Aurélie Névéol
Hercules Dalianis, Sumithra Velupillai,
Guergana Savova, Pierre Zweigenbaum

A scoping review of the literature on Clinical Natural Language Processing for Languages other than English

dépasser les frontières

Why address a variety of languages?

- Access to a larger demographic
  - Access to more patient cohorts
  - Agregate data for rare and other diseases, e.g. autism spectrum disorder in 4 healthcare centers **[Kohane et al. 2012]**

- Apply WHO protocols widely
  - Success story in the making: **IRIS**, a software for automated coding of causes of death
  - Collaboration involving France, Hungary, Japan, Germany, Italy, Sweden, United States

Literature on Clinical NLP is hard to find!

- International Medical Informatics Association (IMIA) Yearbook
  - Clinical NLP section started in 2014
    - Survey paper, synopsis with « best papers » selection
  - For year 2017, 709 articles reviewed
    - ACL anthology: BioNLP, *ACL conferences (34% off topic)
    - Pubmed: natural language processing (35% off topic)
    - Pubmed: text mining (60% off topic)
    - Overall, 31 (4.3%) addressed a language other than English
- Reviewing tools used
  - Bibreview https://pypi.org/project/BibReview/
  - Integrated classifier **[Norman et al. P47 Friday@9:45]**

Literature on Clinical NLP is hard to find!

- American Medical Informatics Association (AMIA)
  - Panels on clinical NLP for languages other than English in 2014, 2017.

Névéol *et al. Journal of Biomedical Semantics* (2018) 9:12
https://doi.org/10.1186/s13326-018-0179-8

Journal of
Biomedical Semantics

REVIEW　Open Access

Clinical Natural Language Processing in languages other than English: opportunities and challenges

Aurélie Névéol[1]*, Hercules Dalianis[2], Sumithra Velupillai[3,4], Guergana Savova[5] and Pierre Zweigenbaum[1]

Abstract

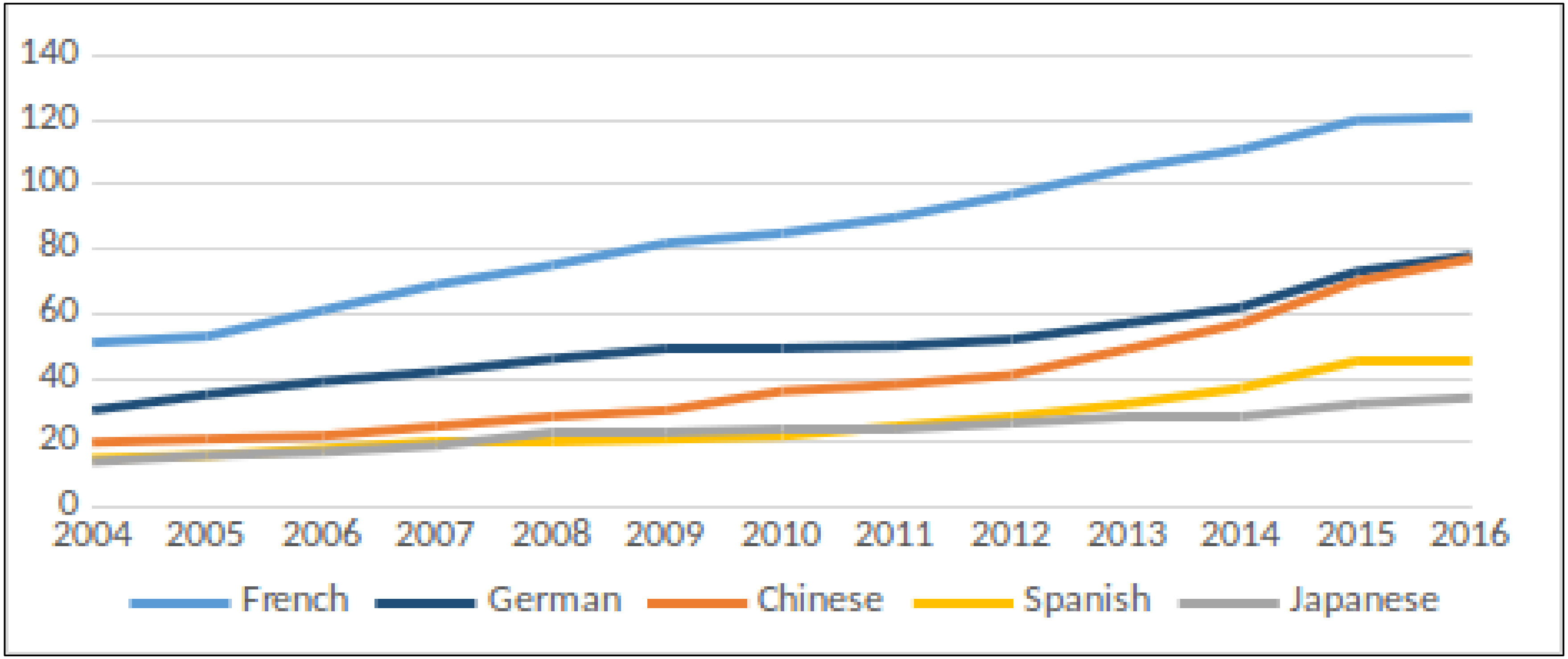
Background: Natural language processing applied to clinical text or aimed at a clinical outcome has been thriving in recent years. This paper offers the first broad overview of clinical Natural Language Processing (NLP) for languages other than English. Recent studies are summarized to offer insights and outline opportunities in this area.

