Traitement Automatique de la Langue et Intégration de Données pour les Réunions de Concertations Pluridisciplinaires en Oncologie

Nesrine Bannour[1], Aurélie Névéol[1], Xavier Tannier[2], Bastien Rance[3]

(1) Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)

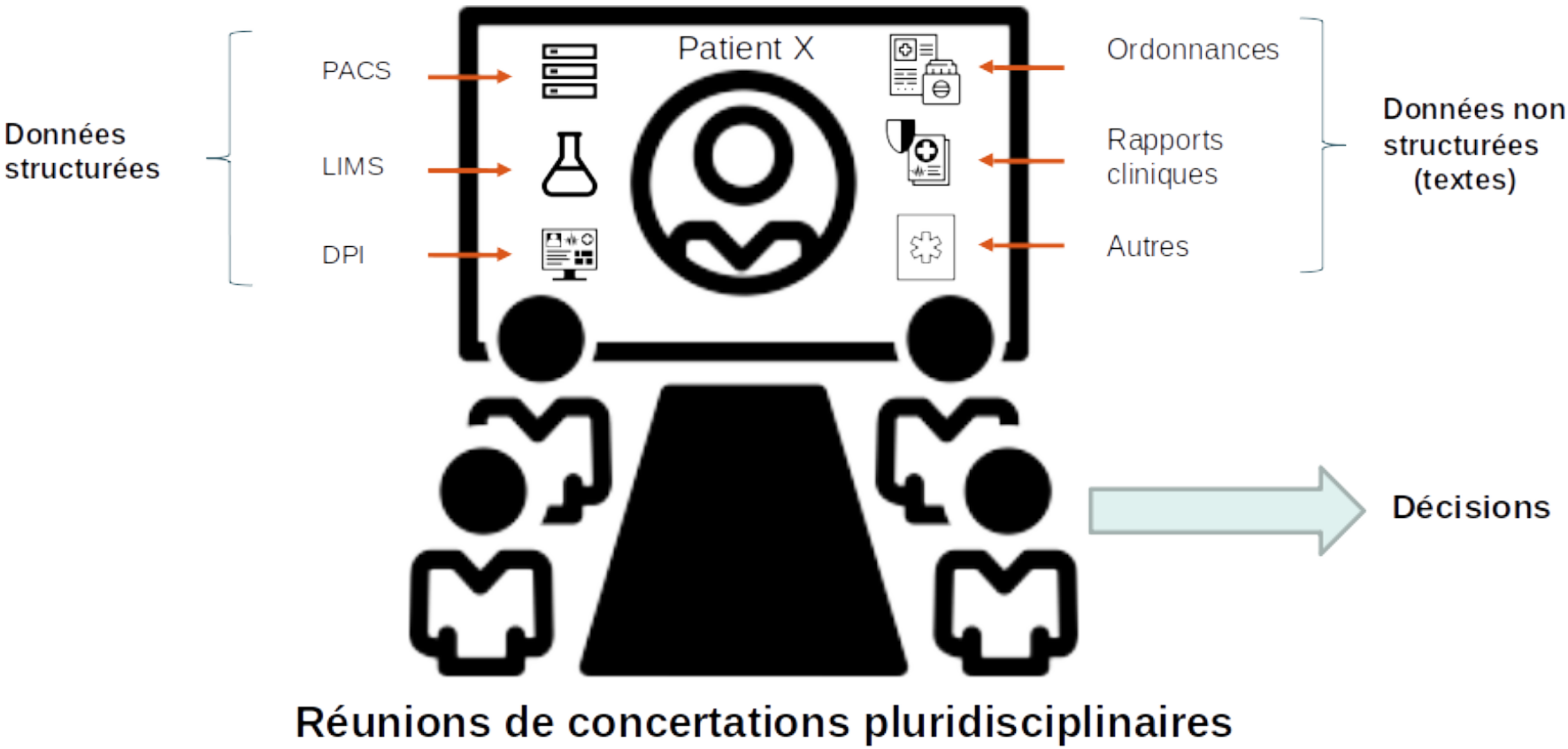
(2) Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS)

