Combining String-based and Embeddings-based Methods for Medical Concept Normalization: LIMSI-CEA-INRA@n2c2 2019

Mohamadou Ba, Robert Bossy, Pauline Brunet, Louise Deleger, Hicham El Boukkouri, Olivier Ferret, Arnaud Ferré, Thomas Lavergne, Claire Nédellec, Pierre Zweigenbaum

Combining String-based and Embeddings-based Methods for Medical Concept Normalization: LIMSI-CEA-INRA@n2c2 2019

Mohamadou Ba,[1] Robert Bossy,[1] Pauline Brunet,[2] Louise Deléger,[1]
Hicham El Boukkouri,[3,4] Olivier Ferret,[2] Arnaud Ferré,[1,3,4]
Thomas Lavergne,[3,4] Claire Nédellec,[1]
Pierre Zweigenbaum[3]

[1]MaIAGE, INRA, Jouy-en-Jausas, France
[2]CEA, LIST, Gif-sur-Yvette, France
[3]LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[4]Université Paris-Sud, Université Paris-Saclay, Orsay, France

1 Methods

We explored methods based upon:

1. String representations of concepts, with incrementally more aggressive string normalizations (lower-casing, stop word removal, etc.). These methods match input mentions to entries (submitted to the same string normalizations) in a string-to-concept dictionary obtained:

   - from the training examples: a "rote learner" returns the most frequent concept seen for a given mention in the training corpus;
   - from the UMLS 2017AB Metathesaurus terms (including synonyms) for all Concepts Unique Identifiers (CUIs) in SNOMED CT and RxNorm, preprocessed and filtered with JuFiT [1] (*Run2* and *Run3*) to remove metalanguage.

   These methods aim at handling the easiest cases with maximal precision. If no concept is found for a mention at a given step, control is passed to the next, possibly less precise, step.

2. MetaMap [2] (*Run1* only), which provides more elaborate matching, including linguistic transformations, against UMLS Metathesaurus terms.

3. Word embeddings of mentions and concept terms (*Run2* and *Run3*), obtained by fastText [3] trained on MIMIC-III [4]. Mention and concept term embeddings are computed using *smooth*

inverse frequency (SIF) [5] which applies an inverse-frequency weighting before removing the first principal component of the term representations. The k nearest neighbor concept terms are retrieved for each mention, using the Faiss library [6] with HNSW indexing [7] for efficiency.

Fusion: We combined these methods in a sieve-style pipeline (Fig. 1). Processing stops as soon as one of the string-based methods returns exactly one concept, as in [8]. If multiple concepts are in a tie, the embeddings-based nearest neighbor among these concepts is selected (*Run2* only). If the string-based methods return no concept, the embeddings-based nearest neighbor is returned (*Run2* and *Run3* only).

2 Results and Discussion

We submitted three runs (Table 1):

- *Run1* implements a more sophisticated dictionary-based approach in which preprocessing and filtering of UMLS terms are more accurate, MetaMap is used, but word embeddings are not used. This gives it a very good precision (P=90.74) and a reasonable recall (R=75.52).
- *Run2* and *Run3* implement fusion with the embeddings method, and obtain higher recall and accuracy but lower precision. *Run2* does not perform disambiguation when several CUIs are selected for a mention while *Run3* does. This only increases scores by 0.1pt.

Run	Features	Acc	Pre	Rec	F1
Run1	RM	75.78	**90.74**	75.52	**82.43**
Run2	RUE	**78.43**	78.18	**79.65**	78.91
Run3	RUE	**78.54**	78.30	**79.77**	79.03

Table 1: Detailed scores on test set. R=Rote, M=MetaMap, U=UMLS, E=Embeddings. To compute precision and recall, we only consider non CUI-less mentions, for both gold standard and system.

Table 2 shows how accuracy increases as more steps are cumulatively tried in *Run3*. The embeddings-based method alone obtains 60.75 accuracy but adds 6pt to the previous step. Similarly, combining it to *Run1* adds 4.4pt (post-submission test: Acc=**80.17**). This shows a clear complementarity of the

Rote	UMLS lower case	UMLS . . . stemming	Embeddings
53.60	67.81	72.56	78.54

Table 2: Cumulative accuracy of best run (*Run3*) as more steps are tried

string-based and embeddings-based methods that we might better exploit with a more sophisticated fusion strategy (e.g., voting or stacking).

