Language Transfer for Identifying Diagnostic Paragraphs in Clinical Notes

Luca Di Liello, Olga Uryupina and Alessandro Moschitti
University of Trento, Italy
{luca.diliello,moschitti}@unitn.it, uryupina@gmail.com

Abstract

English. This paper aims at uncovering the structure of clinical documents, in particular, identifying paragraphs describing "diagnosis" or "procedures". We present transformer-based architectures for approaching this task in a monolingual setting (English), exploring a weak supervision scheme. We further extend our contribution to a cross-lingual scenario, mitigating the need for expensive manual data annotation and taxonomy engineering for Italian.

Italian. *In questo lavoro abbiamo studiato approfonditamente la struttura dei documenti clinici ed, in particolare, abbiamo creato sistemi automatici per l'estrazione di paragrafi contenenti diagnosi e procedure. Attraverso l'utilizzo di modelli basati sull'architettura transformer, abbiamo estratto diagnosi e procedure nel setting monolingua (in inglese). Successivamente, abbiamo esteso la nostra ricerca allo scenario multilingue, riducendo il fabbisogno di larghi dataset in italiano annotati manualmente grazie all'utilizzo di machine translation e transfer learning.*

1 Introduction

Big Data approaches have been shown to yield a breakthrough to a variety of healthcare-related tasks, ranging from eHealth governance and policy making to precision medicine and smart solutions/suites for hospitals or individual doctors. They rely on large-scale and reliable automatic processing of vast amounts of heterogeneous data, i.e., images, lab reports and, most importantly, textual medical documentation.

The current paper focuses on *Medical Discourse Analysis*: imposing structure on digitalized health reports through document segmentation and labeling of relevant segments (e.g., `diagnoses`). Identifying and interpreting discourse fragments is essential for accurate and robust Information Extraction from medical documents. In terms of doctor assistance, such a system could quickly and reliably identify the most crucial parts of voluminous health records, allowing to highlight them for improved visibility and thus reducing cognitive load on doctors. For example, a highlighted problematic diagnosis can alert a doctor perusing a large medical dossier. In terms of automated data analytics, discourse structure is crucial for correct interpretation of extracted information. For example, if we want to study a possible correlation between the use of a specific medicine and some outcome, we should only consider documents where this medicine is mentioned as a part of `therapy`, but not as a part of `allergies`.

Some medical documents are generated using task-specific eHealth software imposing certain discourse structure. In Italy, however, there is no single software adopted at either national or regional levels. While there is a general agreement on the nature of information to be included, there are no guidelines or programmatic implementations for structuring it. In addition, historical records, produced before the adoption of recording software, follow the logic of individual doctors and thus show even more variability. We aim therefore at a statistical model that is able to infer the discourse structure without making any assumptions on the recording software.

An important advantage of our approach is its adaptability to new domains (e.g., radiology reports) or languages as well as its robustness in the (highly probable) scenario where new reportgenerating systems appear at the market.

Several recent studies (Sec. 2) focus on segment labeling for medical records in English. To our knowledge, no approach has been proposed so far to analyze medical discourse structure automatically in other languages, including, most importantly, Italian. The required research is hampered by the lack of resources in other languages, ranging from no data annotated for discourse structure, either for training or for benchmarking, to lack of high-coverage resources, e.g., taxonomies. In our study, we propose a language transfer approach to the problem of medical discourse analysis in Italian. We first investigate possibilities for training robust monolingual models (Sec. 4) and then build upon our monolingual results to transfer the model in another language (Sec. 5).

2 State of the Art

In the past decade, a massive effort has been invested into analyzing automatically textual medical data (clinical notes). The notes' internal logic is crucial for interpreting their underlying semantics, thus enabling better understanding and interoperability. This has given rise to empirical studies on the medical document structure: reliable and interpretable annotation guidelines and systems for automatically segmenting clinical notes and annotating segments with labels such as `allergy` or `diagnosis`.

The most thorough attempt at defining clinical records' structure via a taxonomy of *section headers* has been undertaken by Denny et al. (2008). This study developed SecTag—a hierarchical header terminology, supporting mappings to LOINC and other taxonomies. Table 1 shows some SecTag entries related to `diagnosis` and their parameters relevant for the present study.[1] The SecTag concepts (column 1) are organized hierarchically, with specific diagnoses (e.g., admission or discharge diagnoses) being subnodes (column 2) of the main `diagnosis` concept (SecTag node "5.22"). Different ways of expressing the specific semantics via headers (column 3) are then linked to the corresponding nodes. SecTag advocates a practical data-driven approach, thus listing headers that are not always grammatical (e.g., "admit diagnosis"), provided they are commonly used by practicing clinicians. Most importantly, SecTag goes beyond a superficial view of the task, not only linking easily identifiable headers, (e.g., most common spellings, headers containing important key words), but also organising hierarchically concepts that are normally expressed in very distinct ways (e.g., linking "cause of death" or "gaf" to diagnoses). In total, SecTag provides 94 entries just for `diagnosis`. This shows that a considerable medical expertise is required for creating a similar resource for other languages from scratch.

