A baseline temporal tagger for all languages

J. Strötgen & M. Gertz
2015

Mathilde Guernut - M2LFC
2018/2019

Faisant référence

Temporal processing
- Extraction
- Normalization

A baseline temporal tagger for all languages

J. Strötgen & M. Gertz

Créateurs du temporal tagger HeidelTime
Développement de HeidelTime pour l'allemand

Fonctionnant sur toutes les langues existantes

Temporal processing & temporal tagger

*“Temporal processing is a field of Computational Linguistics which aims to access this dimension and derive a **precise temporal representation of a natural language text** by extracting time expressions, events and temporal relations, and then representing them according to a chosen knowledge framework.”* G. Marsic, 2011

Temporal processing & temporal tagger

2 étapes

- Extraction

Temporal information extraction is the **identification of chunks/tokens corresponding to temporal intervals**, and the extraction and determination of the temporal relations between those.

- Normalisation

Temporal expression normalisation is the grounding of a lexicalisation of a time to a calendar date or other formal temporal representation.

Source: http://nlpprogress.com/english/temporal_processing.html

En théorie

Crucial en NLP

Permet

- Recherche d’informations
- Ordonner chronologiquement les événements d’un texte
- …

Applications dans beaucoup de domaines notamment médical

En pratique

- Temporal tagger - surtout pour une langue spécifique, principalement l’anglais
- HeidelTime = le seul à avoir été développé pour plus d’une langue (manuellement)
- Procédé manuel
- Tentatives de rendre le procédé automatique = ne traite que d’une sous-tâche, soit extraction, soit normalisation

Limites

- Étendre un temporal tagger manuellement à une autre langue demande beaucoup de temps & de ressources

→ certaines langues ne seront jamais traitées

- Chaque langue traitée séparément = variation dans la qualité

Intérêt du projet

Une technique imparfaite mais “existante” permettra de faire progresser la recherche

Permettra de traiter les 2 sous-tâches du temporal processing i.e extraction & normalisation

HeidelTime est open access: permettra contributions extérieures

→ **question de recherche** = est-il possible de développer une technique automatique d’ajout de langue pour le time tagging & comment faire?

HeidelTime

Input: sentence, token, (PoS), (publi date)

Pre-process: Tokenization, Sentence splitting, (PoS tagging)

Output: expressions temporelles en TimeML avec TIMEX3 tags (type, value ++)

Ex:

```
<TIMEX3 tid="t25" type="DATE" value="1989-11-02" temporalFunction="false"
functionInDocument="PUBLICATION_TIME">11/02/89</TIMEX3>
```

HeidelTime

Chaque langue nécessite 3 types d'info: extraction / normalization / rules

- Pattern (extraction) = termes fréquemment utilisés pour expressions temporelles

  ex: JJ/MM/AA

- Normalization informations = règles d'écriture

  ex: January = 01

- Rules = règles des expressions temporelles, contiennent règles d'extraction & normalisation

Ajouter une langue dans HT - manuellement

Processing manuel - 3 étapes

- Pre-processing de la langue ajoutée ie Tokenization, Sentence splitting, PoS tagging
- Traduire tous les patterns & normalization files
- Réécrire toutes les règles

Ajouter une langue dans HT - auto - 1

- Pre-processing de la langue ajoutée ie Tokenization, Sentence splitting PoS tagging retiré car n'existe pas dans toutes les langues

- Réécriture d'un tokenizer & sentence splitter fonctionnant sur toutes les langues, basé sur les espaces entre les mots

- Réécriture des patterns, normalization, rules en 3 catégories

Ajouter une langue dans HT - auto - 2

2 cas : valable pour toutes les langues ; spécifique à la langue

Cas 1- valable pour toutes langues:

- Réécriture générique des patterns & normalization files

→ info fonctionnant sur toutes les langues

ex: ordre d'organisation des tokens dans la date:
DD/MM/YY, YY/MM/DD etc

Ajouter une langue dans HT - auto - 3

Cas 2 - spécifique à la langue:
ex: janvier, january etc

- Pour chaque langue, création de pattern & normalization files remplis à partir des règles de traduction depuis l'anglais basé sur Wiktionary

- Permet backtracking → contrôle des données, les info du fichier original "eng" sont présentes dans le fichier de chaque langue

- Amélioration des patterns & normalization files "eng" en les éprouvant sur corpus largement étudiés

Ajouter une langue dans HT - auto

Chaque langue =

1. Language independent pattern & normalization files
2. Language independent rules
3. Language dependent pattern, normalization files & rules translated from english

