5

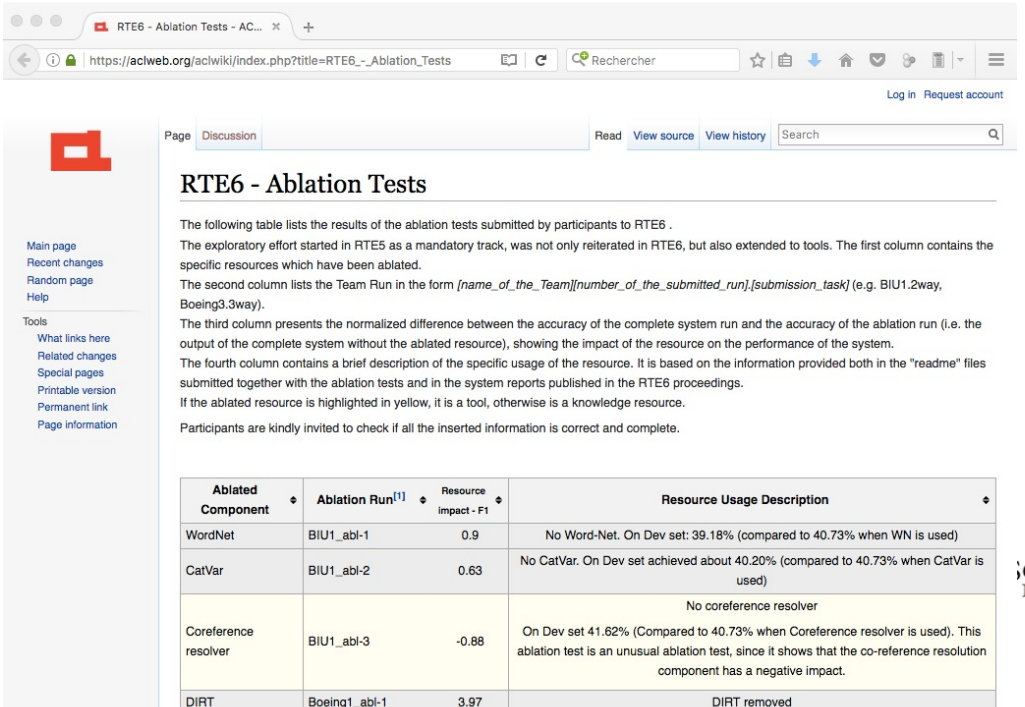
- application scenarios : QA, RE, IR, SUM
- annotateurs humains / crowdsourcing
  mesure κ 0.78 pour RTE 2 (accord satisfaisant)
  voir slides Kappa de Cohen/Fleiss
- 600 à 1000 paires équilibrées (50% Yes, 50% No ; les No sont subdivisés à partir de RTE-5)
- Textes de plus en plus longs
- Mesure : classification $\rightarrow$ accuracy
  ranking $\rightarrow$ average precision

RTE 6-7

- méthode d'équilibrage plus réaliste que 50-50
  $\rightarrow$ évaluation en f-mesure (précision/rappel)

Ablation tests

- évaluation de la contribution de chaque module :
- évaluations successives du même système avec un composant en moins
- limite : phénomènes spécifiques rares (verbes de mouvement, présupposition temporelle...)
- réponse : spécialized datasets



RTE6 - Ablation Tests - AC...

https://aclweb.org/aclwiki/index.php?title=RTE6_-_Ablation_Tests

Rechercher

Log in Request account

Page Discussion Read View source View history Search

Main page
Recent changes
Random page
Help
Tools
What links here
Related changes
Special pages
Printable version
Permanent link
Page information