The SecTag release has led to the development of a related method for automatic identification of sections in clinical notes (Denny et al., 2009), via a combination of NLP techniques, terminology-based rules, and naive Bayes classification.

While the SecTag approach exhibits remarkable performance, creation and maintenance of the header taxonomy is a very expensive task requiring considerable medical expertise. More data-driven approaches have been proposed recently for English (Rosenthal et al., 2019; Dai et al., 2015), among others. These systems, however, require manually labeled data.

3 Data for Identifying Diagnoses and Procedures Segments

3.1 English Data: MIMIC-III

Several large collections of medical data, with partial NLP annotations, have been released recently, for example, MIMIC (Johnson et al., 2016) or I2B2[2]. Unfortunately, none of these resources provide annotation for discourse structure. Our study relies on the MIMIC-III dataset, extending it with an extra layer to label diagnosis and procedure fragments. Our choice follows practical motivations: it is the largest available dataset, most commonly used by the AI community. We only rely on the textual data from MIMIC discharge notes (the NOTESEVENTS table), however, a future work can explore possibilities of joint modeling of textual and numeric data (e.g., lab measurements).

We have built a rule-based algorithm for annotating MIMIC with diagnosis/procedure fragments. We segment a note into fragments and label them based on the headers, looking them up in SecTag (Section 2). For fragments with no header, we propagate the label from the previous fragment. Fragments with headers not

[1] SecTag entires contain 16 parameters, inheriting information from referenced taxonomies such as LOINC, most of them are of no practical relevance in our case and, moreover, are typically set to NULL.

[2] `https://www.i2b2.org/`

concept	**taxonomy tree id**	**header**
diagnoses	5.22	diagnosis
principle_diagnosis	5.22.39	primary diagnoses
diagnosis_at_death	5.22.41	cause of death
admission_diagnosis	5.22.44	admit diagnosis
discharge_diagnosis	5.22.45	discharge_diagnosis
global_assessment_functioning	5.22.49.58.11	gaf

Table 1: Examples of diagnostic headers in the SecTag taxonomy.

	MIMIC discharge	exprivia-10	exprivia-100
total documents	59652	10	100
paragraphs per doc	30.57	7.7	26.77
diagnoses per doc	1.22	0.8	1.28
documents with no diagnosis	8674 (14.5%)	2 (20%)	27 (27%)
procedures per doc	0.71	N/A	N/A
documents with no procedure	20797 (35.86%)	N/A	N/A

Table 2: MIMIC-III discharge (silver annotation with SecTag) vs. Exprivia datasets (gold annotation).

found in SecTag are considered −diagnosis, −procedure. The headers are then removed from the document, thus forcing the model to learn paragraph classification from the textual content, relying on headers as a silver supervision signal.

While a typical MIMIC note has a single diagnostic paragraph, some contain multiple diagnostic fragments: (i) some notes span multiple related reports, where each report comes with its own diagnosis; (ii) some notes contain semantically different diagnostic sections (e.g., "admitting diagnosis" and "discharge diagnosis"); (iii) some notes cover complex cases and the diagnostic section is expressed in several (consecutive) paragraphs.

Since SecTag predates major MIMIC releases, some popular headers are missing—we have therefore manually extended the taxonomy (6.7k headers) to cover another 75 of the most popular headers. The expansion yielded a considerable increase in procedure paragraphs, augmenting drastically the number of positive examples for training the procedure classifier. At the same time, the overall precision improved, eliminating some consistent errors with diagnosis paragraphs. In what follows, we always rely on data preprocessed with expanded SecTag.

3.2 Italian Data: Exprivia Datasets

A large collection of discharge reports in Italian has been provided by Exprivia S.p.a. The documents show some similarity to MIMIC discharge reports: they are typically 0.5-1 page long, they can be split into paragraphs rather reliably, they exhibit a considerable variability in terms of the underlying discourse structure. Each document is associated with a set of ICD-9 codes for discharge diagnoses. Yet, similarly to MIMIC, no inline manual annotation is provided for identifying textual segments referring to diagnoses/procedures.

