Traitement automatique des langues pour la santé : travaux récents au ~~LIMSI~~ LISN

Pierre Zweigenbaum
avec Cyril Grouin, Thierry Hamon, Aurélie Névéol, François Yvon *et al.*

Université Paris-Saclay, CNRS, LISN

TLH-Santé 2021 : La santé et le langage
En ligne, 4/2/2021

→ LISN

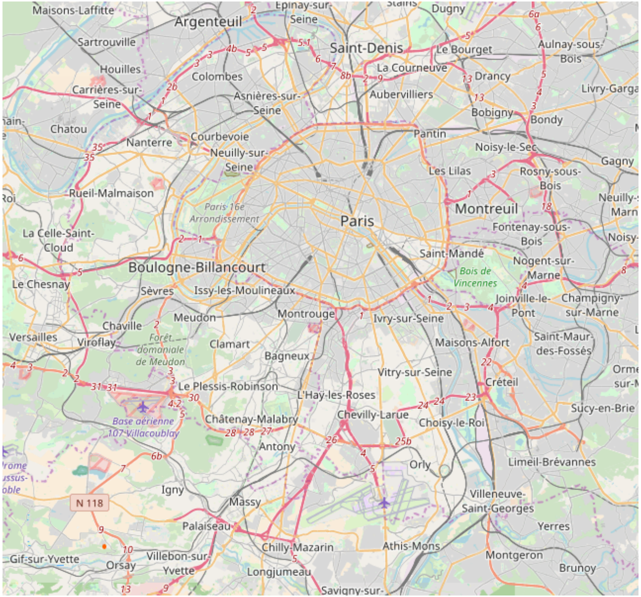
Plan

1. Introduction: LISN, TAL pour la santé
2. TAL pour les professionnels de santé
3. TAL pour les chercheurs en santé
4. TAL pour les étudiants en médecine
5. Textes biomédicaux pour chercheurs en TAL
6. Plongements lexicaux en domaine spécialisé

Plan

1. Introduction: LISN, TAL pour la santé
    - LISN, CNRS, Université Paris-Saclay
    - NLP for e-Health
2. TAL pour les professionnels de santé
    - Désidentification de textes cliniques
    - Normalisation de l'information pour les statistiques de santé
    - Extraction d'information dans les médias sociaux pour la pharmacovigilance
3. TAL pour les chercheurs en santé
    - Natural Language Processing for Systematic Reviews
    - Machine Translation of Biomedical Abstracts
    - Information Extraction from Biomedical Text
4. TAL pour les étudiants en médecine
    - Dialogue personne-machine pour entrainer les étudiants en médecine
5. Textes biomédicaux pour chercheurs en TAL
    - Création de corpus de textes biomédicaux
6. Plongements lexicaux en domaine spécialisé
    - Plongement d'ontologie pour le liage référentiel

LISN in the Greater Paris Area



Limsi

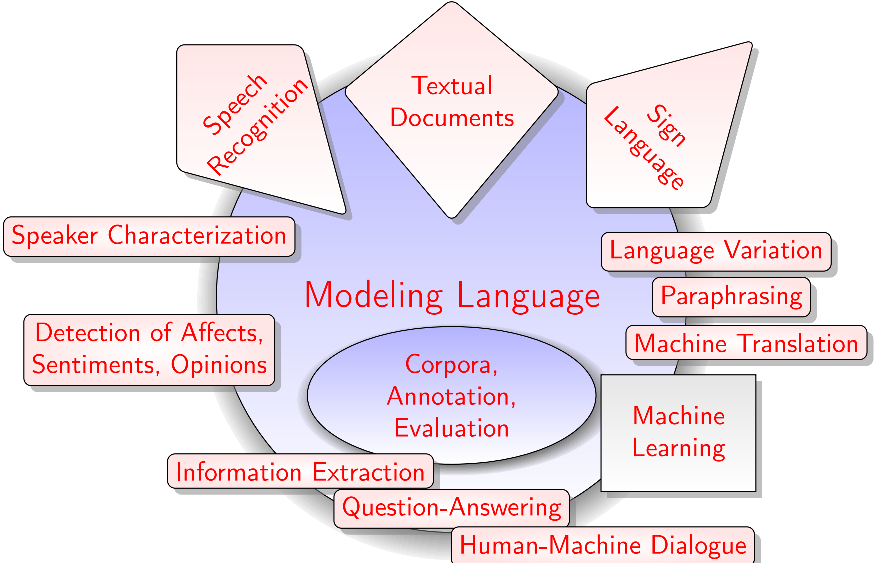
LISN: An Interdisciplinary Laboratory for Digital Sciences

`https://www.lisn.upsaclay.fr/`

A joint CNRS Research Unit on the Université Paris-Saclay Campus: Belvédère buildings

LISN's Language Sciences and Technologies Department

2 teams, about 70 people including 28 permanent researchers



Natural Language Processing of Biomedical Text

Biomedical texts convey information and knowledge

- Natural Language Processing of these texts unlocks that information and knowledge for further computer processing

Overall approach

- Supervised machine learning
- Self-supervised word representations
- Make use of rich pre-existing knowledge in the medical domain:
    - terminologies, ontologies

Natural Language Processing for e-Health

Multiple needs

- Health care professionals
  - Obtain information from patient records
- Health care researchers
  - Obtain knowledge from the literature
- Medical students
  - Receive assistance for training
- Patients
  - Access understandable documentation

General Bibliography: Surveys

- Grouin C, Grabar N; Section Editors for the IMIA Yearbook Section on Natural Language Processing. A Year of Papers Using Biomedical Texts. Yearb Med Inform. 2020 Aug;29(1):221-225. doi: 10.1055/s-0040-1701997.
- Hahn U, Oleynik M. Medical Information Extraction in the Age of Deep Learning. Yearb Med Inform. 2020 Aug;29(1):208-220. doi: 10.1055/s-0040-1702001.
- Névéol A, Zweigenbaum P. Making sense of big textual data for health care: Findings from the section on clinical natural language processing. *Yearb Med Inform*, 26(01):228-233, 2017.
- Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. *J Biomed Semantics*. 2018 Mar 30;9(1):12.
- Demner-Fushman D, Elhadad N. Aspiring to unintended consequences of natural language processing: A review of recent developments in clinical and consumer-generated text processing. *Yearb Med Inform*. 2016 Nov 10;(1):224-233.
- Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128-44.
- Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief Bioinform*. 2016 Jan;17(1):132-44.
- Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. *Brief Bioinform*. 2007 Sep;8(5):358-75. Epub 2007 Oct 30.

Plan

1. Introduction: LISN, TAL pour la santé
   - LISN, CNRS, Université Paris-Saclay
   - NLP for e-Health
2. TAL pour les professionnels de santé
   - Désidentification de textes cliniques
   - Normalisation de l'information pour les statistiques de santé
   - Extraction d'information dans les médias sociaux pour la pharmacovigilance
3. TAL pour les chercheurs en santé
   - Natural Language Processing for Systematic Reviews
   - Machine Translation of Biomedical Abstracts
   - Information Extraction from Biomedical Text
4. TAL pour les étudiants en médecine
   - Dialogue personne-machine pour entrainer les étudiants en médecine
5. Textes biomédicaux pour chercheurs en TAL
   - Création de corpus de textes biomédicaux
6. Plongements lexicaux en domaine spécialisé
   - Plongement d'ontologie pour le liage référentiel

Plan

1. Introduction: LISN, TAL pour la santé
   - LISN, CNRS, Université Paris-Saclay
   - NLP for e-Health
2. TAL pour les professionnels de santé
   - Désidentification de textes cliniques
   - Normalisation de l'information pour les statistiques de santé
   - Extraction d'information dans les médias sociaux pour la pharmacovigilance
3. TAL pour les chercheurs en santé
   - Natural Language Processing for Systematic Reviews
   - Machine Translation of Biomedical Abstracts
   - Information Extraction from Biomedical Text
4. TAL pour les étudiants en médecine
   - Dialogue personne-machine pour entrainer les étudiants en médecine
5. Textes biomédicaux pour chercheurs en TAL
   - Création de corpus de textes biomédicaux
6. Plongements lexicaux en domaine spécialisé
   - Plongement d'ontologie pour le liage référentiel

Text De-identification

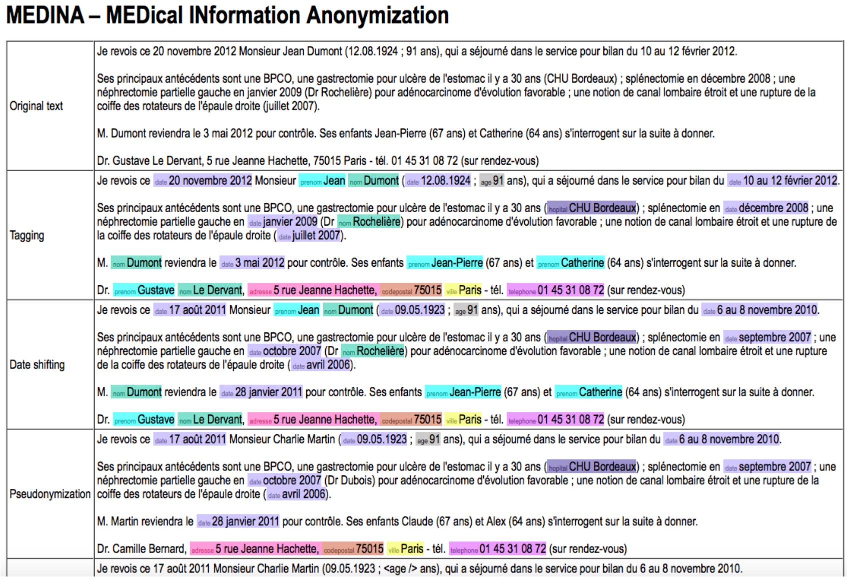
An entity detection task

- MEDINA system
  - Cyril Grouin
  - Rule-based version (`https://medina.limsi.fr/`)
  - Supervised learning version (`https://github.com/grouin/medina`)
- European Project CEF MAPA
  - Manuel Herranz (Pangeanic), Montse Cuadros (Vicomtech), . . . , Lucie Gianola , Cyril Grouin , Thomas Lavergne , Patrick Paroubek , Pierre Zweigenbaum
  - Multilingual BERT for entity detection

---

[1] Grouin C, Névéol A. De-Identification of Clinical Notes in French: towards a Protocol for Reference Corpus Development J Biomed Inform. 2014 Aug;50:151-61.
Eriks Ajausks, Victoria Arranz, Laurent Bié, Aleix Cerdà-i-Cucó, Khalid Choukri, Montse Cuadros, Hans Degroote, Amando Estela, Thierry Etchegoyhen, Mercedes García-Martínez, Aitor García-Pablos, Manuel Herranz, Alejandro Kohan, Maite Melero, Mike Rosner, Roberts Rozis, Patrick Paroubek, Artūrs Vasilevskis, Pierre Zweigenbaum. The Multilingual Anonymisation Toolkit for Public Administrations (MAPA) Project. EAMT, Lisboa, Portugal, 2020:471–472