RTE6 - Ablation Tests

The following table lists the results of the ablation tests submitted by participants to RTE6 .
The exploratory effort started in RTE5 as a mandatory track, was not only reiterated in RTE6, but also extended to tools. The first column contains the specific resources which have been ablated.
The second column lists the Team Run in the form *[name_of_the_Team][number_of_the_submitted_run].[submission_task]* (e.g. BIU1.2way, Boeing3.3way).
The third column presents the normalized difference between the accuracy of the complete system run and the accuracy of the ablation run (i.e. the output of the complete system without the ablated resource), showing the impact of the resource on the performance of the system.
The fourth column contains a brief description of the specific usage of the resource. It is based on the information provided both in the "readme" files submitted together with the ablation tests and in the system reports published in the RTE6 proceedings.
If the ablated resource is highlighted in yellow, it is a tool, otherwise is a knowledge resource.

Participants are kindly invited to check if all the inserted information is correct and complete.

Ablated Component	Ablation Run[1]	Resource impact - F1	Resource Usage Description
WordNet	BIU1_abl-1	0.9	No Word-Net. On Dev set: 39.18% (compared to 40.73% when WN is used)
CatVar	BIU1_abl-2	0.63	No CatVar. On Dev set achieved about 40.20% (compared to 40.73% when CatVar is used)
Coreference resolver	BIU1_abl-3	-0.88	No coreference resolver On Dev set 41.62% (Compared to 40.73% when Coreference resolver is used). This ablation test is an unusual ablation test, since it shows that the co-reference resolution component has a negative impact.
DIRT	Boeing1_abl-1	3.97	DIRT removed

Evaluation : pistes

- Evaluation de la justification produite par le système (tentée en pilote pour RTE3)
- Évaluation par la tâche
  suppose le développement de modules génériques d'entraînement

Plan

Modèle intuitif

(1) a. Jean est allé en train à Montargis et a acheté une C3.

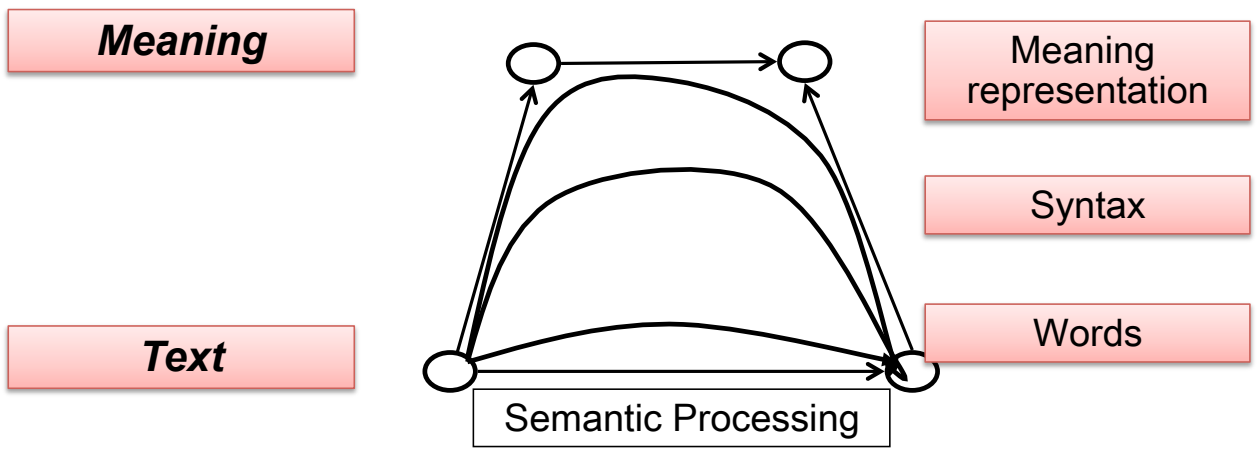
b. `aller_en_train(`e_1`, Jean, Montargis) & acheter(`e_2`, Jean, C3)`

(2) a. Jean est allé en train à Montargis.

b. `aller_en_train(`e_1`, Jean, Montargis)`

Schéma classique Dagan *et al.* (2013)

"Easy-first processing"



- Perform as many inferences over natural language representations as possible
- Resort to formal meaning representation when necessary