Main Body: We envision three groups of intended readers: (1) NLP researchers leveraging experience gained in other languages, (2) NLP researchers faced with establishing clinical text processing in a language other than English, and (3) clinical informatics researchers and practitioners looking for resources in their languages in order to apply NLP techniques and tools to clinical practice and/or investigation. We review work in clinical NLP in languages other than English. We classify these studies into three groups: (i) studies describing the development of new NLP systems or components de novo, (ii) studies describing the adaptation of NLP architectures developed for English to another language, and (iii) studies focusing on a particular clinical application.

Conclusion: We show the advantages and drawbacks of each method, and highlight the appropriate application context. Finally, we identify major challenges and opportunities that will affect the impact of NLP on clinical practice and public health studies in a context that encompasses English as well as other languages.

Keywords: Natural Language Processing, Clinical Decision-Making, Languages other than English

Growth of bio-clinical NLP publications in MEDLINE for the top 5 studied languages other than English



(22 languages covered in review)

Biomedical NLP in a language other than English

- What does it consist in?
  - Data creation: vocabularies, annotated dataset
  - Method development: NLP methods for the biomedical domain, bioNLP tasks
  - Applications
- Is it different from bioNLP in English?
  - Less resources
  - Language, country specificities
  - Multilingual aspects: translation, language adaptation, cross-culture comparisons

Building new systems and resources

Domain-specific NLP components

- Morphological analyzer
  - DeRIF, for French [Namer & Zweigenbaum 2004]
- PoS tagger
  - Experiments conducted for Portuguese [Oleynik et al. 2010], Polish [Marciniak and Mykowiecka 2011], Spanish [Costumero et al. 2014]
- Parser
  - Some work for French [Baud et al. 1999] and Finish [Haverinen et al.], but no public tool
- Entity and concept recognition
  - No equivalent of Metamap or cTAKES
  - Some tools for direct lexical matching, e.g. BioPortal [Jonquet et al.]

Automatic word segmentation and applications

Named Entity Recognition in Chinese [Lei et al. JAMIA 2014], [Xu et al. JAMIA 2014]

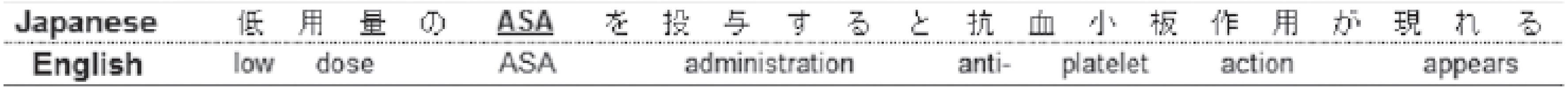
- Word segmentation (vs. character) performs better
- Joint segmentation+NER yields 1-% improvement for both
- F-measure of 90+% for 4 entity types: performance comparable to English, with specific features

Word segmentation in Japanese [Nishimoto et al. Methods Inf Med 2008]

- Addresses the lack of spacing
- A probabilistic model of word segmentation using dictionaries
- Successfully applied to English – can be useful for OCR

Transliteration issues

- Expansion of English Abbreviations in Japanese **[Shinohara et al. Methods Inf Med 2013]**
  - code-switching

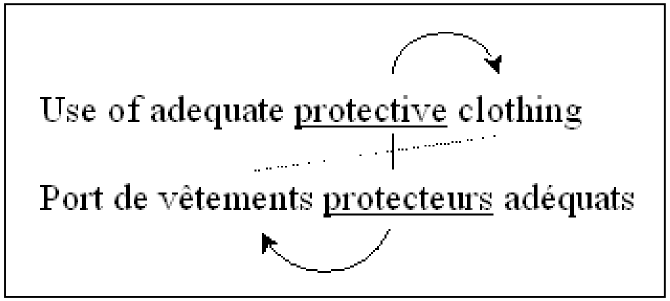


Japanese	低	用量	の	ASA	を	投与する	と	抗	血小板	作用	が	現れる
English	low	dose		ASA		administration		anti-	platelet	action		appears