We also experimented with, but did not have the time to finalize, contextual embeddings [9] and ontology embeddings [10]. We plan to include them before the shared task workshop.

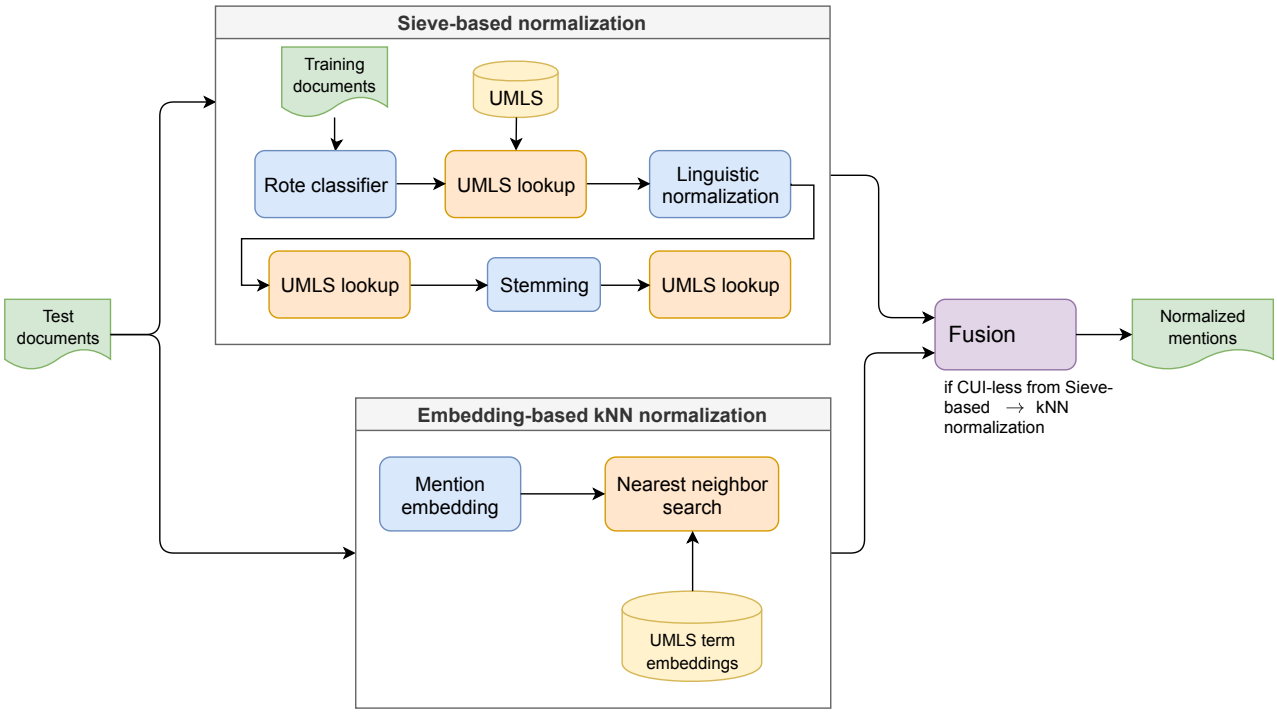

Figure 1: Global architecture of the LIMSI-CEA-INRA@n2c2 system (*Run2* and *Run3*)

References

[1] Johannes Hellrich, Stefan Schulz, Sven Buechel, and Udo Hahn. JuFiT: A configurable rule engine for filtering and generating new multilingual UMLS terms. In *AMIA Annual Symposium*, pages 604–610. American Medical Informatics Association, November 2015.

[2] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association (JAMIA)*, 17(3):229–236, 2010.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[4] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, May 2016.

[5] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations (ICLR 2017), poster session*, Toulon, France, 2017.

[6] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*, 2017.

[7] Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[8] Arnaud Ferré, Louise Deléger, Pierre Zweigenbaum, and Claire Nédellec. Combining rule-based and embedding-based approaches to normalize textual entities with an ontology. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3443–3447, Miyazaki, Japan, May 7-12 2018. European Language Resources Association (ELRA).

[9] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. Embedding strategies for specialized domains: Application to clinical entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Student Research Workshop)*, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Arnaud Ferré, Pierre Zweigenbaum, and Claire Nédellec. Representation of complex terms in a vector space structured by an ontology for a normalization task. In *BioNLP 2017*, pages 99–106, Vancouver, Canada,, August 2017. Association for Computational Linguistics.