To provide accurate test data for our multilingual approach, a human expert has conducted a manual annotation of the Italian set. We have labeled a pilot of 10 notes and a random sample of 100 notes. The annotation only covered diagnosis as our pilot phase revealed that labeling procedure required considerably more elaborate guidelines and medical training.

Table 2 compares document statistics for discharge notes from MIMIC-III and Exprivia datasets. It suggests that the pilot can only be used as a very preliminary sample of the data: the notes are rather small and with few diagnoses. The Italian documents from *exprivia-100* show a striking similarity to MIMIC: there are on average around 25-30 paragraphs per document, 1.2-1.3 of which are diagnostic. The major difference comes from the documents with no diagnosis (27% in Italian, 14.5% in English). We believe that this similarity reflects the fact that, despite differences in national and local healthcare regulations as well as individual practicing/recording approaches, clinical notes reflect a common underlying semantics and thus a language transfer model can be successful for our task, mitigating the need for very time-consuming and costly expert effort on constructing taxonomies similar to SecTag in Italian.

4 Transformer-Based Architectures for Diagnosis and Procedure Extraction

Transformer-based models have recently become the standard in NLP. Models like BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020) showed impressive performance when compared to previous state of the art. These models are based on the Transformer block (Vaswani et al., 2017), which exploits the attention mechanism to find relations between all pairs of tokens in the input text and thus creates deep contextualized representations. Transformer layers can be stacked to create more powerful and refined models. For computational efficiency, we focus on architectures with no more than 12 layers.

Tokenization. Raw text cannot be provided directly to a transformer-based model: it is first tokenized using a fixed-size vocabulary, created via a segmentation algorithm, e.g., *WordPiece*. We extended BERT vocabulary to account for eventual deidentified medical input.

Pre-training and fine-tuning. Transformer-based models are usually trained in a 2-step fashion. The model is first *pretrained* on a huge amount of artificially labelled text taken from sources like Wikipedia or CommonCrawl. At the *fine-tuning* stage, the model is adapted to a specific task, e.g., Question Answering or Diagnosis Extraction. Since the model is already able to create good contextualized representations, the fine-tuning requires only a small amount of manually labelled examples. Following the common transformer fine-tuning practices, we classify paragraphs into ±`diagnosis` with a binary classification head on top of the first token output.

5 Language Transfer for Diagnosis Identification

The main bottleneck for NLP on medical data in Italian lies in the lack of annotated data and professionally created resources, similar to SecTag. To mitigate this issue, we advocate a language transfer approach, combining our transformer models (Section 4) with state-of-the-art machine translation (MT).

We investigate three cross-lingual setting. In the baseline set up, we do not perform any translation, relying on BERT's tokenizer and cross-lingual embeddings to learn informative sub-word clues for diagnostic paragraphs.

Transformer	Language	parameters
BERT-base-uncased	English	109M
BERT-base-cased	English	108M
ELECTRA-small	English	13M
BERT-Ita	Italian	110M
BERTino	Italian	68M

Table 3: Transformer models used in empiric evaluation

Our second cross-lingual pipeline builds directly upon the model presented in Section 4. We use an MT component to translate test documents from Italian into English, run our diagnosis identification model and then port the results to the Italian original via a trivial paragraph-level alignment. Note that this model is trained on high-quality data in English and tested on noisy automatically translated data.

For the third pipeline, we first translate the whole training set from English into Italian, while keeping paragraphs aligned. We follow the methodology from Section 4 to train a new model, operating on Italian directly. Note that, unlike the second pipeline, this approach implies training on noisy automatically translated data while testing on high-quality Italian. The effect of this is twofold: on one hand, the task becomes more difficult to learn, on the other hand, the resulting classifier should be more robust.

To obtain a satisfactory translation using open-source architectures, we rely on the transformer encoder-decoder models (Tiedemann and Thottingal, 2020) trained on the OPUS corpus[3]. While the OPUS corpus is not tailored specifically to the medical domain, its large size and generic nature allow for training very robust MT models. We exploit the two models to translate from English to Italian [4] and from Italian to English [5]. Both are transformer encoder-decoder models trained with the Causal Language Modeling objective.