(3) Hôpital Européen Georges Pompidou (HEGP), AP-HP

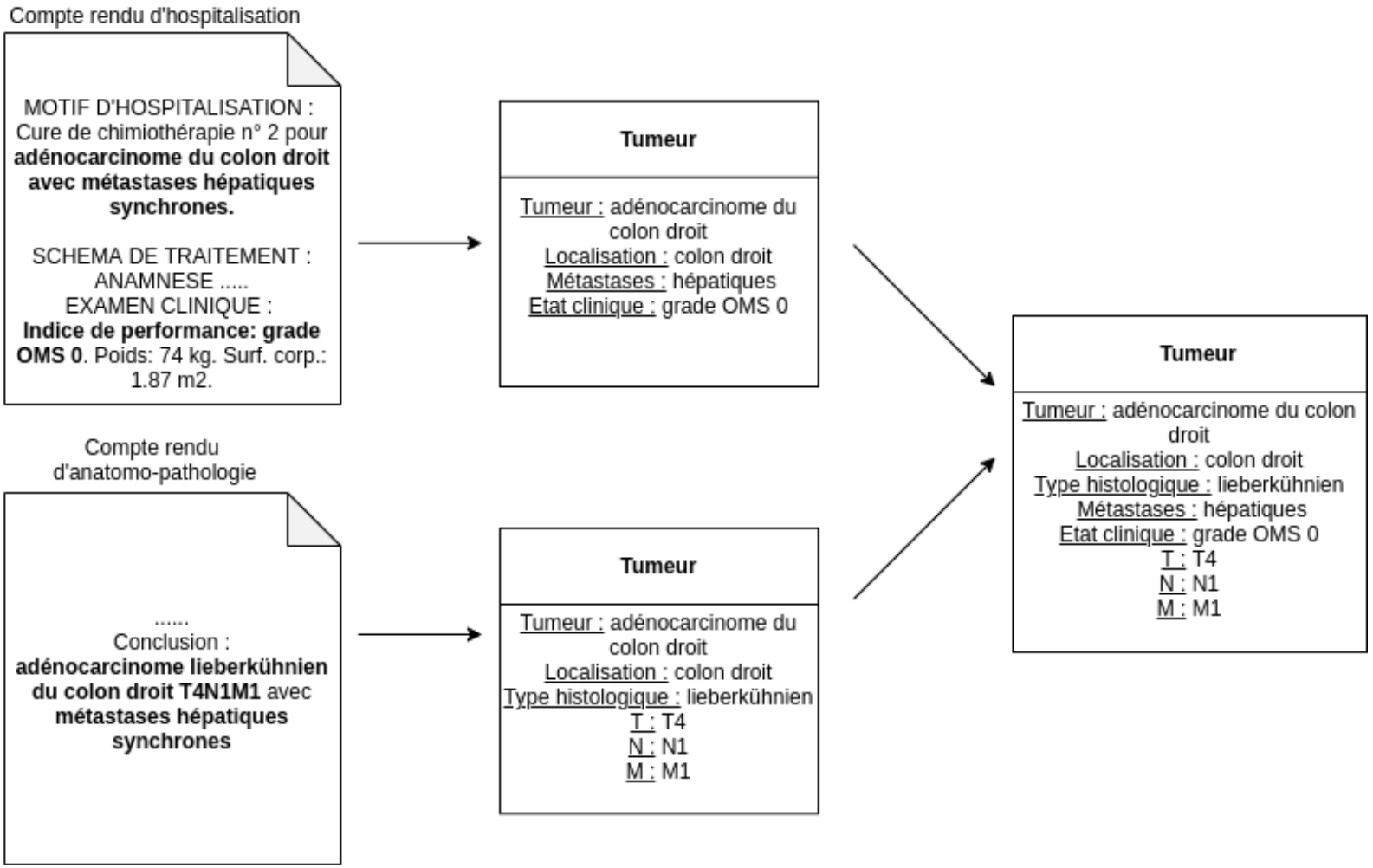
TLH Santé 2021 - 4 Février 2021

Plan de la présentation

Projet TALONCO - ASIMOV (1/2)



Projet TALONCO - ASIMOV (2/2)



Classification de documents → Extraction d'entités → Résolution d'ambiguïté

Objectifs

- Extraire automatiquement les informations pertinentes dans les entrepôts de données et les textes cliniques des dossiers médicaux.
- Intégrer les informations structurées et non structurées extraites dans une plateforme d'intégration de données sur le cancer.
- Réduire le degré de supervision et assurer l'adaptabilité à d'autres domaines dans les méthodes à utiliser.