Évaluation

Precision

Recall / rappel

Accuracy

	HeidelTime 1.9 (manual)					HeidelTime – automatic				
	relaxed extr			value		relaxed extr			value	
language: corpus / domain	P	R	F1	F1	acc.	P	R	F1	F1	acc.
English: TE-3 TimeBank / news (UzZaman et al., 2013)	93.1	90.8	91.9	79.6	86.5	95.6	49.2	64.9	54.7	84.3
English: TE-3 platinum / news (UzZaman et al., 2013)	93.1	88.4	90.7	78.1	86.1	98.7	56.5	71.9	54.4	75.7
English: WikiWars / narrative (Mazur and Dale, 2010)	98.3	86.1	91.8	83.1	90.5	97.9	58.4	73.2	53.4	73.0
Arabic: Arabic test-50* / news (Strötgen et al., 2014)	90.9	90.9	90.9	82.2	90.4	91.7	31.8	47.2	38.0	80.5
Chinese: TE-2 test impro. / news (Li et al., 2014)	95.8	89.3	92.4	79.5	86.0	100	9.5	17.3	7.6	44.0
Croatian: WikiWarsHR / narrative (Skukan et al., 2014)	92.6	90.5	91.5	80.8	88.3	87.3	6.8	12.6	9.7	77.0
French: FR-TimeBank / news (Bittar et al., 2011)	91.9	90.1	91.0	73.6	80.9	87.2	59.5	70.8	54.6	77.1
German: WikiWarsDE / narrative (Strötgen and Gertz, 2011)	98.7	89.3	93.8	83.0	88.5	98.4	64.7	78.1	59.7	76.4
Italian: EVALITA'14 test / news (Caselli et al., 2014)	92.7	86.1	89.3	75.0	84.0	98.5	41.2	58.1	49.3	84.9
Spanish: TempEval-3 test / news (UzZaman et al., 2013)	96.0	84.9	90.1	85.3	94.7	95.5	53.8	68.8	58.5	85.0
Vietnamese: WikiWarsVN / narrative (Strötgen et al., 2014)	98.2	98.2	98.2	91.4	93.1	84.0	45.5	59.0	27.1	45.9
Portuguese: PT-TimeBank test / news (Costa and Branco, 2012)	87.3	75.9	81.2	63.5	78.2	91.5	59.3	72.0	59.4	82.5
Portuguese: PT-TimeBank train / news (Costa and Branco, 2012)	83.3	73.1	77.9	54.5	70.0	88.2	51.0	64.6	50.4	78.0
Romanian: Ro-TimeBank / news (Forascu and Tufis, 2012)	–	–	–	–	–	31.9	11.4	16.9	7.8	46.2

Table 1: Evaluation results for several languages on public corpora. HeidelTime 1.9 results as reported on `https://github.com/HeidelTime/heideltime/wiki/Evaluation-Results`.

Precision, recall, F1 score & accuracy

Précision: ce qu'on a trouvé positif et qui l'est / tout ce qu'on a détecté positif (et qui ne l'est pas forcément)

Parmi toutes les expressions temporelles identifiées, quel % est vraiment une expression temporelle

Recall: ce qu'on a trouvé positif et qui l'est / ce qui est vraiment positif

Parmi toutes les expressions temporelles du corpus, quel % a été identifié

F1: moyenne harmonique de P&R

Accuracy: tous les vrais positifs + tous les vrais négatifs / tout ce qu'on a

ce qu'on a trouvé positif et qui l'est + ce qu'on a trouvé négatif et qui l'est / l'ensemble

Le % d'expressions temporelle correctement détecté (positif & négatif) par le test évalué ie % de résultats pas erronés