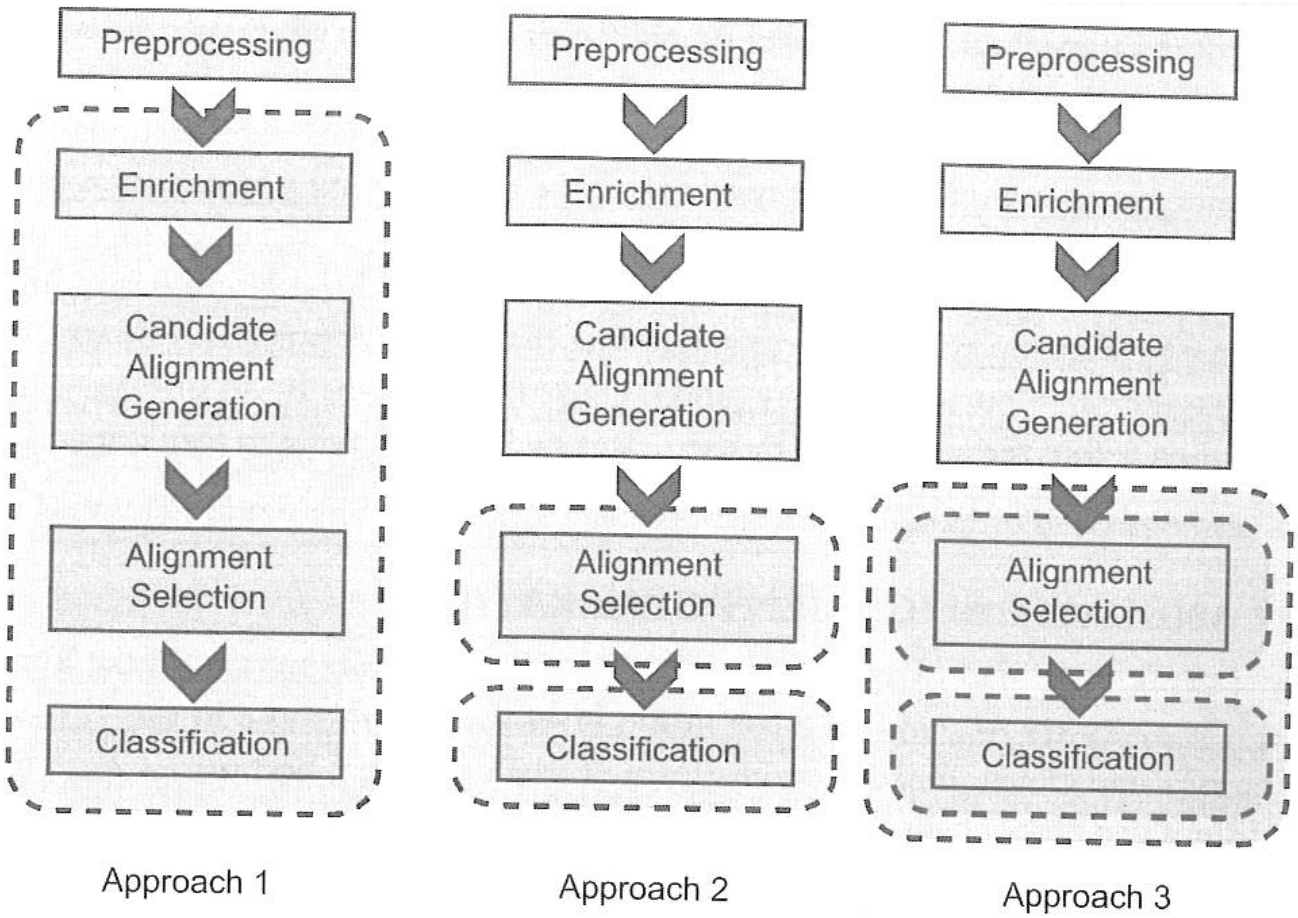
Niveaux de représentation

- Text & Hypothèse représentés dans le même cadre
- bag of words
- strings of words
- SRL structures (Propbank/FrameNet)
- arbre syntaxique (constituant/dépendance)
- représentation sémantique : DRS/ formule logique...

