Aurélie Névéol

Clinical Natural Language Processing for Languages other than English

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
Campus Universitaire Bât. 508 - F-91405 Orsay Cedex
www.limsi.fr

Why address a variety of languages?

- Access to a larger demographic
  - Access to more patient cohorts
  - Aggregate 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. LREC2018]**

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

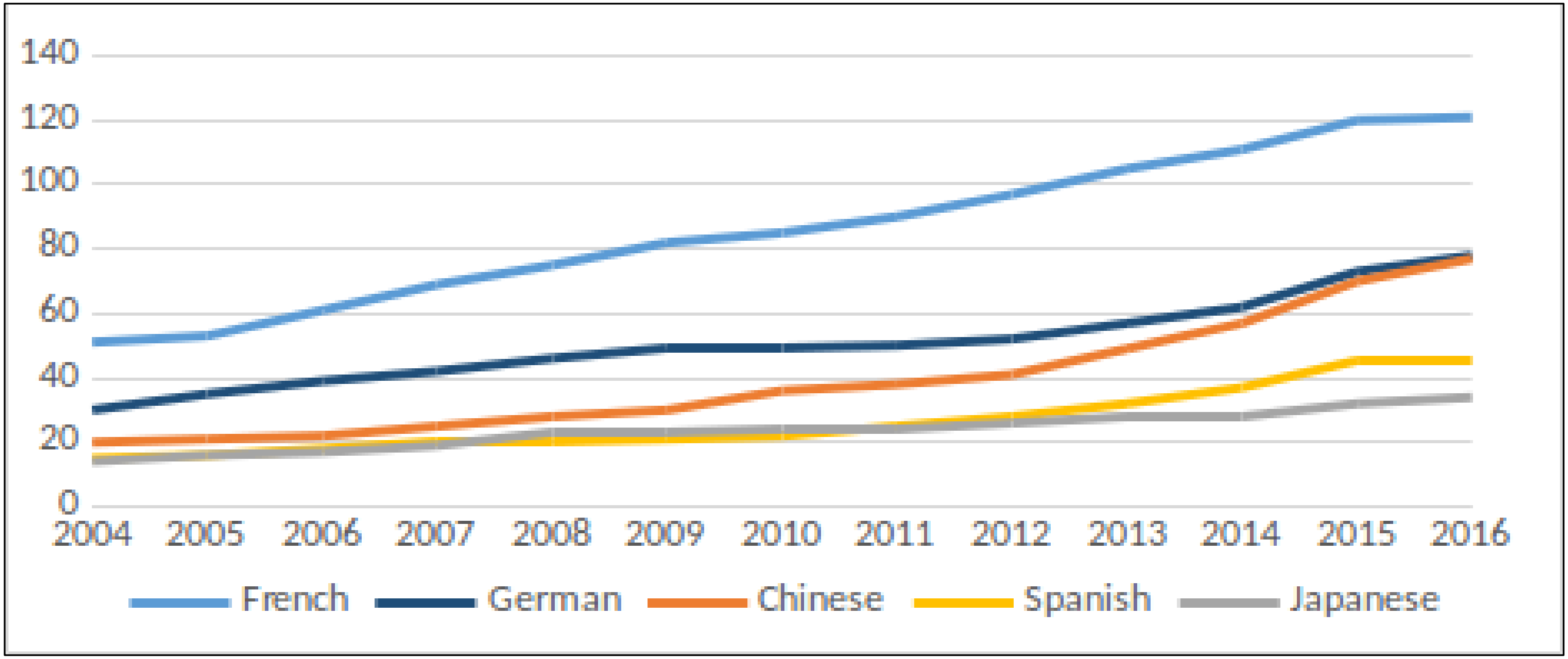
Background: Natural language processing applied to clinical text or aimed at a clinical outcome has been thriving in recent years. This paper offers the first broad overview of clinical Natural Language Processing (NLP) for languages other than English. Recent studies are summarized to offer insights 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
  - Morphologic analyzer : French [Namer and Zweigenbaum 2004]
  - PoS tagger: Portuguese [Oleynik et al. 2010], Polish [Marciniak and Mykowiecka 2011], Spanish [Costumero et al. 2014]
  - Entity and concept recognition: no equivalent of Metamap or cTAKES; some tools for direct lexical matching, e.g. BioPortal [Jonquet et al.]
- Lexicon and terminology development
  - Term translation [Deléger et al. 2010; van Mulligen et al. 2016] mapping of terminologies to the UMLS [Bousquet 2012;Tapi Nzali et al. 2017]

Building new systems and resources (specific to some languages)

- Word segmentation and applications
  - Clinical entity recognition in Chinese [Lei et al. JAMIA 2014], [Xu et al. JAMIA 2014]
  - Word segmentation in Japanese [Nishimoto et al. Methods Inf Med 2008] → **can be usefully transferred back to English (OCR)**
- Transliteration
  - Expansion of English abbreviations in Japanese [Shinohara et al. Methods Inf Med 2013]
  - Identification of translitterated words for word segmentation in Hebrew [Cohen et al. Methods Inf Med 2010]

Corpora and annotations (for Romance Languages)

- Monolingual Corpora

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

- Parallel Corpora
  - Scielo, EDP, UFAL, … [Névéol et al. 2018]

Adapting NLP architectures developed for English

- 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
  - Relation extraction, English to French [Tourille et al. 2017]
  - Temporal expressions: Swedish [Vellupilai et al 2014] French [Tapi Nzali et al 2015]

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

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, 2018: abbreviation resolution in Spanish
- Support the creation of modular, multilingual NLP suites

Thank you!

Hercules Dalianis, Sumithra Velupillai,
Guergana Savova, Pierre Zweigenbaum

CABeRneT ANR-13-JS02-0009-01

EU H2020 Marie Sklodowska-Curie grant No 676207