De-identification of Clinical Text



MEDINA – MEDical INformation Anonymization

Original text	Je revois ce 20 novembre 2012 Monsieur Jean Dumont (12.08.1924 ; 91 ans), qui a séjourné dans le service pour bilan du 10 au 12 février 2012. Ses principaux antécédents sont une BPCO, une gastrectomie pour ulcère de l'estomac il y a 30 ans (CHU Bordeaux) ; splénectomie en décembre 2008 ; une néphrectomie partielle gauche en janvier 2009 (Dr Rochelière) pour adénocarcinome d'évolution favorable ; une notion de canal lombaire étroit et une rupture de la coiffe des rotateurs de l'épaule droite (juillet 2007). M. Dumont reviendra le 3 mai 2012 pour contrôle. Ses enfants Jean-Pierre (67 ans) et Catherine (64 ans) s'interrogent sur la suite à donner. Dr. Gustave Le Dervant, 5 rue Jeanne Hachette, 75015 Paris - tél. 01 45 31 08 72 (sur rendez-vous)
Tagging	Je revois ce date 20 novembre 2012 Monsieur prenom Jean nom Dumont (date 12.08.1924 ; age 91 ans), qui a séjourné dans le service pour bilan du date 10 au 12 février 2012. Ses principaux antécédents sont une BPCO, une gastrectomie pour ulcère de l'estomac il y a 30 ans (hopital CHU Bordeaux) ; splénectomie en date décembre 2008 ; une néphrectomie partielle gauche en date janvier 2009 (Dr nom Rochelière) pour adénocarcinome d'évolution favorable ; une notion de canal lombaire étroit et une rupture de la coiffe des rotateurs de l'épaule droite (date juillet 2007). M. nom Dumont reviendra le date 3 mai 2012 pour contrôle. Ses enfants prenom Jean-Pierre (67 ans) et prenom Catherine (64 ans) s'interrogent sur la suite à donner. Dr. prenom Gustave nom Le Dervant, adresse 5 rue Jeanne Hachette, codepostal 75015 ville Paris - tél. telephone 01 45 31 08 72 (sur rendez-vous)
Date shifting	Je revois ce date 17 août 2011 Monsieur prenom Jean nom Dumont (date 09.05.1923 ; age 91 ans), qui a séjourné dans le service pour bilan du date 6 au 8 novembre 2010. Ses principaux antécédents sont une BPCO, une gastrectomie pour ulcère de l'estomac il y a 30 ans (hopital CHU Bordeaux) ; splénectomie en date septembre 2007 ; une néphrectomie partielle gauche en date octobre 2007 (Dr nom Rochelière) pour adénocarcinome d'évolution favorable ; une notion de canal lombaire étroit et une rupture de la coiffe des rotateurs de l'épaule droite (date avril 2006). M. nom Dumont reviendra le date 28 janvier 2011 pour contrôle. Ses enfants prenom Jean-Pierre (67 ans) et prenom Catherine (64 ans) s'interrogent sur la suite à donner. Dr. prenom Gustave nom Le Dervant, adresse 5 rue Jeanne Hachette, codepostal 75015 ville Paris - tél. telephone 01 45 31 08 72 (sur rendez-vous)
Pseudonymization	Je revois ce date 17 août 2011 Monsieur Charlie Martin (date 09.05.1923 ; age 91 ans), qui a séjourné dans le service pour bilan du date 6 au 8 novembre 2010. Ses principaux antécédents sont une BPCO, une gastrectomie pour ulcère de l'estomac il y a 30 ans (hopital CHU Bordeaux) ; splénectomie en date septembre 2007 ; une néphrectomie partielle gauche en date octobre 2007 (Dr Dubois) pour adénocarcinome d'évolution favorable ; une notion de canal lombaire étroit et une rupture de la coiffe des rotateurs de l'épaule droite (date avril 2006). M. Martin reviendra le date 28 janvier 2011 pour contrôle. Ses enfants Claude (67 ans) et Alex (64 ans) s'interrogent sur la suite à donner. Dr. Camille Bernard, adresse 5 rue Jeanne Hachette, codepostal 75015 ville Paris - tél. telephone 01 45 31 08 72 (sur rendez-vous)
	Je revois ce 17 août 2011 Monsieur Charlie Martin (09.05.1923 ; <age /> ans), qui a séjourné dans le service pour bilan du 6 au 8 novembre 2010.

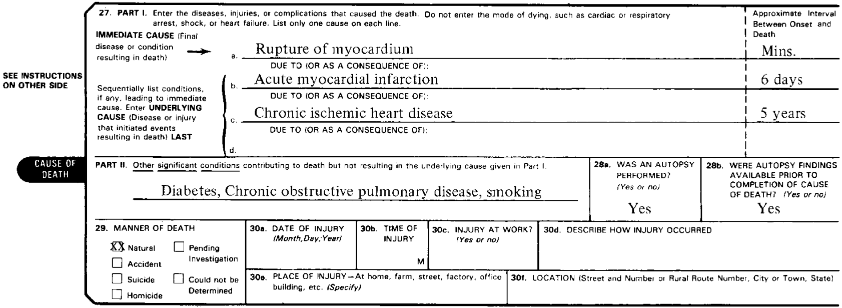
Coding Causes of Death

- Classification: with Thomas Lavergne (LISN)
- CLEF eHealth 2016–2018 Shared Tasks with
  - Cyril Grouin , Thomas Lavergne , Aurélie Névéol (LISN)
  - Aude Robert, Grégoire Rey (INSERM, French CépiDc data)
  - Robert Anderson (US CDC data)
  - Francesco Grippo and Chiara Orsi (Italian data)
  - László Pelikán (Hungarian data)

Coding Death Certificates for Health Statistics

Text classification: From the text of a certificate to one or mode labels (codes)

Needed in about every country for public health statistics



SEE INSTRUCTIONS ON OTHER SIDE

CAUSE OF DEATH

27. PART I. Enter the diseases, injuries, or complications that caused the death. Do not enter the mode of dying, such as cardiac or respiratory arrest, shock, or heart failure. List only one cause on each line.

Approximate Interval Between Onset and Death

IMMEDIATE CAUSE (Final disease or condition resulting in death) → a. Rupture of myocardium — Mins.

DUE TO (OR AS A CONSEQUENCE OF):

b. Acute myocardial infarction — 6 days

DUE TO (OR AS A CONSEQUENCE OF):

c. Chronic ischemic heart disease — 5 years

DUE TO (OR AS A CONSEQUENCE OF):

d.

Sequentially list conditions, if any, leading to immediate cause. Enter **UNDERLYING CAUSE** (Disease or injury that initiated events resulting in death) **LAST**

PART II. Other significant conditions contributing to death but not resulting in the underlying cause given in Part I.

Diabetes, Chronic obstructive pulmonary disease, smoking

28a. WAS AN AUTOPSY PERFORMED? *(Yes or no)*

Yes

28b. WERE AUTOPSY FINDINGS AVAILABLE PRIOR TO COMPLETION OF CAUSE OF DEATH? *(Yes or no)*

Yes

29. MANNER OF DEATH

☒ Natural
☐ Accident
☐ Suicide
☐ Homicide
☐ Pending Investigation
☐ Could not be Determined

30a. DATE OF INJURY *(Month, Day, Year)*

30b. TIME OF INJURY

M

30c. INJURY AT WORK? *(Yes or no)*

30d. DESCRIBE HOW INJURY OCCURRED

30e. PLACE OF INJURY—At home, farm, street, factory, office building, etc. *(Specify)*

30f. LOCATION (Street and Number or Rural Route Number, City or Town, State)

I23.3 Rupture of cardiac wall without haemopericardium as current complication following acute myocardial infarction (伴随急性心肌梗塞的无即期心包积血的心壁破裂)

I21.9 Acute myocardial infarction, unspecified (未明确的急性心肌梗塞)

I25.9 Chronic ischaemic heart disease, unspecified (未明确的慢性缺血性心脏病)

Challenges

- Source language
  - Large variation in expression
  - Context-dependent coding
  - Spelling errors, abbreviations
- Target ontology
  - Large number of target concepts ($> 3,000$ disease codes)
  - Highly unbalanced examples
  - Concept labels use meta-language

Methods: Supervised machine learning

- ‘Pre-processing’: Normalize text as much as possible
    - Spelling correction
    - Compound word normalization
    - Stemming
- Generate features:
    - Text features: token and character n-grams of normalized words
    - Predictions by lookup in pre-existing dictionary
    - Structured data features
        - Age of the deceased
        - Coding year
- Multi-label classification
    - SVM, one-vs-all

Outcomes

Top results on CLEF eHealth 2017 data

- English (US) death certificates
  - P=90, R=81, F=85
- French death certificates
  - P=89, R=77, F=83
- French death certificates, line-aligned
  - P=87, R=87, F=87

Collaboration with users for take-up

- CepiDc: National epidemiology center on medical causes of death (WHO coding center)
  - Insert into coding workflow
- Santé Publique France: National public health agency
  - Detect "syndromic groups" for early outbreak surveillance and public health monitoring

Pharmacovigilance from Social Media

François Morlane-Hondère, Leonardo Campillos-Llanos, Cyril Grouin (LISN)
Funding: ANSM: National Agency for Drug Safety

Vigi4med (2014–2017)

- PI: Dr Cédric Bousquet
  - St-Étienne Medical School
  - LIMICS, Inserm, Paris
- 2 pharmacovigilance centers
- Detect adverse drug reactions
- Proof of concept, batch processing

PHARES (2017–2019)

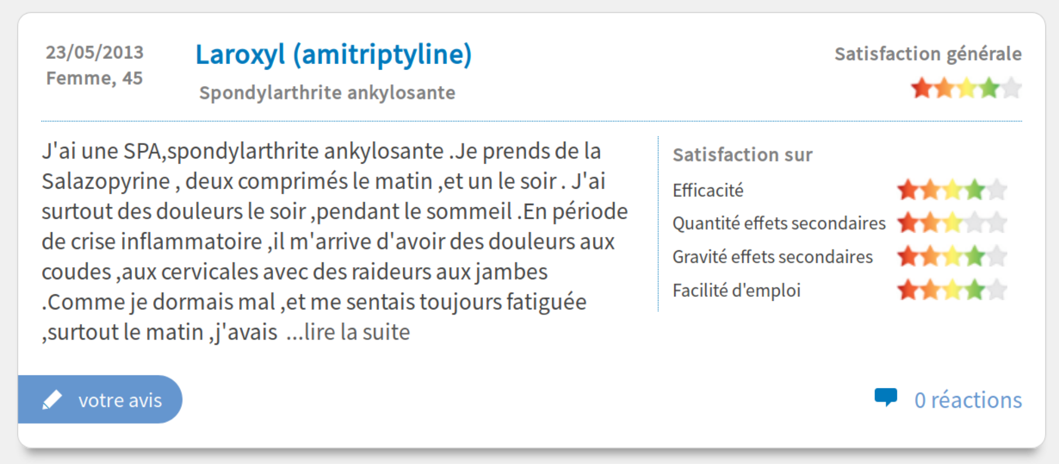
- PI: Dr Marie-Christine Jaulent
  - LIMICS, Inserm, Paris
- 1 pharmacovigilance center
- Strengthen causal link detection
- Explore drug misuse
- On-line processing