Approches existantes d'extraction d'informations cliniques

1. Approches à base de règles [Childs et al., 2009], [Deléger et al., 2010], [Khalifa et al., 2016]

2. Approches statistiques
   - Apprentissage automatique [Sarker and Gonzalez, 2015], [Jiang et al., 2015], [Henriksson et al., 2017]
   - Apprentissage profond [Li and Huang, 2016], [Liu et al., 2017]

3. Approches hybrides [Tang et al., 2013], [Castro et al., 2017], [Lerner et al., 2020]

4. **En oncologie :**
   - Localisation de l'événement tumoral [Savova et al., 2017]
   - Classification histologique [Nguyen et al., 2017]
   - Classification TNM [Nguyen et al., 2010]
   - Stade du cancer [Gupta et al., 2019]

Ressources médicales existantes

Corpus	Documents cliniques #	Entités uniques #
LERUDI	138,000	-
MERLOT[1] (annoté)	500	13,830
Quaero [2] (annoté) - EMEA	38	1,880
Quaero (annoté) - MEDLINE	2,498	5,895
DEFT [3] (annoté)	167	8,725

- Nos méthodes seront évalués par la suite sur des dossiers patients suivis pour un cancer du poumon, du colon ou un cancer du rein métastatique de l'EDS de l'AP-HP.

[1] [Campillos et al., 2017]
[2] [Névéol et al., 2014]
[3] https://deft.limsi.fr/2020/

Création d'un corpus spécifique aux informations cancer (1/2)

- **Idée :** Augmenter le corpus annoté MERLOT avec d'autres documents non annotés du corpus LERUDI contenant des informations cancer
  → Classification de documents : Régression logistique

- **Données :** Récupération de comptes rendus d'hospitalisation et de RCP du corpus MERLOT en les annotant en exemples positifs et négatifs.

Données	Exemples positifs	Exemples négatifs
Entraînement_MERLOT	60	237
Test_MERLOT	11	51

Création d'un corpus spécifique aux informations cancer (2/2)

- Récupérer les exemples du LERUDI les mieux classés en tant que positifs et négatifs par le classifieur après la première itération, corriger l'annotation et les ajouter aux données d'entraînement et de test de MERLOT

  → Quatre itérations effectuées

Données de test	Précision	Rappel	F_mesure
Test_augmenté	0.982	0.849	0.936

Performance du modèle lors de la quatrième itération

Données	Exemples positifs	Exemples négatifs
Entraînement_augmenté	298	427
Test_augmenté	179	107

Description des données augmentées

Système de reconnaissance d'entités nommées (1/2)

- Transformer les données de MERLOT annotées en brat vers le format CoNLL.
- Utilisation des transformers pré-entrainés mis à disposition par Hugging Face [Wolf et al., 2019], en particulier le modèle CamemBERT [Martin et al., 2019].

Données	Nombre de documents
Entraînement_MERLOT	320
Validation_MERLOT	80
Test_MERLOT_pertinent	17

Description des données utilisées

Système de reconnaissance d'entités nommées (2/2)

