Closing the Gap: Joint De-Identification and Concept Extraction in the Clinical Domain

Lukas Lange[1,2,3]   **Heike Adel**[1]   **Jannik Strötgen**[1]

[1] Bosch Center for Artificial Intelligence, Renningen, Germany
[2] Spoken Language Systems (LSV), Saarland University, Saarbrücken, Germany
[3] Saarbrücken Graduate School of Computer Science, Saarbrücken, Germany
`{Lukas.Lange,Heike.Adel,Jannik.Stroetgen}@de.bosch.com`

Abstract

Exploiting natural language processing in the clinical domain requires de-identification, i.e., anonymization of personal information in texts. However, current research considers de-identification and downstream tasks, such as concept extraction, only in isolation and does not study the effects of de-identification on other tasks. In this paper, we close this gap by reporting concept extraction performance on automatically anonymized data and investigating joint models for de-identification and concept extraction. In particular, we propose a stacked model with restricted access to privacy-sensitive information and a multi-task model. We set the new state of the art on benchmark datasets in English (96.1% F1 for de-identification and 88.9% F1 for concept extraction) and Spanish (91.4% F1 for concept extraction).

1 Introduction

In the clinical or biomedical domain, natural language processing (NLP) could significantly improve the efficiency and effectiveness of processes. For example, the extraction of structured information from clinical narratives can help in decision making or drug repurposing (Marimon et al., 2019). However, the automatic processing of documents with privacy-sensitive content is restricted due to the necessity of applying anonymization techniques.

Text anonymization, also called de-identification, aims at detecting and replacing protected health information (PHI),[1] such as patient names, age and phone numbers, as shown in the upper part of Figure 1. Recent studies show that automatic de-identification leads to
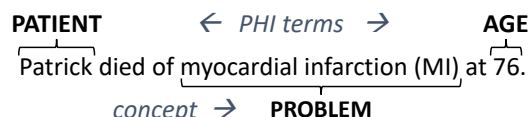

Figure 1: Sentence with annotations of the two tasks.

promising results (Stubbs and Uzuner, 2015; Marimon et al., 2019). A severe limitation of current approaches in research, however, is that de-identification is typically addressed in isolation but not together with a downstream task, such as concept extraction (CE) from medical texts (Gonzalez-Agirre et al., 2019; Uzuner et al., 2011). Instead, the downstream task models are trained and evaluated on the non-anonymized data, and it remains unclear how de-identification affects their performance in real-world settings.

In this paper, we argue that to evaluate the effectiveness of NLP in the medical domain, the tasks of de-identification and information extraction should be analyzed together. Our contributions are as follows: We close this gap and analyze the effect of de-identification on clinical concept extraction. Moreover, we consider the two tasks jointly and propose two end-to-end models: A multitask model that shares the input representation across tasks, and a stacked model that trains a pipeline of de-identification and concept extraction in an end-to-end manner. For the stacked model, we propose to use a masked embedding layer to restrict the access of the concept detector to privacy-sensitive information and train it on an anonymized version of the data. To make the model differentiable, we use the Gumbel softmax trick (Maddison et al., 2017; Jang et al., 2017).

We conduct experiments on clinical benchmark datasets in English and Spanish. Our results indicate that de-identification does not affect CE models negatively, but has even a slight positive effect

---

[1] PHI types are typically defined by governments, for instance in the Health Insurance Portability and Accountability Act (HIPAA) of the United States.
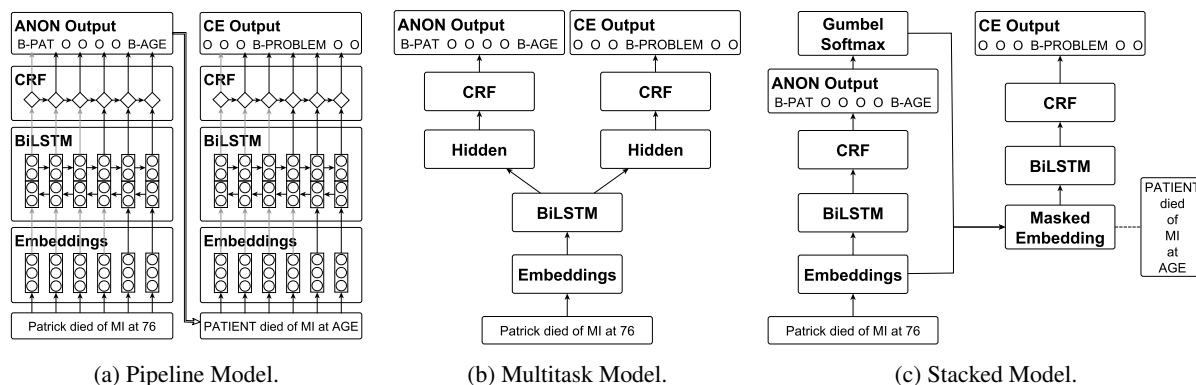
Figure 2: Overview of our different model architectures. While the CE model in the multitask setting has access to all privacy-sensitive information, the access of the pipeline CE model and the stacked CE model is restricted by the ANON output. "PAT" stands for Patient. The labels are encoded in BIO format.