LISN: NLP

Follow-up projects: BiFold (DFKI, Berlin); ANR-DFG-JST KEEPHA (DFKI, Berlin; Riken, Tokyo)

Social Networks

Patient forums: New sources of information on health



23/05/2013
Femme, 45

Laroxyl (amitriptyline)

Spondylarthrite ankylosante

Satisfaction générale ★★★★☆

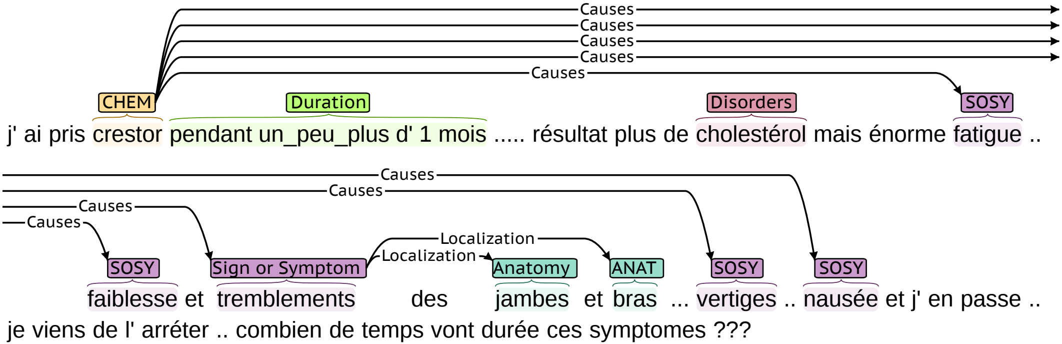
J'ai une SPA,spondylarthrite ankylosante .Je prends de la Salazopyrine , deux comprimés le matin ,et un le soir . J'ai surtout des douleurs le soir ,pendant le sommeil .En période de crise inflammatoire ,il m'arrive d'avoir des douleurs aux coudes ,aux cervicales avec des raideurs aux jambes .Comme je dormais mal ,et me sentais toujours fatiguée ,surtout le matin ,j'avais ...lire la suite

Satisfaction sur

Efficacité	★★★★☆
Quantité effets secondaires	★★★☆☆
Gravité effets secondaires	★★★★☆
Facilité d'emploi	★★★★☆

Objectives

Spotting adverse drug event signals: pharmacovigilance, needed for drug safety



Causes
Causes
Causes
Causes
Causes

CHEM Duration Disorders SOSY

j' ai pris crestor pendant un_peu_plus d' 1 mois ..... résultat plus de cholestérol mais énorme fatigue ..

Causes
Causes
Causes
Causes
Localization
Localization

SOSY Sign or Symptom Anatomy ANAT SOSY SOSY

faiblesse et tremblements des jambes et bras ... vertiges .. nausée et j' en passe ..

je viens de l' arréter .. combien de temps vont durée ces symptomes ???

Challenges

- Source language
  - Non-standard vocabulary
  - Lay vocabulary
  - Creative use of language on social networks
- Target ontology
  - Large number of target classes
  - MedDRA: $> 22,000$ sign and symptom (PT) codes

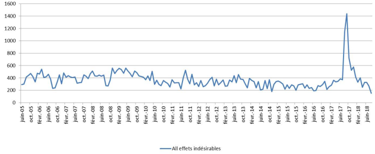
Methods

- Large terminology resources: e.g., drug names
  - Collect based on UMLS semantic classes
- Robustly learn to recognize non-standard vocabulary
  - Word embeddings to collect non-standard variants
    → gynécologue, gynécolgue, gynécoloque, gynéco, gynécho, génico, gyné, gygy, etc.
  - Leverage syntactic parsing
    - Drug-taking verbs
    - I take aspirin every day
    → prendre, prescrire, être sous, donner, avoir. . .
  → Drug names absent from our lists
    - e.g., food complements, Canadian drug names
  → Intentional variants
    - Lévothyrox → lévo
    - Duphaston → dudu
    - pilule du lendemain → pdl
  → Spelling errors
- Supervised machine learning for entity and relation detection

Outcomes

Example

- Analyzed 100 million posts
- Studies published by pharmacovigilance experts
  - Example: Baclophen study[1]
  - Example: Levothyrox study[2]
- Providing server to National Agency for Drug Safety



---

[1] Lillo-Le Louet A, et al. *Baclofen safety and its use in social media: a preliminary study.* International Society of Pharmacovigilance, Oct 2017

[2] Audeh B, Grouin C, Zweigenbaum P, Bousquet C, Jaulent MC, Benkebil M, Lillo-Le Louet A. *French Levothyrox® Crisis: Retrospective Analysis of French web forums dedicated to thyroid pathology.* International Society of Pharmacovigilance, Oct 2019

Baclofen study

- 18570 posts mentioning baclofen and a clinical finding retrieved from NLP output database
  - Exclude alcohol dependence → 2621 posts
- Reviewed by two evaluators from pharmacovigilance center
  - 782 (29.2%) mentioned potential adverse drug reactions
- Compared to cases in French National Database
  - 717 ADR cases with baclofen
- Main findings
  - Information about baclofen dosage more frequent in posts (27.7% vs 9.2%)
  - 75% of the patients had a daily dosage over 90 mg, higher than the maximum recommended daily dosage
  - Demographic information (gender, age) scarce in posts

---

[0] Lillo-Le Louet A, et al. *Baclofen safety and its use in social media: a preliminary study.* International Society of Pharmacovigilance, Oct 2017

Levothyrox study

First retrospective analysis of social media following a drug health crisis

- Levothyrox: used by 3 million patients in France
    - New Levothyrox formulation introduced in March 2017
- Examined all posts (900,000) published in the forums of the *Vivre sans thyroïde* Web site
- Posts mentioning Levothyrox or levothyroroxine and a clinical finding were retrieved from NLP output database
    - Identified medical problems most mentioned with Levothyrox
    - Reviewed 100 most mentioned verbatims to create categories
    - Distribution and temporal evolution of problem categories

---

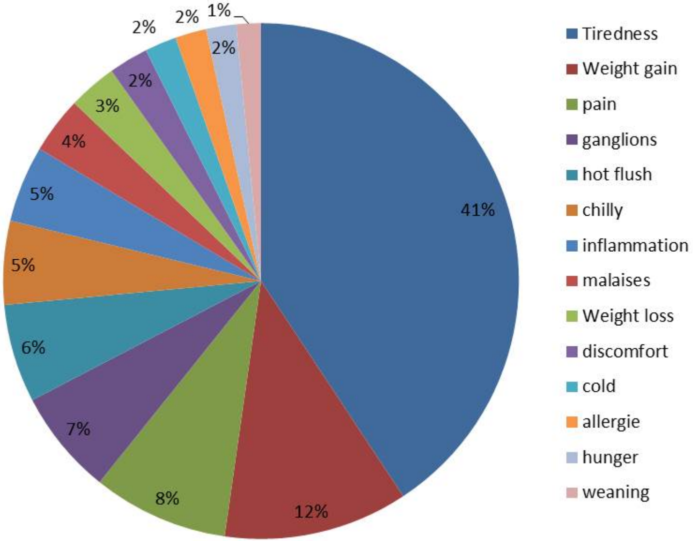
[0] Audeh B, Grouin C, Zweigenbaum P, Bousquet C, Jaulent MC, Benkebil M, Lillo-Le Louet A. *French Levothyrox® Crisis: Retrospective Analysis of French web forums dedicated to thyroid pathology.* International Society of Pharmacovigilance, Oct 2019

Levothyrox: Distribution of problem categories

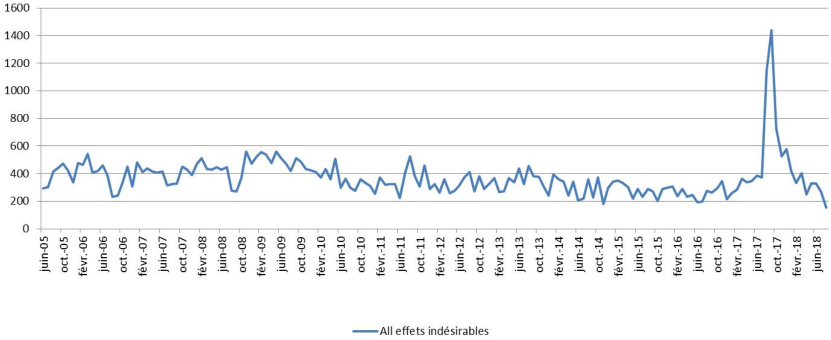
Problem Category	**Examples**	**%**
Drug Indication	*nodules, thyroid cancer*	25%
General event	*inflammation, tiredness*	25%
Neuro-psychiatric event	*depression, stress*	18%
Cardiovascular event	*hypertension, palpitation*	5%
Osteomuscular event	*cramps, osteoporosis*	
Gastrointestinal event	*nausea, constipation*	
Miscellaneous	*emergency, overdose*	
incl. Seriousness	*emergency, crisis*	7%

Levothyrox: Most mentioned problems

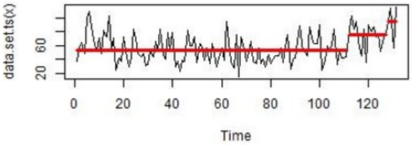
read clockwise



Levothyrox: Temporal evolution of number of adverse effects



Time-series analysis using change-point analysis (CPA):



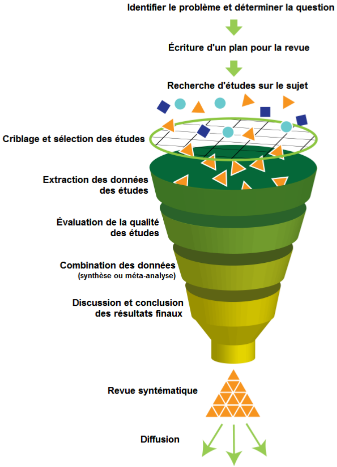
Plan

1. Introduction: LISN, TAL pour la santé
    - LISN, CNRS, Université Paris-Saclay
    - NLP for e-Health
2. TAL pour les professionnels de santé
    - Désidentification de textes cliniques
    - Normalisation de l'information pour les statistiques de santé
    - Extraction d'information dans les médias sociaux pour la pharmacovigilance
3. TAL pour les chercheurs en santé
    - Natural Language Processing for Systematic Reviews
    - Machine Translation of Biomedical Abstracts
    - Information Extraction from Biomedical Text
4. TAL pour les étudiants en médecine
    - Dialogue personne-machine pour entrainer les étudiants en médecine
5. Textes biomédicaux pour chercheurs en TAL
    - Création de corpus de textes biomédicaux
6. Plongements lexicaux en domaine spécialisé
    - Plongement d'ontologie pour le liage référentiel