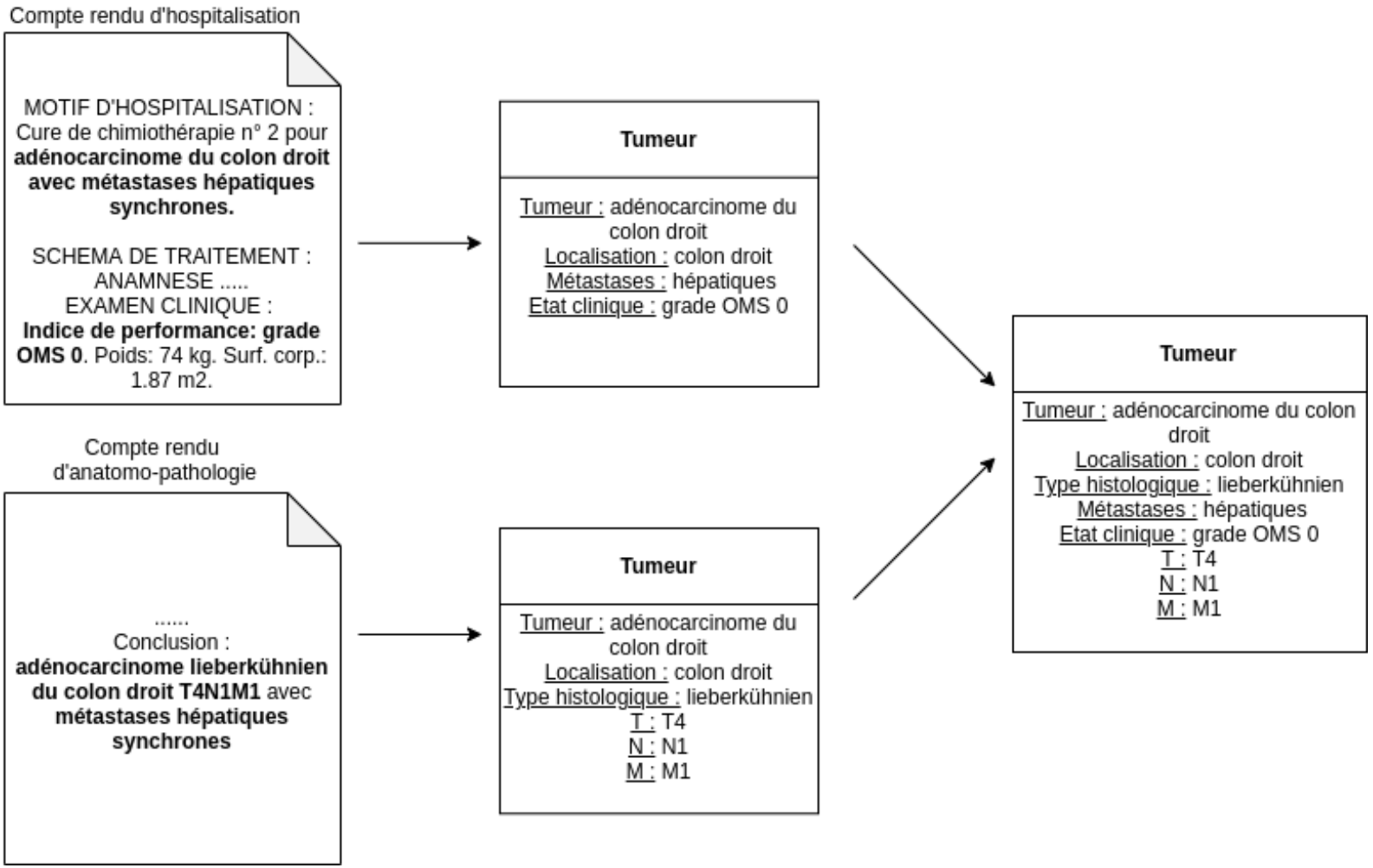
Données de test	Précision	Rappel	F_mesure
Test_MERLOT_pertinent	0.8145	0.8356	0.8249

Résultats de NER sur les données pertinentes de test de MERLOT

Entité	Précision	Rappel	F_mesure	Support
Anatomy	0.69	0.77	0.73	56
Disorder	0.76	0.79	0.77	178
Localization	0.00	0.00	0.00	5
Measurement	0.73	0.81	0.77	229
Temporal	0.91	0.93	0.92	240

Résultats de NER sur les données pertinentes de test de MERLOT pour certaines entités

Extraction d'informations



Classification de documents → Extraction d'entités → Résolution d'ambiguïté

Intégration de données et un degré minimal de supervision

- Intégrer les informations extraites des textes cliniques avec les informations extraites des entrepôts de données dans CARPEM [Rance et al., 2016]

- Définir des méthodes d'adaptation du modèle générique à un domaine particulier en exploitant peu d'exemples annotés, par exemple par supervision distante [Fries et al., 2017], par apprentissage actif [Settles et al., 2008] ou en exploitant des techniques d'apprentissage par transfert [Papernot et al., 2017]

Conclusion et travaux à venir (1/2)

Travaux réalisés

- Mise en place d'un corpus spécifique aux informations cancer en augmentant le corpus MERLOT,
- Extraction d'entités cliniques à partir d'un corpus annoté en utilisant un système de reconnaissance d'entités nommées.

Conclusion et travaux à venir (2/2)

Travail futur

- Utilisation des ressources médicales annotées (Quaero, DEFT) pour entraîner l'outil d'extraction d'informations,
- Appliquer cet outil d'extraction d'informations sur les données de l'EDS de l'AP-HP,
- Extraction d'informations temporelles pour la construction des chronologies médicales de patients.

Merci!

Merci de votre attention !

Campillos, L., Deléger, L., Grouin, C., Hamon, T., Ligozat, A.-L., and Névéol, A. (2017).
A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT).
Language Resources and Evaluation, 52(2):571–601.