on the results, probably because de-identification homogenizes the input for CE. Modeling both tasks jointly leads to better results than treating de-identification as a pure preprocessing step, resulting in new state-of-the-art performance for CE.

For future research, we publish our code.[2]

2 Related Work

While many works propose joint training for other NLP tasks (i.a., Finkel and Manning, 2009; Miwa and Sasaki, 2014), including multitask learning (i.a., Collobert and Weston, 2008; Klerke et al., 2016; Søgaard and Goldberg, 2016) and stacking of pipeline components (i.a., Miwa and Bansal, 2016), we are to the best of our knowledge the first to combine de-identification with an information extraction task. In this section, we report related work in those two fields.

2.1 De-Identification

The increasing importance of de-identification is reflected in the number of shared tasks (Uzuner et al., 2007; Stubbs and Uzuner, 2015; Marimon et al., 2019). State-of-the-art methods for de-identification typically rely on recurrent neural networks (RNNs) (Dernoncourt et al., 2016; Lange et al., 2019b; Kajiyama et al., 2018).

Feutry et al. (2018) and Friedrich et al. (2019) create pseudo-de-identified text representations with adversarial training. In particular, they replace personal information, such as names, by other names. Zhao et al. (2018) augment the training data by creating more general text skeletons, e.g., by replacing rare words, such as names, by a

special unknown token. Compared to these works, we exploit the advantages of both approaches and replace personal information by their class names as placeholders. This approach is not only common for de-identification (Johnson et al., 2016), but also for relation extraction where entities are often either replaced by their type or enriched with type information (i.a., Zhang et al., 2017; Miwa and Sasaki, 2014). We further motivate our choice in Section 3.2. Another difference to the above mentioned works is that we do not augment the training data for our de-identification model.

2.2 Medical Information Extraction

Analogously, there have been a series of shared tasks for information extraction in the clinical and biomedical domain (Uzuner et al., 2011; Sun et al., 2013; Krallinger et al., 2015; Gonzalez-Agirre et al., 2019). Models for these tasks often either rely on hand-crafted features (Leaman et al., 2015; Xu et al., 2012) or RNNs (Hemati and Mehler, 2019; Korvigo et al., 2018; Tourille et al., 2018). Newman-Griffis and Zirikly (2018) study the performance of RNNs for medical named entity recognition in the context of patient mobility and find that they benefit from domain adaption.

In contrast to previous work, we investigate the usage of de-identified texts as input for clinical concept extraction models and propose to jointly model de-identification and concept extraction.

3 Model

In this section, we present our systems for the two individual tasks and our proposed joint models. Figure 2 shows the respective architectures.

3.1 General Architecture

We model both document anonymization (ANON) and clinical concept extraction (CE) as sequence labeling problems and apply a bidirectional long short-term memory (BiLSTM) network (Hochreiter and Schmidhuber, 1997) with a conditional random field (CRF) output layer (Lafferty et al., 2001), similar to Lample et al. (2016). In recent works on clinical de-identification and CE, this architecture was shown to be very promising (Marimon et al., 2019; Gonzalez-Agirre et al., 2019).

Each token is represented with a concatenation of different pre-trained language-specific embeddings: byte-pair-encoding (Heinzerling and Strube, 2018), fastText (Bojanowski et al., 2017) and FLAIR (Akbik et al., 2018). For Spanish, we also include multilingual BERT embeddings (Devlin et al., 2019). Further, we include the following domain-specific embeddings: clinical BERT for English pre-trained on discharge summaries (Alsentzer et al., 2019) and clinical fastText for Spanish pre-trained on the Spanish E-health corpus from the Scielo archive (Soares et al., 2019).