Natural Language Processing for Systematic Reviews

A Multiple-Stage Process

- Christopher Norman (defended in Feb 2020), Mariska Leeflang (U. Amsterdam), Aurélie Névéol
- European Project MiRoR
- Motivation: explore NLP methods to assist with systematic review preparation
  - Focus on Diagnostic Test Accuracy (DTA) studies
- Practical user needs
  - International Collaboration for the Automation of Systematic Reviews (ICASR)
  - Update of the Core Outcome Measures in Effectiveness Trials (COMET) database https://www.comet-initiative.org/



Identifier le problème et déterminer la question

Écriture d'un plan pour la revue

Recherche d'études sur le sujet

Criblage et sélection des études

Extraction des données des études

Évaluation de la qualité des études

Combination des données (synthèse ou méta-analyse)

Discussion et conclusion des résultats finaux

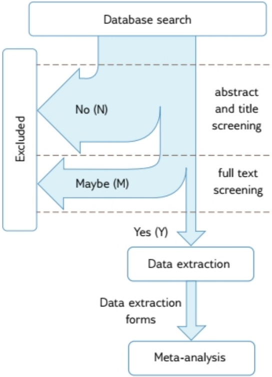
Revue syntématique

Diffusion

Text Classification for Systematic Reviews

Article Triage for Systematic Reviews

- A highly imbalanced classification problem
  - Analyzed the contribution of available supervision information
  - Addressed reviews in different subfields
  - Studied the impact on resulting meta-analysis



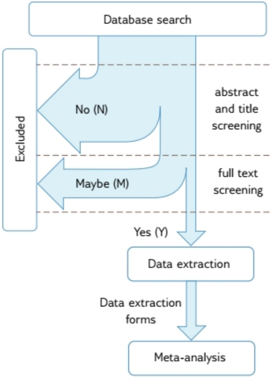
---

[1] Norman C, Leeflang M, Zweigenbaum P, Névéol A. Automating Document Discovery in the Systematic Review Process: How to Use Chaff to Extract Wheat. Language and Resource Evaluation Conference, LREC 2018. 2018:3681-3687
Norman CR, Gargon E, Leeflang MMG, Névéol A, Williamson PR. Evaluation of an automatic article selection method for timelier updates of the Comet Core Outcome Set database. Database (Oxford). 2019 Jan 1;2019. pii: baz109.
Norman C, Leeflang M, Porcher R, Névéol A. Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. Syst Rev. 2019 Oct 28;8(1):243.

Information Extraction for Systematic Reviews

Data Extraction for Systematic Reviews

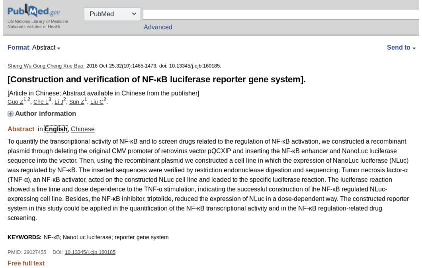
- Data extraction with low supervision
  - Created a dataset for studying data extraction
  - Evaluated the reproducibility of data extraction and meta-analysis
  - Assessed automatic extraction of Index Test, Reference Standard and Target Condition



---

[1] Norman C, Leeflang M, Névéol A. Data Extraction and Synthesis in Systematic Reviews of Diagnostic Test Accuracy: A Corpus for Automating and Evaluating the Process. Proc AMIA Annu Symp. 2018:817-826.
Norman C, Leeflang M, Spijker R, Kanoulas E, Névéol A. A distantly supervised dataset for automated data extraction from diagnostic studies. ACL BioNLP Workshop. 2019:105-114.

Machine Translation of Scientific Abstracts


PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Advanced
Format: Abstract
Send to

Sheng Wu Gong Cheng Xue Bao. 2016 Oct 25;32(10):1465-1473. doi: 10.13345/j.cjb.160185.

[Construction and verification of NF-κB luciferase reporter gene system].

[Article in Chinese; Abstract available in Chinese from the publisher]

Guo Z^{1,2}, Che L^{3}, Li J^{2}, Sun Z^{1}, Liu C^{2}.

Author information

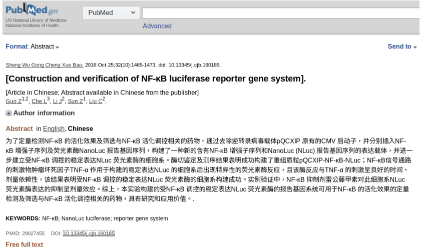
Abstract in **English**, Chinese

To quantify the transcriptional activity of NF-κB and to screen drugs related to the regulation of NF-κB activation, we constructed a recombinant plasmid through deleting the original CMV promoter of retrovirus vector pQCXIP and inserting the NF-κB enhancer and NanoLuc luciferase sequence into the vector. Then, using the recombinant plasmid we constructed a cell line in which the expression of NanoLuc luciferase (NLuc) was regulated by NF-κB. The inserted sequences were verified by restriction endonuclease digestion and sequencing. Tumor necrosis factor-α (TNF-α), an NF-κB activator, acted on the constructed NLuc cell line and leaded to the specific luciferase reaction. The luciferase reaction showed a fine time and dose dependence to the TNF-α stimulation, indicating the successful construction of the NF-κB regulated NLuc-expressing cell line. Besides, the NF-κB inhibitor, triptolide, reduced the expression of NLuc in a dose-dependent way. The constructed reporter system in this study could be applied in the quantification of the NF-κB transcriptional activity and in the NF-κB regulation-related drug screening.

KEYWORDS: NF-κB; NanoLuc luciferase; reporter gene system

PMID: 29027455 DOI: 10.13345/j.cjb.160185

Free full text


PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Advanced
Format: Abstract
Send to

Sheng Wu Gong Cheng Xue Bao. 2016 Oct 25;32(10):1465-1473. doi: 10.13345/j.cjb.160185.

[Construction and verification of NF-κB luciferase reporter gene system].

[Article in Chinese; Abstract available in Chinese from the publisher]

Guo Z^{1,2}, Che L^{3}, Li J^{2}, Sun Z^{1}, Liu C^{2}.

Author information

Abstract in English, **Chinese**

为了定量检测NF-κB 的活化效果及筛选与NF-κB 活化调控相关的药物，通过去除逆转录病毒载体pQCXIP 原有的CMV 启动子，并分别插入NF-κB 增强子序列及荧光素酶NanoLuc 报告基因序列，构建了一种新的含有NF-κB 增强子序列和NanoLuc (NLuc) 报告基因序列的表达载体，并进一步建立受NF-κB 调控的稳定表达NLuc 荧光素酶的细胞系。酶切鉴定及测序结果表明成功构建了重组质粒pQCXIP-NF-κB-NLuc；NF-κB信号通路的刺激物肿瘤坏死因子TNF-α 作用于构建的稳定表达NLuc 的细胞系后出现特异性的荧光素酶反应，且该酶反应与TNF-α 的刺激呈良好的时间、剂量依赖性，该结果表明受NF-κB 调控的稳定表达NLuc 荧光素酶的细胞系构建成功。实例验证中，NF-κB 抑制剂雷公藤甲素对此细胞系NLuc 荧光素酶表达的抑制呈剂量效应。综上，本实验构建的受NF-κB 调控的稳定表达NLuc 荧光素酶的报告基因系统可用于NF-κB 的活化效果的定量检测及筛选与NF-κB 活化调控相关的药物，具有研究和应用价值。.

KEYWORDS: NF-κB; NanoLuc luciferase; reporter gene system

PMID: 29027455 DOI: 10.13345/j.cjb.160185

Free full text

- WMT Biomedical Shared Tasks: Aurélie Névéol with many others
- WMT 2020 EN–FR Machine Translation System: François Yvon with Sadaf Abdul Rauf *et al.*

WMT Shared Tasks: Machine Translation of Biomedical Text (2016–2020+)

http://statmt.org/wmt20/biomedical-translation-task.html

- WMT 2020 biomedical task organizers
  - Rachel Bawden, Giorgio Di Nunzio, Christian B Grozea, Iñigo Inanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol , Mariana Neves, Maite Oronoz, Olatz Perez de Viñaspre, Massimo Piccardi, Rolland Roller, Amy Siu, Philip Thomas, Federica Vezzani, Maika Vincente Navarro, Dina Wieman, Lana Yeganova
- Motivation: Improve access to health information for non native speakers
  - Access to "health in my language"
  - Provide writing assistance and leverage tools available for English only
- Data:
  - Literature (translated by authors), other texts (translated by professionals)
  - Aligned parallel corpora in 15 language pairs, to and from English

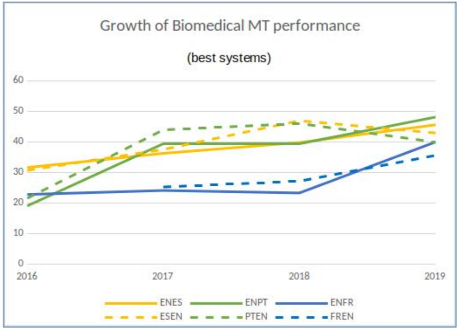
---

[1] Related publications (other than task overviews and participations):
Névéol A, Jimeno Yepes A, Neves ML. MEDLINE as a parallel corpus: a survey to gain insight on French-, Spanish- and Portuguese-speaking authors' abstract writing practice. Language Resources and Evaluation Conference, LREC 2020. 2020:3676-3682.
Névéol A, Jimeno Yepes A, Neves ML, Verspoor K. Parallel Corpora for the Biomedical Domain. Language Resources and Evaluation Conference, LREC 2018. 2018:286-291.

WMT: Overview of the results for recurring language pairs (2016-2019)

Indicative comparison, test sets were different each year

- Participation
  - Up to 20 participating teams from up to 10 countries every year
  - Includes teams with no speaker of the languages involved
- Material and Methods
  - Participants increasingly rely on neural machine translation toolkits
  - A wide variety of in-domain, out-of domain, monolingual and parallel corpora are used



Contributions of the WMT biomedical task

see also: *Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Chapman et al., J Am Med Inform Assoc. 2011 Sep-Oct;18(5):540-3*

- Engage the scientific community on a specialized domain
  - Provide data to study translation in a specialized domain
  - Encourage general purpose translation
- Advance biomedical natural language processing
  - Increase availability and knowledge of biomedical data
  - Develop methods and understanding of NLP processes

LIMSI's WMT 2020 EN–FR Machine Translation System

- Sadaf Abdul Rauf (post-doc), François Yvon *et al.*
- Resource-heavy system to translate medical abstracts from English into French
  - Biomedical corpora
    - Back-translated texts
    - Terminological resources
  - Pre-trained representations
  - Multiple pre-processing pipelines
  - Transformer architecture

---

[1] Sadaf Abdul Rauf, José Carlos Rosales, Pham Minh Quang, François Yvon. LIMSI @ WMT 2020. WMT 2020.