Resultats

F1: score plus faible auto que manuel
Or F1 = moyenne P&R

P: haute dans les 2

R: score plus faible auto que manuel

Acc: haute dans les 2

HT manuel

HT auto

language: corpus / domain	HeidelTime 1.9 (manual) relaxed extr P	R	F1	value F1	acc.	HeidelTime – automatic relaxed extr P	R	F1	value F1	acc.
English: TE-3 TimeBank / news (UzZaman et al., 2013)	93.1	90.8	91.9	79.6	86.5	95.6	49.2	64.9	54.7	84.3
English: TE-3 platinum / news (UzZaman et al., 2013)	93.1	88.4	90.7	78.1	86.1	98.7	56.5	71.9	54.4	75.7
English: WikiWars / narrative (Mazur and Dale, 2010)	98.3	86.1	91.8	83.1	90.5	97.9	58.4	73.2	53.4	73.0
Arabic: Arabic test-50* / news (Strötgen et al., 2014)	90.9	90.9	90.9	82.2	90.4	91.7	31.8	47.2	38.0	80.5
Chinese: TE-2 test impro. / news (Li et al., 2014)	95.8	89.3	92.4	79.5	86.0	100	9.5	17.3	7.6	44.0
Croatian: WikiWarsHR / narrative (Skukan et al., 2014)	92.6	90.5	91.5	80.8	88.3	87.3	6.8	12.6	9.7	77.0
French: FR-TimeBank / news (Bittar et al., 2011)	91.9	90.1	91.0	73.6	80.9	87.2	59.5	70.8	54.6	77.1
German: WikiWarsDE / narrative (Strötgen and Gertz, 2011)	98.7	89.3	93.8	83.0	88.5	98.4	64.7	78.1	59.7	76.4
Italian: EVALITA'14 test / news (Caselli et al., 2014)	92.7	86.1	89.3	75.0	84.0	98.5	41.2	58.1	49.3	84.9
Spanish: TempEval-3 test / news (UzZaman et al., 2013)	96.0	84.9	90.1	85.3	94.7	95.5	53.8	68.8	58.5	85.0
Vietnamese: WikiWarsVN / narrative (Strötgen et al., 2014)	98.2	98.2	98.2	91.4	93.1	84.0	45.5	59.0	27.1	45.9
Portuguese: PT-TimeBank test / news (Costa and Branco, 2012)	87.3	75.9	81.2	63.5	78.2	91.5	59.3	72.0	59.4	82.5
Portuguese: PT-TimeBank train / news (Costa and Branco, 2012)	83.3	73.1	77.9	54.5	70.0	88.2	51.0	64.6	50.4	78.0
Romanian: Ro-TimeBank / news (Forascu and Tufis, 2012)	–	–	–	–	–	31.9	11.4	16.9	7.8	46.2

Table 1: Evaluation results for several languages on public corpora. HeidelTime 1.9 results as reported on `https://github.com/HeidelTime/heideltime/wiki/Evaluation-Results`.

Comparaison HT manuel / HT auto / corpus de référence (i.e 100% acc, entièrement manuel)

Resultats

- P. haute = peu de faux positifs ie ce qui est détecté comme une expression temporelle l'est vraiment

→ dans certains cas, + précis en auto qu'en manuel (chinois)

- R. bas = beaucoup de faux négatifs ie beaucoup d'expressions temporelles n'ont pas été détectées par HT
- Problème de détection des expressions temporelles mais quand détectées, juste
- Acc haute = malgré R bas & nombre importants de faux négatifs

Discussion

- Résultats moins bons que HT manuel
- Cas particuliers du chinois, croate, roumain
- Résultats liés aussi à la disponibilité des traductions dans Wiktionary

2 limites concernant les langues ajoutées

- Liée à la construction de la langue:
  - langue riche morphologiquement
  - langue sans white space tokenization
- Liée à la trad dans Wiktionary
  - seulement 34 langues traduites à 75% dans wiktionary pour les patterns, 83 langues à 50%

Perspectives

→ **question de recherche** = est-il possible de développer une technique automatique d'ajout de langue pour le time tagging & comment faire?

YEP! \o/

Mais **limites**

Les auteurs proposent de l'utiliser comme base pour amélioration du process (open source), soit comme point de comparaison à d'autres process

Processus dynamique: HT 2.0 ; wiktionary = mis à jour constante ; augmentation des références, étendu à wikipedia:

"In the future, we thus plan to constantly update HeidelTime's automatically created resources."

Aujourd'hui

2018: HT 2.2.1 (last update sept 2016)

Version 2.2.1

SHA: f7e4c3f

- added: temponym tagging functionality［Core］
- added: English temponym resources［Resources］
- fixed: parameter pos set to "no" (POSTagger.NO) works for all languages now (AllLanguagesTokenizer)［Standalone］
- fixed: several minor issues

Version 2.1

SHA: 378e476

- fixed: TreeTaggerWrapper no longer creates temporary files which speeds up processing［Core］
- fixed: Various improvements to Maven support［Maven］
- fixed: Errors in Arabic resources［Resources］
- fixed: Values in IntervalTagger that were switched for a while［Core］

Version 2.0

SHA: b9248e0

- added: automatically-created resources for 200+ languages［Resources］
- added: AllLanguagesTokenizer, a simple, generic, whitespace-based tokenizer that can be used with all languages
- fixed: Minor rule improvements for some languages［Resources］

→ Mars 2018: changements mineurs dans Readme

Source: https://github.com/HeidelTime/heideltime/wiki/Changelog#version-21

Références

Marsic, G. (2011). *Temporal Processing of News : Annotation of Temporal Expressions, Verbal Events and Temporal Relations*. University of Wolverhampton. Retrieved from http://clg.wlv.ac.uk/papers/marsic-thesis.pdf

Sanampudi SK, Kumari GV. Temporal Reasoning in Natural Language Processing: A Survey. Int J Comput Appl. 2010;1(4):68–72.

http://nlpprogress.com/english/temporal_processing.html

https://github.com/HeidelTime/heideltime/wiki/Changelog#version-21