6 Experiments

6.1 Setup

Data processing. We split the MIMIC III discharge dataset into training, development and test-

[3] https://opus.nlpl.eu

[4] https://huggingface.co/Helsinki-NLP/opus-mt-en-it

[5] https://huggingface.co/Helsinki-NLP/opus-mt-it-en

Task	Filt. Accuracy	Precision@1
Paragraph-level granularity		
Diagnosis	92.4	95.9
Procedure	97.1	98.4

Table 4: Diagnosis and procedure discourse segments identification, monolingual setting (English), document-level view: training, fine-tuning and testing on subsets of MIMIC-III discharge.

ing sets (60%, 20% and 20% respectively). We used the first for training all the models presented in this study, while we use the other two for checkpoint selection, hyper-parameter tuning (batch size and learning rate) and evaluating the monolingual model. We used the *exprivia-10* set for validation and *exprivia-100* set for testing in the cross-lingual (language transfer) experiments.

Transformer Models. We run most experiments in two modes: (i) with powerful transformer components comprising a large number of parameters and providing top performance such as BERT (Devlin et al., 2019) and BERT-ita[6] and (ii) with small and efficient transformer models such as ELECTRA small (Clark et al., 2020) and BERTino (Muffo and Bertino, 2020). The objective of this setup was to measure the performance/efficiency trade-off.

Table 3 presents all the used transformer models with the respective number of parameters.

Evaluation metrics. Diagnosis/Procedure classification task shows a very skewed label distribution. For this reason, we approach it from an information retrieval viewpoint, i.e., we rank paragraphs based on their probability of containing a diagnosis. We use Mean Average Precision and Precision@1 to evaluate the ranking quality. The former takes into account the whole ranking and is therefore the best indicator of the ranking quality. The latter indicates the number of times a correct diagnosis is returned in the first position. To provide a better comparison, we report MAP and P@1 averaging only over the documents that contain at least one diagnosis. We also report model accuracy in recognizing documents with no diagnoses (Filtering Accuracy). This metric was introduced because a relevant fraction of documents did not contain a diagnosis, see Table 2.

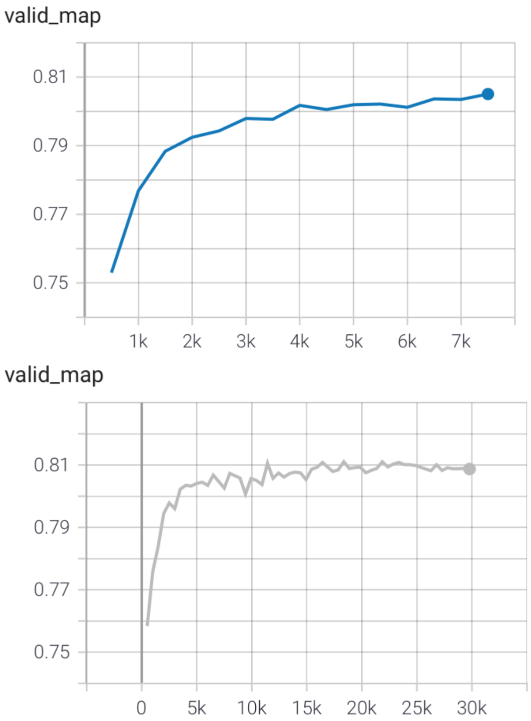

Figure 1: Learning curves on the *exprivia-10* validation set in the Italian pipeline: BERT-Ita (top) vs. BERTino (bottom). MAP (y-axis) for a given number of training steps (x-axis).

6.2 Results

Monolingual results. Table 4 summarizes the English results. The numbers refer to a BERT-base-cased model fine-tuned with a batch size of 64 and a learning rate of $2 * 10^{-6}$. The model is able to identify very accurately documents with no diagnoses/procedures (92.4% and 97.1% accuracy respectively). Moreover, the binary classification of paragraphs into diagnoses (or not), and procedures (or not) is very reliable: 95.9% and 98.4% P@1 at document level.

Cross-lingual experiments. Table 5 shows the results of our language transfer experiments. A moderate performance (58.8% Filtering Accuracy, 49.2% P@1) can be achieved via a BERT model trained on English MIMIC data and directly tested on the Italian *exprivia-100* set. Multilingual-BERT does slightly better as it was trained on 104 languages, English and Italian included. This approach relies on joint multilingual embeddings and fine tokenization. It can, for example, identify and align stems of Latin origin for some disease names. However, it cannot go much beyond: it is not able to model deep semantics related to medi-