WMT 2020 EN–FR Machine Translation System

Resources

- In-domain parallel corpora
  - Provided by the WMT organizers
  - Additional
    - Cochrane bilingual parallel corpus
    - TAUS Corona Crisis corpus
    - More selected by information retrieval from OPUS corpora News Commentary, Books, and Wikipedia
- In-domain monolingual sources
  - Abstracts of medical papers published by Elsevier from the Lissa portal[1]
  - A collection of research articles collected from various sources[2]
  - Back-translated into French using a neural FR-EN engine trained with the official WMT data sources

---

[1] `https://www.lissa.fr/dc/#env=lissa`

[2] `https://crtt.univ-lyon2.fr/les-corpus-medicaux-du-crtt-613310.kjsp`

WMT 2020 EN–FR Machine Translation System

Methods

- Sentence alignment: Microsoft bilingual aligner (Moore, 2002)
- Tokenization and Byte-pair encoding
  - M: Moses tools and subword-nmt
  - H: HuggingFace BERT API
- FT: Fine-tuning on MEDLINE abstracts and Cochrane corpora
- PT: Pre-translation of medical terms
  - French-English version of the MeSH thesaurus
  - 30k English terms + preferred French translation
  - Extended to 40k by extracting more MeSH term pairs occurring in the training corpus
  - Use the resulting sentence pairs in a second training phase
- Translation framework: seq-2-seq (Facebook's fairseq)
  - Basic Transformers
  - B: BERT-fused transformers (BERT-NMT)

WMT 2020 EN–FR Machine Translation System

Results: BLEU=32.9–37.0

- Test sets: WMT 2018, 2019, 2020 (official)
- Adding additional corpora (1M sentences) to the provided corpora (2.9M sentences): improves BLEU score (+0.6-5.3 BLEU pt)
- Fine-tuning: always improves the Moses pipeline and the 2018 test set
- Back-translated monolingual corpora: small gain (+0.2–0.8 BLEU) for HuggingFace tokenization pipeline
- Pre-translation of medical terms: small drop (–0–0.8 BLEU)
  - The majority were already well translated
  - MeSH translations less similar to “natural” translations
- IR-selected additional corpora: improvement +0.5-0.9 BLEU for HuggingFace pipeline, +0.2–0.8 BLEU for BERT pipeline

Detection of Food-Drug Interaction in Scientific Papers

- Tsanta Randriatsitohaina (3rd year PhD student), Cyril Grouin , Thierry Hamon
- ANR project MIAM
- Example: Grapefruit juice increases effect of other dihydropyridine calcium antagonists
- Sparse information
  - Build annotated corpus with project partners
  - Experiment with Silver Standard corpus created by transfer from annotated drug-drug interaction corpora

Information Extraction for Spin Detection

- Information Extraction for Spin Detection
  - Anna Koroleva (defended in Jan 2020), Patrick Bossuyt (U. Amsterdam), Patrick Paroubek
  - European Project MiRoR
  - The Notion of Spin in Scientific Papers

Drug Repurposing: Detection of Adverse Drug Reactions in Drug Labels

- Jingbo Xia (HZAU), Pierre Zweigenbaum
- Drug repurposing: finding novel uses of drugs
- Hypothesis
  - A cancer immunotherapy drug may induce adverse reactions through off-target protein interactions
  - If there is an identical side-effect phenotype between different drugs, it reveals that these drugs share a target
  - Analyzing the side-effects of drugs can reveal new targets for drug action
  - Entity detection helps collect adverse reactions from DailyMed drug labels
- Results
  - Here, suggests the BTK gene as an off-target site

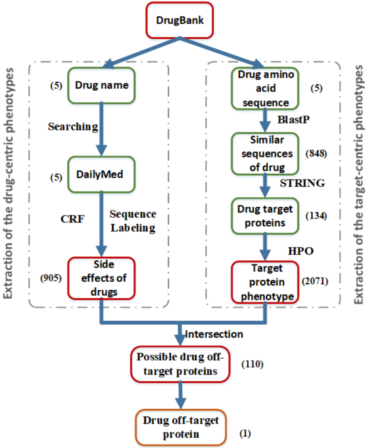
---

[1] Yuyu Zheng, Xiangyu Meng, Pierre Zweigenbaum, Lingling Chen, Jingbo Xia. Hybrid phenotype mining method for investigating off-target protein and underlying side effects of anti-tumor immunotherapy BMC Medical Informatics and Decision Making 2020, 20(Suppl 3):133 https://doi.org/10.1186/s12911-020-1105-4

Entity Detection: a Component in a Larger Drug Repurposing Method

A hybrid phenotype mining method

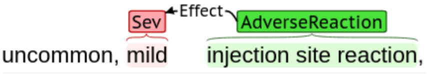
- Select 5 monoclonal antibody drugs with similar mechanism
- Information Extraction
  - Train CRF entity detection on TAC 2017 ADR dataset
  - Use to extract adverse drug reactions from drug labels
- Find target proteins of these drugs and their phenotypes
  - Sequence similarity: find homologous amino acid sequences (BLAST)
  - Expand through protein-protein interaction network (STRIN
  - Obtain phenotype terms from Human Phenotype Ontology
- Select matching adverse drug reactions and phenotype term
  - Among 110 detected, investigated mutation of gene BTK
  - Gene Ontology analysis: filter genes with function related to T-cell proliferation activation, immune regulation
  - Results in 3 genes: AKT1, ACTG2, BTK, the latter more relevant according to a literature search



Training Corpus: TAC 2017 Adverse Drug Reactions in Drug Labels

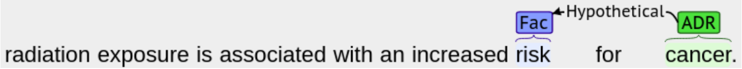
`https://bionlp.nlm.nih.gov/tac2017adversereactions/`

AdverseReaction, Severity, Factor, etc.



uncommon, mild injection site reaction,

Relations between these entities: Effect, Hypothetical, etc.



radiation exposure is associated with an increased risk for cancer.

---

[1] Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 adverse reaction extraction from drug labels track. In: TAC 2017.

Zhou K, Zhang X, Zweigenbaum P, Liang R, Jiang Y, Xia J. Adverse reaction identification driven by semantic information. In: TAC 2017.

Plan

1. Introduction: LISN, TAL pour la santé
   - LISN, CNRS, Université Paris-Saclay
   - NLP for e-Health
2. TAL pour les professionnels de santé
   - Désidentification de textes cliniques
   - Normalisation de l'information pour les statistiques de santé
   - Extraction d'information dans les médias sociaux pour la pharmacovigilance
3. TAL pour les chercheurs en santé
   - Natural Language Processing for Systematic Reviews
   - Machine Translation of Biomedical Abstracts
   - Information Extraction from Biomedical Text
4. TAL pour les étudiants en médecine
   - Dialogue personne-machine pour entrainer les étudiants en médecine
5. Textes biomédicaux pour chercheurs en TAL
   - Création de corpus de textes biomédicaux
6. Plongements lexicaux en domaine spécialisé
   - Plongement d'ontologie pour le liage référentiel

A Virtual Patient to Train Medical Students

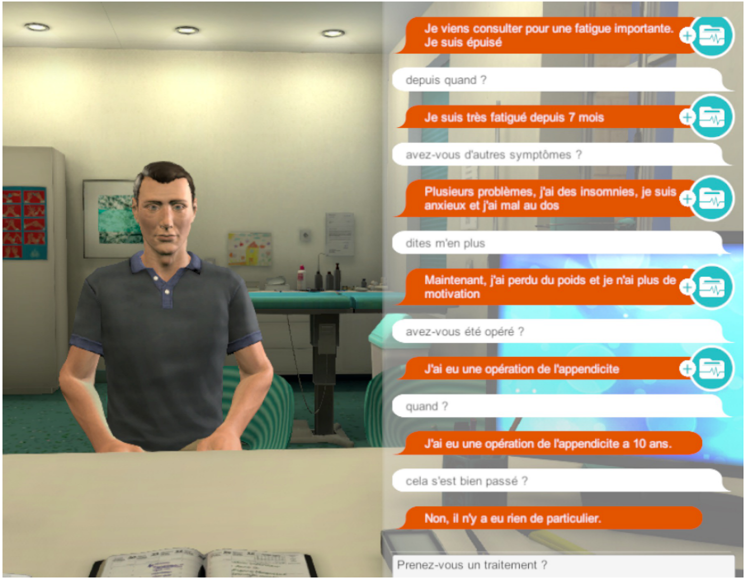
Providing rich specialized terminology and concepts to a dialogue system

- LISN, CNRS
    - Sophie Rosset , Pierre Zweigenbaum , Éric Bilinski
    - Leonardo Campillos Llanos (post-doc, now researcher in Madrid), Catherine Thomas, Dhouha Bouamor
- Funding: FUI / BPI Patient Genesys; SATT Paris-Saclay PVdial
- Partners: Interaction Healthcare / SimForHealth, CHU d'Angers, Voxygen, Vidal

---

[1] Campillos Llanos, L., C. Thomas, É. Bilinski, P. Zweigenbaum, and S. Rosset (2019) Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. Natural Language Engineering, 2019, pp. 1-38.

A Virtual Patient to Train Medical Students



Je viens consulter pour une fatigue importante. Je suis épuisé

depuis quand ?

Je suis très fatigué depuis 7 mois

avez-vous d'autres symptômes ?

Plusieurs problèmes, j'ai des insomnies, je suis anxieux et j'ai mal au dos

dites m'en plus

Maintenant, j'ai perdu du poids et je n'ai plus de motivation

avez-vous été opéré ?

J'ai eu une opération de l'appendicite

quand ?

J'ai eu une opération de l'appendicite a 10 ans.

cela s'est bien passé ?

Non, il n'y a eu rien de particulier.

Prenez-vous un traitement ?