Inférence

- Similarity-based approaches
  - lexical : overlap des tokens/lemmes/mots similaires
  - syntaxique : distance d'arbres
- Alignment-based approaches
  - lexical : recherche de facteurs
  - syntaxique
- Proof-theoretic approaches (theorem-prover)
- Transformation-based approaches

Architecture conceptuelle



Dagan *et al.* (2013)

Preprocessing

- segmentation (mot/phrases)
- lemmatisation
- POS-tagging
- (chunking)
- constituant/dependancy parsing
- Named Entity Recognition
- Coreference Resolution
- Semantic Role Labelling

Défis

Les deux problèmes principaux à ce jour :

- Knowledge acquisition bottleneck
- Noise tolerant architectures

Critiques et extensions

(Sammons *et al.* , 2010)

- Utilisation de RTE pour évaluation externe d'autres tâches
- Interaction problem
- Sparseness problem
- Augmentation de la tâche avec une "explication" (évaluation plus pertinente)

(Bowman *et al.* , 2015)

- SNLI : 570 000 paires produites par crowdsourcing
- permet de comparer les performances d'un classifier avec des traits lexicaux et celle d'un LSTM : très similaires ($\approx$ 76% acuracy sur la classification ternaire)

References

Androutsopoulos, Ion, & Malakasiotis, Prodromos. 2010. A survey of paraphrasing and textual entailment methods. *J. Artif. Int. Res.*, **38**(1), 135–187.

Artstein, Ron, & Poesio, Massimo. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, **34**(4), 555–596.

Bergmair, Richard. 2009. A proposal on evaluation measures for RTE. *Pages 10–17 of : Proceedings of the 2009 Workshop on Applied Textual Inference*. TextInfer '09. Stroudsburg, PA, USA : Association for Computational Linguistics.

Bowman, Samuel R., Angeli, Gabor, Potts, Christopher, & Manning, Christopher D. 2015. A large annotated corpus for learning natural language inference. *CoRR*, **abs/1508.05326**.

Dagan, Ido, Roth, Dan, Sammons, Mark, & Zanzotto, Fabio Massimo. 2013. *Recognizing Textual Entailment : Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

de Marneffe, Marie, Grimm, Scott, & Potts, Christopher. 2009. Not a Simple Yes or No : Uncertainty in Indirect Answers. *In : Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue*. Queen Mary University of London : ACL.

Harabagiu, Sanda, & Hickl, Andrew. 2006. Methods for using textual entailment in open-domain question answering. *Pages 905–912 of : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Harabagiu, Sanda, Hickl, Andrew, & Lacatusu, Finley. 2007. Satisfying information needs with multi-document summaries. *Information Processing & Management*, **43**(6), 1619–1642.

Petit Robert. 1990. *Dictionnaire alphabétique et analogique de la langue française*. 9e édition, Le Robert.

Romano, Lorenza, Kouylekov, Milen, Szpektor, Idan, Dagan, Ido, & Lavelli, Alberto. 2006. Investigating a Generic Paraphrase-Based Approach for Relation Extraction. *In : EACL*.

Sammons, Mark, Vydiswaran, V.G.Vinod, & Roth, Dan. 2010. "Ask Not What Textual Entailment Can Do for You...". *Pages 1199–1208 of : Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden : Association for Computational Linguistics.

Tâches de TAL

SUM Text summarization :

in :	texte	out :	texte résumé

Variante : multidocument

IR Information Retrieval :

in :	requête	out :	ensemble de textes

IE Information Extraction :

in :	formulaire	out :	formulaire rempli

QA Question Answering :

in :	question	out :	réponse tirée de bd/documents