  - Pilot study on 8 short forms associated to 2 or more long forms
  - Character segmentation (vs. word) performs better

- Word segmentation in Hebrew **[Cohen et al. Methods Inf Med 2010]**
  - Identification of transliterated words improves medical term extraction by 29%

Lexicons and terminology development



Use of adequate protective clothing

Port de vêtements protecteurs adéquats

- Term translation
  - Through term alignment using parallel **[Deléger et al. 2010]** or comparable **[Chiao and Zweigenbaum 2002]** corpora
  - Using automatic translation systems **[van Mulligen et al. 2016]**

- Mapping of terminologies to the UMLS
  - ATC **[Merabti et al. 2011]** or CCAM **[Bousquet et al. 2012]**
  - French consumer vocabulary **[Tapi Nzali et al. 2017]**

Corpora and annotations (for Romance Languages)

- Monolingual Corpora

	Corpus	Text type	Annotations	Availability
Spanish	BARR Oronoz et al. IULA	Literature EHR EHR	Abbreviations Entities (ADR) Negation	Open Restricted Open
French	CépiDC QUAERO MERLOT Sequoia Tapi Nzali et al.	Death certificate Literature EHR Drug inserts Social Media	ICD10 Concepts E+R+M PoS Sentiment	under DUA Open Restricted Open Restricted
Portugese	Aluisio et al.	Patient speech	classification	Restricted
Italian	Attardi et al.	EHR	Silver entities	From authors
Romanian	BioRo	Literature, lecture notes	**[Mitrofan P16**	**Wed 4:35]**

- Parallel Corpora
  - Used in the WMT campaigns: Scielo, EDP, UFAL

Adapting NLP architectures developed for English

Rule-based systems

- Negation
  - – Adaptation of NegEx to French [Chapman et al. 2013], Swedish [Skeppstedt 2011], German [Cotik et al. 2016], Dutch [Afzal et al. 2014] and Spanish [Costumero et al. 2014] [Cotik et al. 2016]

Absence of [evidence to suggest acute cardiac process]

Absence de [ganglions métastasiques]

- De-identification
  - – Adaptation of De-ID to French [Grouin et al. 2009]
- Temporal analysis
  - – Heideltime adapted to French [Tapi Nzali et al. 2015] and Swedish [Vellupilai et al. 2014]

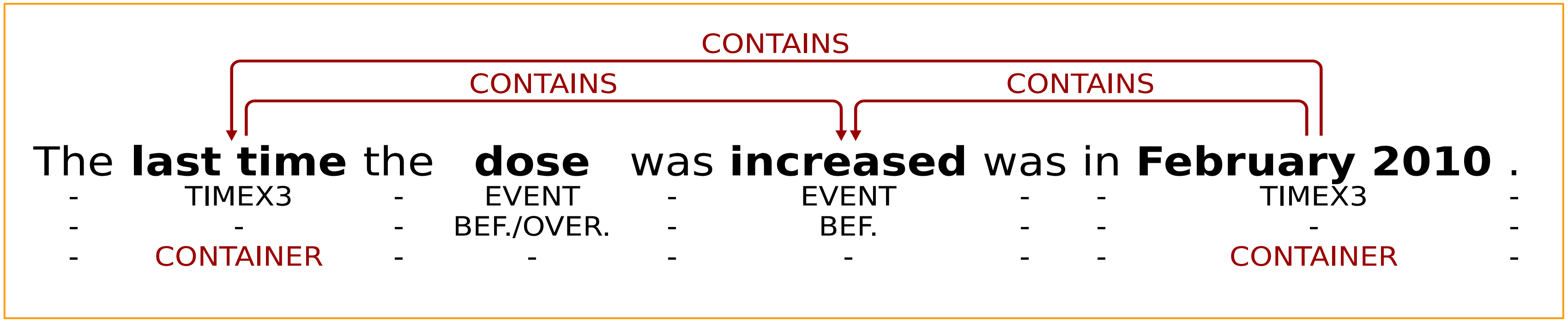
Temporal relation extraction: Clinical TempEval task, THYME corpus

- Temporal "container" relations [Bethard et al. 2016]