A Virtual Patient to Train Medical Students

English version: D (Doctor) is the user input
P (Patient) sentences are generated by our system in real time

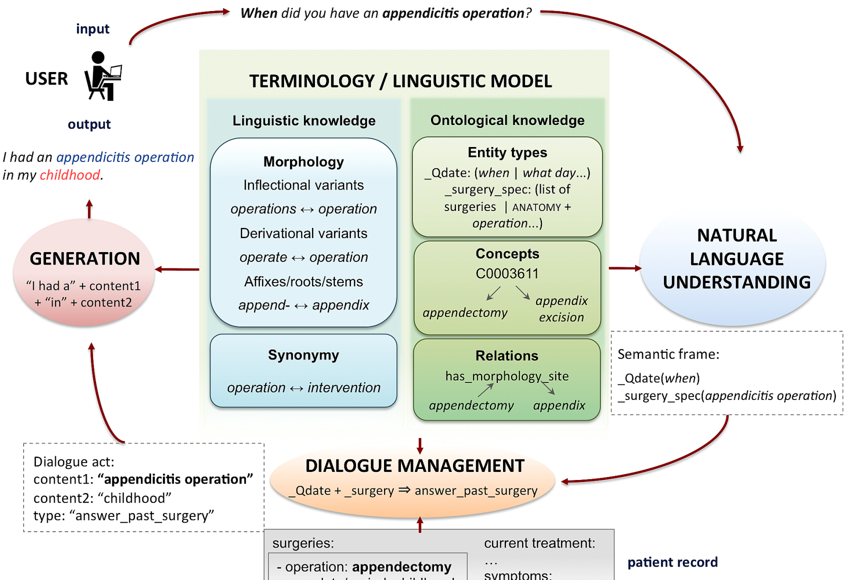
D: *How are you feeling?*	
P: *Several things, I have **chest pain**, fever and sweating.*	LEXICAL SIMPLIFICATION
D: ***And what else?***	PROCESSING INFORMATION STATE
P: *I currently have a cough, yellow sputum and difficulty breathing.*	
D: *Since when are you feeling **feverish**?*	DERIVATIONAL VARIANTS
P: *The fever started minutes after the pain.*	
D: *Describe me your **coughing.***	INFLECTIONAL VARIANTS
P: *I have a dry cough which started around 23 hs.*	
D: *Are you **breathing with difficulty**?*	MULTIWORD ENTITIES
P: *Yes.*	
D: ***Since when**?*	ELLIPSIS
P: *I have difficulty breathing since yesterday night.*	
D: *Tell me more about your **sweating**.*	TERM VARIATION
P: *I **perspire** because of the fever.*	
D: *Have you had chest pain for a long time?*	
P: *I have had chest pain since yesterday evening at 20.*	
D: *Are you allergic to something?*	
P: *I am only allergic to kiwi*	
D: *Do you have any treatment for **that**?*	ANAPHORA
P: *I do not take anything for my allergy.*	
D: *Do you have a **cardiovascular disease**?*	ONTOLOGICAL RELATIONS
P: *I have **hypertension**.*	LEXICAL SIMPLIFICATION
D: *Do you take a pill for your **tension problems**?*	TERM VARIATION
P: *I take an antihypertensive (**oral**).*	
D: *Have you ever been operated?*	LEXICAL SIMPLIFICATION
P: *I had an **appendix operation.***	
D: *When did you have your **appendix** out?*	AFFIXES
P: *I had an appendix operation in my childhood.*	

Limsi

NLP challenges

- Terminology
  - Match user vocabulary to patient case vocabulary
  - Large concept and vocabulary spaces
- Includes question-answering
  - Need for precision
- Dialogue management
  - Context-based processing
  - Instantaneous answer
- Must run on new specialties and cases with no human intervention

System Architecture



When did you have an *appendicitis operation*?

input

USER

output

I had an appendicitis operation in my childhood.

TERMINOLOGY / LINGUISTIC MODEL

Linguistic knowledge

Morphology

Inflectional variants

operations ↔ operation

Derivational variants

operate ↔ operation

Affixes/roots/stems

append- ↔ appendix

Synonymy

operation ↔ intervention

Ontological knowledge

Entity types

_Qdate: (*when* | *what day*...)
_surgery_spec: (list of surgeries | ANATOMY + *operation*...)

Concepts

C0003611

appendectomy *appendix excision*

Relations

has_morphology_site

appendectomy *appendix*

NATURAL LANGUAGE UNDERSTANDING

Semantic frame:
_Qdate(*when*)
_surgery_spec(*appendicitis operation*)

DIALOGUE MANAGEMENT

_Qdate + _surgery ⇒ answer_past_surgery

GENERATION

"I had a" + content1 + "in" + content2

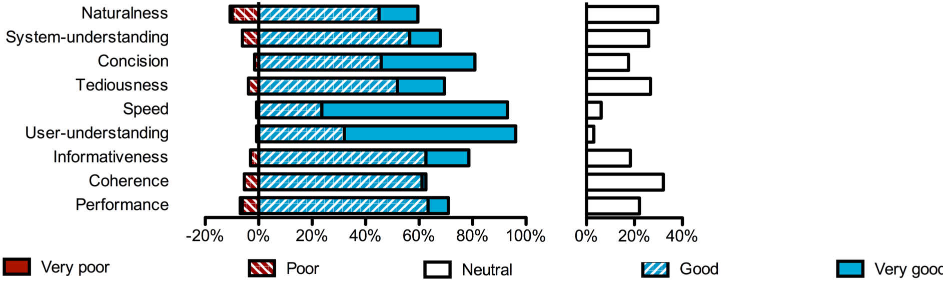
Dialogue act:
content1: **"appendicitis operation"**
content2: "childhood"
type: "answer_past_surgery"

surgeries:
- operation: **appendectomy**

current treatment:
…
symptoms:

patient record

Outcomes

- Functional system
  - Good user evaluations
    - 

  - Robust across medical specialties
  - Robust on unseen patient cases
- Ready for transfer to company

---

[1] Campillos Llanos, L., C. Thomas, É. Bilinski, P. Zweigenbaum, and S. Rosset (2019) Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. Natural Language Engineering, 2019:1–38

Plan

1. Introduction: LISN, TAL pour la santé
   - LISN, CNRS, Université Paris-Saclay
   - NLP for e-Health
2. TAL pour les professionnels de santé
   - Désidentification de textes cliniques
   - Normalisation de l'information pour les statistiques de santé
   - Extraction d'information dans les médias sociaux pour la pharmacovigilance
3. TAL pour les chercheurs en santé
   - Natural Language Processing for Systematic Reviews
   - Machine Translation of Biomedical Abstracts
   - Information Extraction from Biomedical Text
4. TAL pour les étudiants en médecine
   - Dialogue personne-machine pour entrainer les étudiants en médecine
5. **Textes biomédicaux pour chercheurs en TAL**
   - **Création de corpus de textes biomédicaux**
6. Plongements lexicaux en domaine spécialisé
   - Plongement d'ontologie pour le liage référentiel

Annotated Corpora and Shared Tasks

Collection and annotation of text corpora for natural language processing

- Natural Language Processing needs annotated text
    - System development / training
    - System evaluation
    - Organization of shared tasks

Some Annotated Biomedical Corpora (1/3)

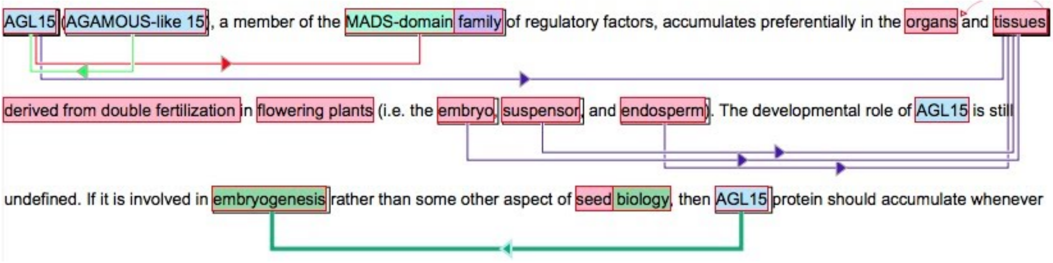
- Translated scientific abstracts
  - WMT 2017–2020 biomedical shared tasks
  - Document-level alignment
  - Aurélie Névéol *et al.*
- Quaero Medical Corpus
  - MEDLINE paper titles, EMEA drug information
  - Entity mentions, 10 types, UMLS concepts (`https://quaerofrenchmed.limsi.fr/`)
  - CLEF eHealth 2016 shared task
  - Aurélie Névéol , Cyril Grouin , Thierry Hamon , Pierre Zweigenbaum *et a*

---

[1] Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing - BioTxtM2014. 2014:24-30

Some Annotated Biomedical Corpora (2/3)

- SeeDEV: plant biology papers
  - Seedev BioNLP shared tasks (2016, 2019) (`http://2016.bionlp-st.org/tasks/seedev`, `https://sites.google.com/view/seedev2019/`)
  - Relation and event extraction
  - Estelle Chaix (INRAE), Claire Nédellec (INRAE), . . . , Pierre Zweigenbaum *et al.*



AGL15 (AGAMOUS-like 15), a member of the MADS-domain family of regulatory factors, accumulates preferentially in the organs and tissues
derived from double fertilization in flowering plants (i.e. the embryo, suspensor, and endosperm). The developmental role of AGL15 is still
undefined. If it is involved in embryogenesis rather than some other aspect of seed biology, then AGL15 protein should accumulate whenever

Some Annotated Biomedical Corpora (3/3)

- CAS: published case reports
  - DEFT 2019–2020 shared tasks
  - Entity mentions, types
  - Natalia Grabar (STL, CNRS, Lille), Cyril Grouin , Thierry Hamon *et al.*
- CépiDc: death certificates
  - CLEF eHealth 2016–2018 shared tasks (`https://clefehealth.imag.fr/?page_id=315`)
  - ICD-10 coding: Text classification / Concept detection (linking)
  - Aude Robert (Inserm), Grégoire Rey (Inserm), Aurélie Névéol , Cyril Grouin , Thomas Lavergne , Thierry Hamon , Pierre Zweigenbaum *et al.*

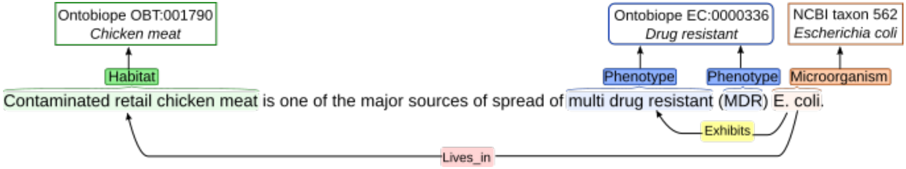
Plan

1. Introduction: LISN, TAL pour la santé
    - LISN, CNRS, Université Paris-Saclay
    - NLP for e-Health
2. TAL pour les professionnels de santé
    - Désidentification de textes cliniques
    - Normalisation de l'information pour les statistiques de santé
    - Extraction d'information dans les médias sociaux pour la pharmacovigilance
3. TAL pour les chercheurs en santé
    - Natural Language Processing for Systematic Reviews
    - Machine Translation of Biomedical Abstracts
    - Information Extraction from Biomedical Text
4. TAL pour les étudiants en médecine
    - Dialogue personne-machine pour entrainer les étudiants en médecine
5. Textes biomédicaux pour chercheurs en TAL
    - Création de corpus de textes biomédicaux
6. Plongements lexicaux en domaine spécialisé
    - Plongement d'ontologie pour le liage référentiel

Hybrid Concept and Word Embeddings for Entity Linking

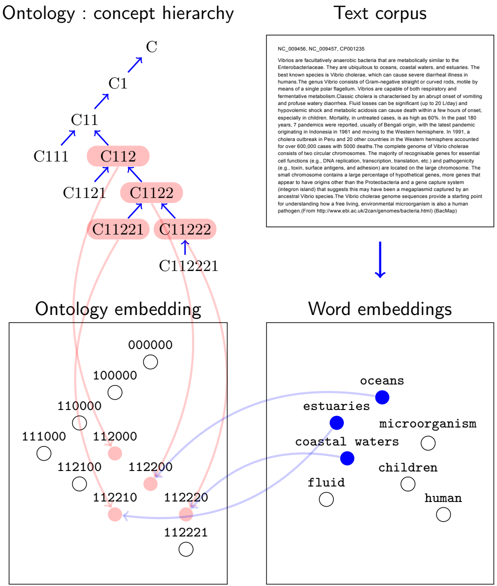
BioNLP Bacteria Biotopes shared task (Bossy et al., BioNLP OST 2019)

- c-norm
  - Co-supervised with Claire Nédellec, INRAE
  - Funding: Paris-Saclay Interdisciplinary Doctoral Initiative
-
- Entity linking (= concept normalization)
  - Link entity mention to ontology concept
- BioNLP ST Bacteria Biotope
  - Bacteria
  - Biotopes: where bacteria live (OntoBiotope)
  - Phenotypes: bacteria properties



Contaminated retail chicken meat is one of the major sources of spread of multi drug resistant (MDR) E. coli.