Castro, S. M., Tseytlin, E., Medvedeva, O., Mitchell, K. J., Visweswaran, S., Bekhuis, T., and Jacobson, R. S. (2017).
Automated annotation and classification of bi-rads assessment from radiology reports.
Journal of biomedical informatics, 69:177–187.

Childs, L. C., Enelow, R., Simonsen, L., Heintzelman, N. H., Kowalski, K. M., and Taylor, R. J. (2009).
Description of a Rule-based System for the i2b2 Challenge in Natural Language Processing for Clinical Data.

Journal of the American Medical Informatics Association, 16(4):571–575.


Deléger, L., Grouin, C., and Zweigenbaum, P. (2010). Extracting medication information from french clinical texts. *Studies in health technology and informatics*, 160:949–53.


Fries, J. A., Wu, S., Ratner, A., and Ré, C. (2017). Swellshark: A generative model for biomedical named entity recognition without labeled data. *CoRR*, abs/1704.06360.


Gupta, K., Thammasudjarit, R., and Thakkinstian, A. (2019). NLP automation to read radiological reports to detect the stage of cancer among lung cancer patients. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 138–141, Florence, Italy. Association for Computational Linguistics.


Henriksson, A., Kvist, M., and Dalianis, H. (2017).

Detecting protected health information in heterogeneous clinical notes.
Studies in health technology and informatics, 245:393–397.


Jiang, J., Guan, Y., and Zhao, C. (2015).
Wi-enre in clef ehealth evaluation lab 2015: Clinical named entity recognition based on crf.
In *CLEF*.


Khalifa, A., Velupillai, S., and Meystre, S. (2016).
UtahBMI at SemEval-2016 task 12: Extracting temporal information from clinical text.
In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1256–1262, San Diego, California. Association for Computational Linguistics.


Lerner, I., Paris, N., and Tannier, X. (2020).
Terminologies augmented recurrent neural network model for clinical named entity recognition.

Journal of Biomedical Informatics, 102:103356.

Li, P. and Huang, H. (2016).
UTA DLNLP at SemEval-2016 task 12: Deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports.
In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1268–1273, San Diego, California. Association for Computational Linguistics.

Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., and Xu, H. (2017).
Entity recognition from clinical texts via recurrent neural network.
BMC Medical Informatics and Decision Making, 17.

Martin, L., Müller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2019).
Camembert: a tasty french language model.

CoRR, abs/1911.03894.


Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014).
The quaero french medical corpus : A ressource for medical entity recognition and normalization.


Nguyen, A., Moore, J., O'Dwyer, J., and Philpot, S. (2017).
Automated cancer registry notifications: Validation of a medical text analytics system for identifying patients with cancer from a state-wide pathology repository.
AMIA Annual Symposium Proceedings, 2016:964–973.


Nguyen, A. N., Lawley, M. J., Hansen, D. P., Bowman, R. V., Clarke, B. E., Duhig, E. E., and Colquist, S. (2010).
Symbolic rule-based classification of lung cancer stages from free-text pathology reports.
Journal of the American Medical Informatics Association, 17(4):440–445.


Papernot, N., Abadi, M., Úlfar Erlingsson, Goodfellow, I., and Talwar, K. (2017).
Semi-supervised knowledge transfer for deep learning from private training data.


Rance, B., Canuel, V., Countouris, H., Laurent-Puig, P., and Burgun, A. (2016).
Integrating heterogeneous biomedical data for cancer research: the carpem infrastructure.
Appl Clin Inform, 7:260–74.


Sarker, A. and Gonzalez, G. (2015).
Portable automatic text classification for adverse drug reaction detection via multi-corpus training.
Journal of Biomedical Informatics, 53:196 – 207.


Savova, G. K., Tseytlin, E., Finan, S., Castine, M., Miller, T., Medvedeva, O., Harris, D., Hochheiser, H., Lin, C., Chavan, G., et al. (2017).

Deepphe: a natural language processing system for extracting cancer phenotypes from clinical records.
Cancer research, 77(21):e115–e118.


Settles, B., Craven, M., and Friedland, L. (2008).
Active learning with real annotation costs.


Tang, B., Wu, Y., Jiang, M., Chen, Y., Denny, J., and Xu, H. (2013).
A hybrid system for temporal information extraction from clinical text.
Journal of the American Medical Informatics Association : JAMIA, 20.


Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019).
Huggingface's transformers: State-of-the-art natural language processing.

CoRR, abs/1910.03771.