  Between pairs of events

  Between events and temporal expressions

  Between events and document creation time



The	**last time**	the	**dose**	was	**increased**	was	in	**February 2010**	.
-	TIMEX3	-	EVENT	-	EVENT	-	-	TIMEX3	-
-	-	-	BEF./OVER.	-	BEF.	-	-	-	-
-	CONTAINER	-	-	-	-	-	-	CONTAINER	-

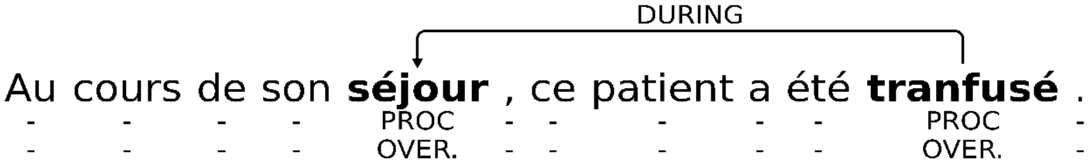
- Participation in the challenge with a neural system

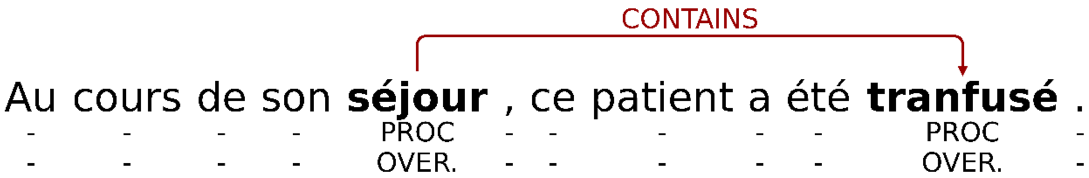
  [Tourille et al. 2016]

Temporal relation extraction: English to French Language Adaptation

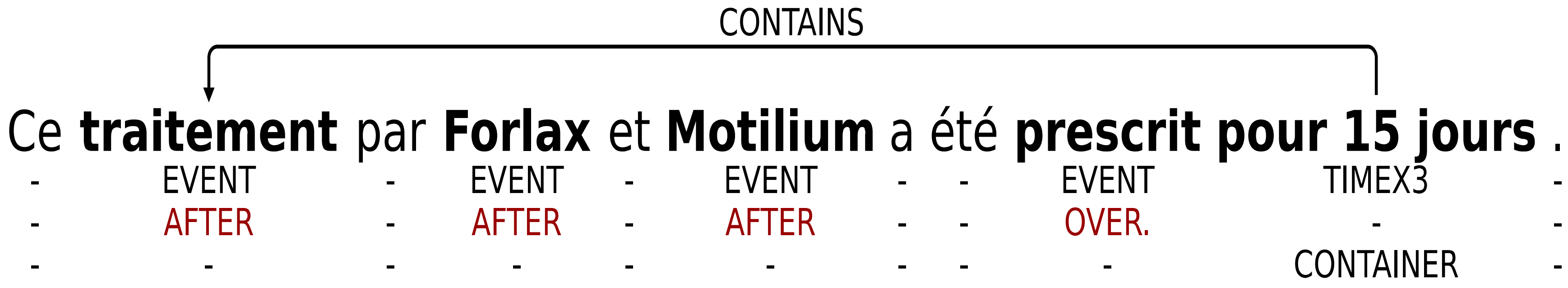
- Use of the MERLOT French clinical corpus, with entity and relation annotations [Campillos et al. 2017]

- Need to convert temporal relations
  - From TimeML to “container” relations
  - Also include related relations, e.g. “reveals”


DURING
Au cours de son **séjour** , ce patient a été **tranfusé** .
- - - - PROC - - - - - PROC -
- - - - OVER. - - - - - OVER. -