[6] `https://huggingface.co/dbmdz/bert-base-italian-xxl-cased`

Model	Development set	Test set performance		
		Filt. Accuracy	Precision1	MAP
Cross-Lingual BERT				
BERT-base-uncased	exprivia-10	58.8	49.2	58.5
Multilingual-BERT-cased	exprivia-10	51.2	73.5	75.6
MT-based pipeline-2, train on English (MIMIC), test on English translation of exprivia-100				
BERT-cased	exprivia-10	31.8 (7.6)	67.4 (6.8)	69.2 (3.3)
BERT-cased	MIMIC dev	53.1 (9.0)	**73.9** (6.6)	**73.3** (4.9)
ELECTRA-small	exprivia-10	**64.6** (9.5)	60.5 (12.6)	71.2 (9.0)
ELECTRA-small	MIMIC dev	54.2 (8.7)	62.4 (11.2)	73.2 (7.9)
MT-based pipeline-3, train on Italian translation of MIMIC, test on Italian (exprivia-100)				
BERT-ita	exprivia-10	69.8 (6.2)	**78.6** (7.3)	81.5 (3.8)
BERT-ita	MIMIC dev	67.1 (7.8)	73.7 (3.0)	77.2 (3.1)
BERTino	exprivia-10	**72.0** (7.5)	74.9 (2.9)	81.9 (2.6)
BERTino	MIMIC dev	67.7 (4.1)	77.3 (2.5)	**83.3** (1.9)

Table 5: Language transfer models, fine-tuning on the MIMIC training set and evaluation on *exprivia-100* test set; boldface indicates the best results. Standard deviation across 5 runs shown in brackets.

cal processes.

The use of MT shows considerable improvement over the baseline. The results suggest a better performance for the setting where the training set is translated into Italian and the diagnosis extraction model is then learned on (noisy) Italian data. Moreover, this approach is much faster when used as a service, as it directly operates on Italian input.

We performed all the MT-based experiments 5 times using random seeds to enable a better statistical assessment of the results. While in general the standard deviation is rather small considering the very small test set, the setting with a translated test set leads to unstable benchmarking, especially for the smaller ELECTRA transformer.

Finally, smaller transformer models, especially BERTino, exhibit very small performance drops compared to larger transformers. This suggests that they are robust enough to capture paragraph-level diagnosis semantics. Therefore, it is possible to run the extraction service with low computational resources, e.g., using CPUs. Figure 1 shows the stability of the learning with translated training data. Small models are able to match the performance of larger models, being also faster to converge. We believe that smaller models overfit less the MIMIC training data, thus providing a final better performance on the Exprivia data. Note that training was stopped after a fixed amount of time for every experiment. BERTino, being smaller, is able to do more steps in the same amount of time.

7 Conclusion

We present a language transfer approach to unraveling discourse structure of clinical notes, focusing on diagnosis and procedure. We combine transformer-based paragraph modeling with state-of-the-art MT architectures in a novel application, that is essential for eHealth big data analytics. Most importantly, our language transfer approach helps mitigate the need for expensive and time-consuming medical resource creation (annotated train data as well as header taxonomy) in Italian.

We empirically investigate two translation-based architectures, showing that both of them outperform a generic cross-lingual pipeline. The approach based on translating train data is more robust and efficient (at runtime) compared to translating the test data, yielding more stable performance.

In future, we plan to expand our study to other discourse segments, such as allergy or history. However, our first experiments with procedure segments show that, unlike diagnosis, modeling and even annotating other headers require a more tight collaboration with medical experts.

8 Acknowledgements

The research presented in this paper has been supported by the Autonomous Province of Trento (project CareGenius). The computational power has been provided by the High Performance Computing department of the CINECA Consortium (ISCRA project CareGeni).

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Hong-Jie Dai, Shabbir Syed-Abdul, Chih-Wei Chen, and Chieh-Chen Wu. 2015. Recognition and evaluation of clinical section headings in clinical documents using token-based formulation with conditional random fields. *BioMed Research International*, 2015.

Joshua Denny, Randolph Miller, Kevin Johnson, and Anderson Spickard. 2008. Development and evaluation of a clinical note section header terminology. In *Proceeding of AMIA Annual Symposium*, pages 156–160.

Joshua Denny, Anderson Spickard, Kevin Johnson, Neeraja Peterson, Josh Peterson, and Randolph Miller. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association : JAMIA*, 16(6):806–15.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Matteo Muffo and E. Bertino. 2020. Bertino: An italian distilbert model. In *CLiC-it*.

Sara Rosenthal, Ken Barker, and Zhicheng Liang. 2019. Leveraging medical literature for section prediction in electronic health records. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4864–4873, Hong Kong, China, November. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.