Mapping Word Embeddings to Concept Embeddings



NC_009456, NC_009457, CP001235

Vibrios are facultatively anaerobic bacteria that are metabolically similar to the Enterobacteriaceae. They are ubiquitous to oceans, coastal waters, and estuaries. The best known species is Vibrio cholerae, which can cause severe diarrheal illness in humans.The genus Vibrio consists of Gram-negative straight or curved rods, motile by means of a single polar flagellum. Vibrios are capable of both respiratory and fermentative metabolism.Classic cholera is characterised by an abrupt onset of vomiting and profuse watery diaorrhea. Fluid losses can be significant (up to 20 L/day) and hypovolemic shock and metabolic acidosis can cause death within a few hours of onset, especially in children. Mortality, in untreated cases, is as high as 60%. In the past 180 years, 7 pandemics were reported, usually of Bengali origin, with the latest pandemic originating in Indonesia in 1961 and moving to the Western hemisphere. In 1991, a cholera outbreak in Peru and 20 other countries in the Western hemisphere accounted for over 600,000 cases with 5000 deaths.The complete genome of Vibrio cholerae consists of two circular chromosomes. The majority of recognisable genes for essential cell functions (e.g., DNA replication, transcription, translation, etc.) and pathogenicity (e.g., toxin, surface antigens, and adhesion) are located on the large chromosome. The small chromosome contains a large percentage of hypothetical genes, more genes that appear to have origins other than the Proteobacteria and a gene capture system (integron island) that suggests this may have been a megaplasmid captured by an ancestral Vibrio species.The Vibrio cholerae genome sequences provide a starting point for understanding how a free living, environmental microorganism is also a human pathogen.(From http://www.ebi.ac.uk/2can/genomes/bacteria.html) (BacMap)

C-Norm Methods

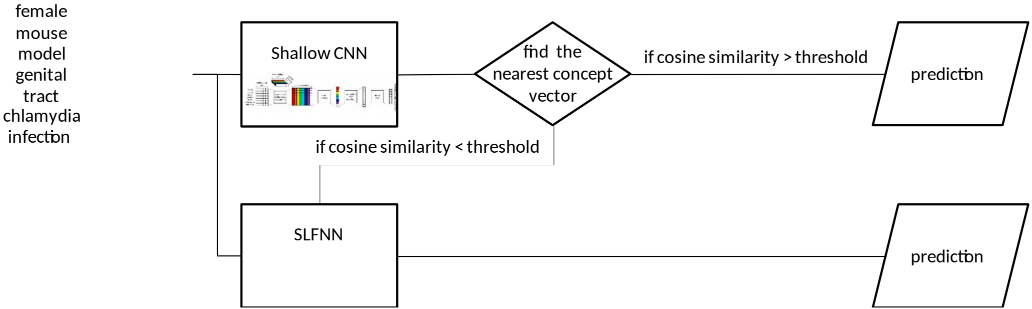
- Representation of words and concepts
  - Input: word2vec word embeddings
  - Output: ontology concept vectors
- Two entity linking methods
  - Single-layer feedforward neural network (SLFNN)
  - Shallow convolutional neural network CNN
  - Ensemble: Sieve, Average
- Ontology Embedding: Concept Vector Creation
  - One dimension per concept
  - Inherit parent vectors, with decay factor
- Training data
  - Supervision: BioNLP ST Bacteria Biotope shared task annotated corpus
  - Weak supervision: concept labels in ontology

---

[1] Arnaud Ferré, Louise Deleger, Robert Bossy, Pierre Zweigenbaum, Claire Nédellec. C-Norm: a Neural Approach to Few-Shot Entity Normalization. BMC Bioinformatics. 2020 Dec 29;21(Suppl 23):579. doi: 10.1186/s12859-020-03886-8

Sieve Combination

Sieve combination: priority to shallow CNN



female
mouse
model
genital
tract
chlamydia
infection

Shallow CNN

find the nearest concept vector

if cosine similarity > threshold

prediction

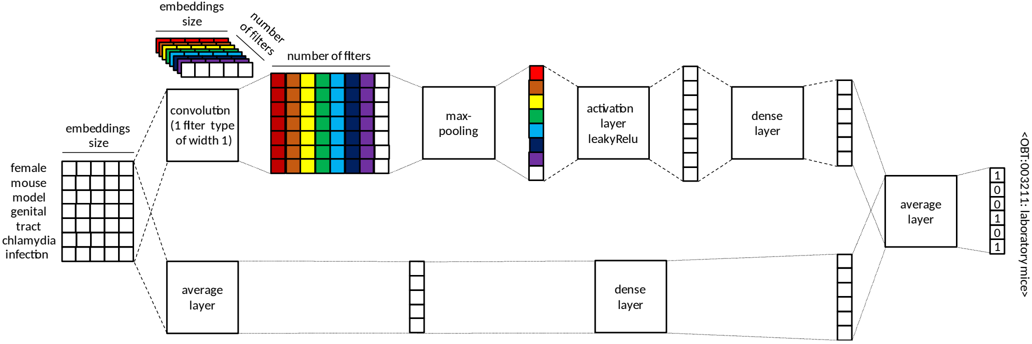
if cosine similarity < threshold

SLFNN

prediction

C-Norm Architecture

End-to-end, average combination


embeddings size
number of filters
number of filters
embeddings size
female mouse model genital tract chlamydia infection
convolution (1 filter type of width 1)
max-pooling
activation layer leakyRelu
dense layer
average layer
dense layer
average layer
1
0
0
0
1
0
1
<OBT:003211: laboratory mice>

C-Norm Results

Strong contribution of ensemble (on development set)

	BB-norm Habitats		BB-norm Phenotypes	
	Wang score	Strict score	Wang score	Strict score
SLFNN	0.654 ±0.003	0.325 ±0.004	0.814 ±0.013	0.537 ±0.011
S-CNN	0.696 ±0.003	0.510 ±0.007	0.782 ±0.005	0.501 ±0.013
Sieve (threshold=0.4)	0.725 ±0.003	0.508 ±0.005	0.807 ±0.008	0.527 ±0.009
C-Norm	**0.819** ±0.004	**0.633** ±0.009	**0.854** ±0.011	**0.620** ±0.024

C-Norm Results

Strong contribution of weak supervision (on development set, habitats)

	Wang score	Strict score
C-Norm standard	0.698 ±0.003	0.473 ±0.004
C-Norm standard+weak	**0.819** ±0.004	**0.633** ±0.009

C-Norm Results

Outperforms state of the art on test set: Bacteria Biotope participants

	Habitats		Phenotypes	
	Wang [95% CI]	Strict [95% CI]	Wang [95% CI]	Strict [95% CI]
Baseline	0.559 [0.543, 0.576]	0.224 [0.199, 0.250]	0.581 [0.559, 0.604]	0.091 [0.056, 0.127]
AmritaCen	0.522 [0.497, 0.548]	0.347 [0.314, 0.376]	0.646 [0.595, 0.698]	0.512 [0.448, 0.571]
BOUN-ISIK	0.687 [0.667, 0.710]	0.428 [0.395, 0.459]	0.566 [0.520, 0.610]	0.315 [0.259, 0.373]
BLAIR GMU	0.615 [0.596, 0.632]	0.211 [0.185, 0.237]	0.646 [0.607, 0.685]	0.313 [0.254, 0.373]
PADIA	0.684 [0.661, 0.709]	0.488 [0.456, 0.519]	0.758 [0.716, 0.803]	0.618 [0.556, 0.676]
CONTES	0.715 [0.694, 0.736]	0.500 [0.467, 0.529]	0.799 [0.760, 0.835]	0.616 [0.553, 0.675]
HONOR	0.737 [0.716, 0.759]	0.531 [0.499, 0.563]	0.778 [0.738, 0.814]	0.578 [0.519, 0.640]
C-Norm	**0.777** [0.755, 0.797]	**0.604** [0.574, 0.635]	**0.881** [0.855, 0.907]	**0.700** [0.643, 0.755]

Error Analysis on Development Set

- Incorrect: 37.4%, divided into:
  - Partially correct: 17.5%
  - Concepts too high (28.1%) or too low (15.8%) in the hierarchy
  - Other: 38.6%, including:
    - Syntactic structure of mention: 35.2%
    - Ambiguities: 18.2%

Improving Word Embeddings in Specialized Domains

- Hicham El Boukkouri : PhD thesis (2018–2021)
- Co-supervision
  - Olivier Ferret (CEA)
  - Thomas Lavergne (LISN)
  - Pierre Zweigenbaum (LISN)
- Funding: ANR project ADDICTE
- CharacterBERT: collaboration with Hiroshi Noji and Junichi Tsujii (AIRC, AIST)

Pre-training *vs* Fine-tuning for a Specialized Domain

Should one pre-train BERT from scratch for a specialized domain?

- Typical procedure: use pre-trained model, resume pre-training on in-domain corpus, then fine-tune on task
- Would it be better to pre-train on in-domain corpus directly, with in-domain vocabulary?