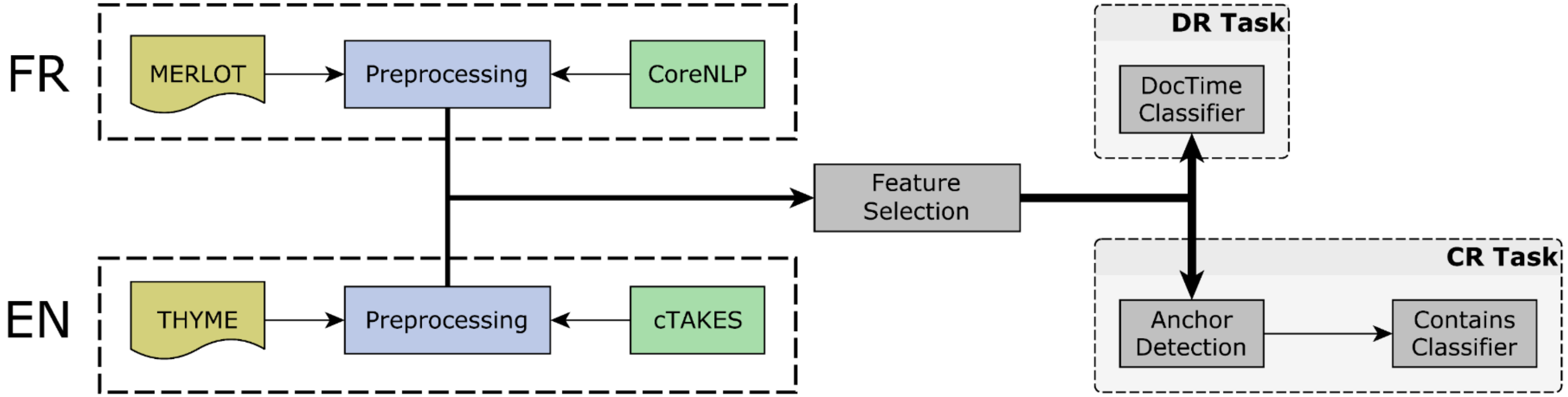

CONTAINS
Au cours de son **séjour** , ce patient a été **tranfusé** .
- - - - PROC - - - - - PROC -
- - - - OVER. - - - - - OVER. -

Experiments: English to French Language Adaptation



Ce	**traitement**	par	**Forlax**	et	**Motilium**	a	été	**prescrit**	**pour 15 jours**	.
-	EVENT	-	EVENT	-	EVENT	-	-	EVENT	TIMEX3	-
-	AFTER	-	AFTER	-	AFTER	-	-	OVER.	-	-
-	-	-	-	-	-	-	-	-	CONTAINER	-

- Generic framework with language dependent features



Results: Temporal relation extraction in English and French

1. DCT relations

	French (MERLOT)			English (THYME)		
	P	R	F1	P	R	F1
baseline	0.67	0.67	0.67	0.47	0.47	0.47
bef./over.	0.68	0.69	0.69	0.73	0.60	0.66
before	0.81	0.60	0.69	0.88	0.88	0.88
after	0.79	0.69	0.73	0.84	0.84	0.84
overlap	0.88	0.92	0.90	0.88	0.90	0.89
micro-average	0.83	0.84	0.83	0.87	0.87	0.87

2. Contains relations

	French (MERLOT)			English (THYME)		
	P	R	F1	P	R	F1
baseline	0.43	0.15	0.22	0.55	0.06	0.11
no-relation	0.99	1.00	0.99	0.96	0.98	0.97
contains	0.75	0.57	0.65	0.61	0.47	0.53
micro-average	0.98	0.98	0.98	0.93	0.94	0.93

- In spite of lower resources for French, performance is similar
- Performance is comparable to inter-annotator agreement

[Tourille et al. EACL 2017]

Applications

Biomedical NLP tasks addressed

- Text classification
  - – Healthcare associated infections in Swedish patient records [Jacobson and Dalianis 2016]
  - – Multiple Myeloma in German records [Löpprich et al 2016]

- Information extraction used for computing clinical scores
  - – Cardiovascular score (French) [Grouin et al. 2012]
  - – Memory scores (Japanese) [Takano et al. 2017]

Multilingual Corpora

Improve access to medical information

- Off-the-shelf automatic translation, e.g. Google translate, Babelfish **[Zeng-Treitler et al. 2010] [Wu et al. 2011]**
- Medical Speech translation **[Bouillon et al. 2007]**

Crosslingual Information Retrieval by query translation

- French, knowledge-based **[Thirion et al. 2010]**
- French/Czech/German, MT based **[Pecina et al. 2014]**

Study of clinical cultural differences

- Breast cancer information in Germany vs. UK **[Weissenberger et al. 2004]**
- Clinical records **[Wu et al. 2013]** and doctor reviews **[Hao et al. 2017]** in China vs. US

Challenges and opportunities

How can we advance clinical NLP In languages other than English?

- Mapping NLP efforts:
  - Track progress through literature review
- Resource development
  - Terminologies: increase coverage of the UMLS
  - Annotated corpus
- Clinical Shared Tasks
  - CLEF 2018: ICD10 coding for French, Hungarian, and Italian
  - BARR 2017: abbreviation resolution in Spanish
- Support the creation of modular, multilingual NLP suites

Thank you!

CABeRneT ANR-13-JS02-0009-01

EU H2020 Marie Sklodowska-Curie grant No 676207