Domain	**Corpora**	**# of documents**	**# of words**
General	Wikipedia (EN)	11,9 million	2,14 billion
	OpenWebText	3,15 million	1,28 billion
Medical	MIMIC-III	4,17 million	0,5 billion
	PubMed	4,65 million	0,5 billion

---

[1] Hicham El Boukkouri. Ré-entraîner ou entraîner soi-même ? Stratégies de pré-entraînement de BERT en domaine médical. In: RECITAL, Nancy, France, 2020:29–42. `https://www.aclweb.org/anthology/2020.jeptalnrecital-recital.3/`

Pre-training *vs* Fine-tuning for a Specialized Domain

Pre-training on medical corpus better; But if retraining on medical corpus, not much different. Better than BERT and BlueBERT

Model			Evaluation task			
V	**C_1**	**C_2**	**i2b2/VA 2010**	**MEDNLI**	**ChemProt**	**DDI**
general	general	∅	85,66 ± 0,18	77,31 ± 0,71	67,47 ± 0,99	75,81 ± 1,02
general	general	medical	89,00 ± 0,17	**84,91** ± 0,46	72,29 ± 0,58	78,82 ± 1,11
medical	medical	∅	88,80 ± 0,10	83,54 ± 0,43	71,30 ± 0,51	79,40 ± 1,15
medical	medical	medical	**89,20** ± 0,20	84,32 ± 0,73	**72,97** ± 0,46	**80,11** ± 0,79
BERT (base)			86,42 ± 0,31	77,85 ± 0,63	69,22 ± 0,56	77,89 ± 0,92
BlueBERT (base)			88,70 ± 0,21	84,53 ± 0,76	68,35 ± 0,61	77,89 ± 0,65

WordPieces and Specialized Domains

- BERT and Transformer-based state of the art
  - use WordPiece tokenization
- Specialized models (e.g., BioBERT, BlueBERT)
  - resume pre-training from general domain
  - keep general-domain WordPiece vocabulary

---

[1] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, Junichi Tsujii. CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters. COLING 2020

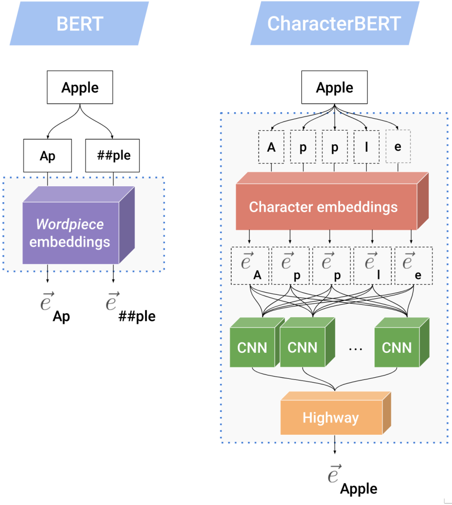
WordPieces: Related Issues

- WordPieces can be inconvenient
  - multiple vectors for unknown tokens
  - word similarity tasks: aggregation issue
  - sequence labeling tasks: only label first wordpiece
- General WordPieces not suited for specialized domains
  - [para, ce, tam, ol]
  - [cho, led, och, oli, thi, asi, s]
  - [bo, rb, ory, gm, i]

CharacterBERT: a Word-based BERT

Character-CNN module like ELMO

- No wordpieces
- Character-level
- 1 token = 1 embedding
- Open-vocabulary
- Whole-word masking for free



CharacterBERT: Pre-training Corpora

General and Medical corpora

Corpus	Composition	# documents	# tokens
General	Wikipedia (EN)	5.99×10^6	2.14×10^9
	OpenWebText	1.56×10^6	1.28×10^9
Medical	MIMIC-III	2.09×10^6	0.51×10^9
	PMC OA abstracts	2.33×10^6	0.52×10^9

General : train on general corpus

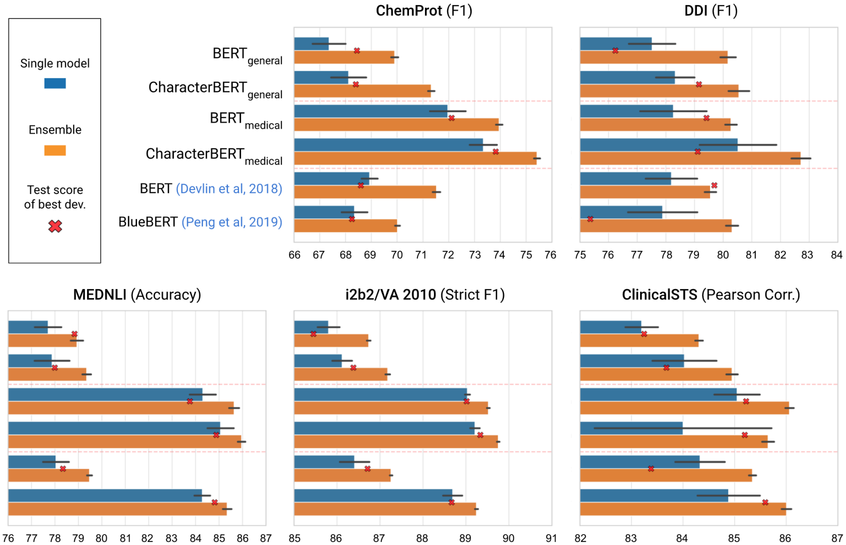
Medical : re-train general on medical corpus

CharacterBERT: Evaluation

- Five tasks
  - Entity detection: i2b2/VA 2010
  - Relation Extraction: ChemProt and DDI
  - Natural Language Inference: MEDNLI
  - Text similarity: ClinicalSTS
- 10 random seeds / 15 training epochs / keep best on validation set
  - Final performance: avg+/– std on test set
  - Ensemble performance: majority vote; avg. score for STS

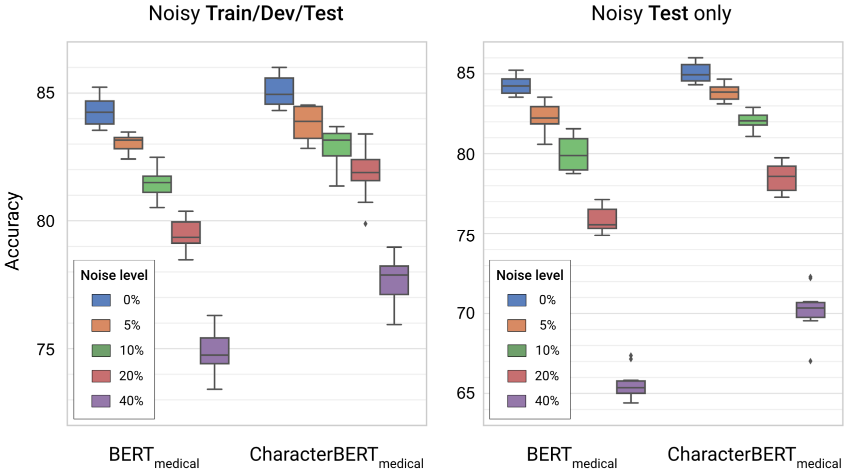
CharacterBERT: Results

Overall better performance from CharacterBERT



CharacterBERT: Robustness to Noise

Improved robustness to noisy versions of MEDNLI test



CharacterBERT: Conclusion

- Drops wordpiece system altogether
- Produces open-vocabulary representations by consulting a token's characters
- Globally outperforms BERT when specialized for medical domain
- More robust to misspellings

Plan

1. Introduction: LISN, TAL pour la santé
   - LISN, CNRS, Université Paris-Saclay
   - NLP for e-Health
2. TAL pour les professionnels de santé
   - Désidentification de textes cliniques
   - Normalisation de l'information pour les statistiques de santé
   - Extraction d'information dans les médias sociaux pour la pharmacovigilance
3. TAL pour les chercheurs en santé
   - Natural Language Processing for Systematic Reviews
   - Machine Translation of Biomedical Abstracts
   - Information Extraction from Biomedical Text
4. TAL pour les étudiants en médecine
   - Dialogue personne-machine pour entrainer les étudiants en médecine
5. Textes biomédicaux pour chercheurs en TAL
   - Création de corpus de textes biomédicaux
6. Plongements lexicaux en domaine spécialisé
   - Plongement d'ontologie pour le liage référentiel

Conclusion: Traitement automatique des langues pour la santé

- Extraction d'information
  - dans des textes cliniques
  - dans les médias sociaux
  - dans les publications scientifiques
- Dialogue en langue naturelle
  - avatar dialogant pour la formation des étudiants en médecine
- Liage à des sources de connaissances pré-existantes
  - dans les certificats de décès pour les statistiques (inter)nationales
  - dans les certificats de décès pour la surveillance syndromique réactive
- Traduction automatique
  - pour faciliter l'accès à des publications dans des langues diverses
  - ... simplification de texte
- Méthodes hybrides qui prennent en compte les connaissances existantes
  - plongement d'ontologie pour le liage référentiel
  - dictionnaires etc. dans les méthodes supervisées
- Plongements de mots et de concepts spécialisés

Merci !

Campillos Llanos, L., L. Deléger, C. Grouin, T. Hamon, A.-L. Ligozat, and A. Névéol. "A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)". Language Resources and Evaluation (2017), pp. 1–31.

Campillos Llanos, L., C. Thomas, É. Bilinski, P. Zweigenbaum, and S. Rosset (2019) "Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation". Natural Language Engineering, 2019, pp. 1-38.

Ferré, A., P. Zweigenbaum, and C. Nédellec. "Representation of complex terms in a vector space structured by an ontology for a normalization task". In: BioNLP workshop, ACL. Vancouver CA, 2017, pp. 99-106.

Grabar, N. and T. Hamon. "Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux". TAL 57.1 (2016), pp. 85–109.

Grouin, C., and P. Zweigenbaum, "Automatic De-Identification of French clinical records: comparison of rule-based and machine-learning approaches". In: Proc MEDINFO 2013, Studies in Health Technology and Informatics. Amsterdam, IOS Press, 2013, pp. 476–480.

Hamon, T. and N. Grabar. "Adaptation of cross-lingual transfer methods for the building of medical terminology in Ukrainian". In: CICLING, LNCS. Springer, 2016.

Hamon, T., T. Vincent, F. Mougin, N. Grabar, and F. Thiessard. "POMELO: Medline corpus with manually annotated food-drug interactions". In: Proceedings of the Biomedical NLP Workshop, Recent Advances in Natural Language Processing (RANLP). Varna BG, 2017, pp. 73–80.

Ive, J., A. Max, F. Yvon, and P. Ravaud. "Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Edition Operations". In: LREC Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem. Portorož SI, 2016, 8p.

Kamath, S., B. Grau, and Y. Ma. "Verification of the expected answer type for biomedical question answering". In: First International Workshop on Hybrid Question Answering with Structured and Unstructured Knowledge, The Web Conference. Lyon FR, 2018, pp. 1093–1097.

Koroleva, A. and P. Paroubek. "On the contribution of specific entity detection and comparative construction to automatic spin detection in biomedical scientific publications". In: Workshop on Processing Emotions, Decisions and Opinions. Poznan PL, 2017, 5p.

Lavergne, T., A. Névéol, A. Robert, C. Grouin, G. Rey, and P. Zweigenbaum. "A dataset for ICD-10 coding of death certificates: Creation and usage". In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016, COLING). Osaka, JP, 2016, pp. 60–69.

Morlane-Hondère, F., C. Grouin, and P. Zweigenbaum. "Identification of Drug-Related Medical Conditions in Social Media". In: LREC. Portorož SI, 2016, pp. 2022–2028.

Neifar, W., T. Hamon, P. Zweigenbaum, M. Ellouze Khemakhem, and L. Hadrich Belguith. "Adaptation of a Term Extractor to Arabic Specialised Texts: First Experiments and Limits". In: CICLING, LNCS. Springer, 2016, 12p.

Norman, C., M. Leeflang, P. Zweigenbaum, and A. Névéol. "Automating Document Discovery in the Systematic Review Process: How to Use Chaff to Extract Wheat". In: LREC. Miyazaki JP, 2018, 7p.

Paris, N., M. Mendis, C. Daniel, S. Murphy, X. Tannier, and P. Zweigenbaum. "i2b2 implemented over SMART-on-FHIR". In: AMIA Jt Summits Transl Sci Proc. San Francisco US, 2018, pp. 369–378.

Tourille, J., O. Ferret, X. Tannier, and A. Névéol. "Temporal information extraction from clinical text". In: Conference of the European Chapter of the Association for Computational Linguistics. Valencia ES, 2017, pp. 739–745.