3.2 Pipeline Model

To assess the effects of de-identification on CE, we first apply the de-identification model to anonymize the CE dataset and then evaluate the CE model on the anonymized data. We refer to this approach as PIPELINE model (see Figure 2a). For anonymization, we replace each detected privacy-sensitive term with a placeholder of its PHI type, i.e., there is one placeholder per type.

This replacement choice has advantages over the alternatives described in Section 2. Compared to replacing personal information with alternative names, it leads to a more general text and thus, homogenizes the input for the downstream-task classifier. Compared to replacing all personal information with the same token, the resulting text is more specific, allowing the downstream-task classifier to take into account which kind of personal information was mentioned. Thus, the approach is a trade-off between more homogeneous input and more fine-grained information for the downstream-task classifier.

3.3 Joint Models

Instead of using a sequential pipeline, we propose to train both tasks jointly. For this, we test two approaches: a multitask model and a stacked model.
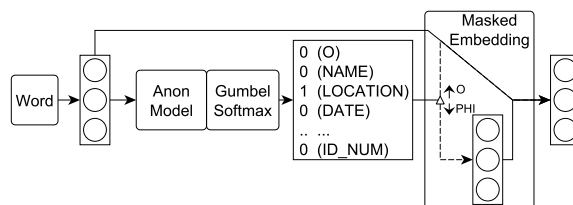

Figure 3: Masked Embedding.

3.3.1 Multitask Model

In the MULTITASK model (Figure 2b), the weights up to the BiLSTM layer are shared across both tasks. For each task, we add a task-specific hidden layer with ReLU activation and a CRF output layer. Note that in this architecture, the CE model has access to the original, privacy-sensitive data.

3.3.2 Stacked Model

We also propose a STACKED model (Figure 2c), where only the de-identification part has access to the privacy-sensitive information. The access of the CE part is restricted by a masked embedding layer as described in the following.

Masked Embedding Layer. The masked embedding layer ensures that the CE model does not have access to privacy-sensitive information by replacing the input embeddings of privacy-sensitive tokens by PHI-class embeddings which are randomly initialized and fine-tuned during training. This is depicted in Figure 3.

Gumbel Softmax Trick. The masked embedding layer requires a discrete output from the de-identification part. In order to ensure that the model stays fully differentiable, we use the Gumbel softmax trick (Maddison et al., 2017; Jang et al., 2017). It approximates categorical samples with a continuous distribution on the simplex and computes gradients for backpropagation with the reparameterization trick. The Gumbel softmax function has the following form:

$$y_k^\tau = \frac{\exp((\log \alpha_k + G_k)/\tau)}{\sum_{i=1}^{K} \exp((\log \alpha_i + G_i)/\tau)} \quad (1)$$

with $\alpha_1, ...\alpha_K$ being the unnormalized output scores from the de-identification layer, $G_1, ..., G_K$ being i.i.d samples drawn from Gumbel(0, 1) and τ being a temperature. For $\tau \to 0$, the distribution becomes identical to the categorical distribution.

The masked embedding layer takes the output of the Gumbel softmax (i.e., an anonymization label) and if the label is a PHI class and requires anonymization, the masked embedding layer uses

	English (i2b2)		Spanish	
	ANON	CE	ANON	CE
# classes	24	3	22	3
train (# tokens)	45,793	16,315	15,903	8,068
dev (# tokens)	5,088	-	8,277	3,748
test (# tokens)	32,587	27,625	7,966	3,930

Table 1: Dataset statistics. # classes denotes the number of classes without the negative 'O' class.

4 Experiments

In this section, we describe the datasets used in our experiments, and training details for our models. Finally, we present our results and analysis.

4.1 Data and Model Training and Evaluation

We evaluate our models on corpora from the clinical domain in English and Spanish. For English, we use the data from the i2b2 2010 CE task (Uzuner et al., 2011) and the i2b2 2014 de-identification task (Stubbs and Uzuner, 2015). For Spanish, we use the MEDDOCAN (Marimon et al., 2019) corpus for de-identification and the PharmaCoNER corpus (Gonzalez-Agirre et al., 2019) for CE. As PharmaCoNER is a subset of MEDDOCAN, we have both gold-standard concept and de-identification annotations for this data.

Data Preprocessing. We used the preprocessing scripts from Alsentzer et al. (2019) for the English i2b2 corpora and the Spanish Clinical Case Corpus tokenizer (Intxaurrondo, 2019) for both Spanish corpora. We noticed that the Spanish tokenizer sometimes merges multi-word expressions into a single token joined with underscores for contiguous words. As a result, some tokens cannot be aligned with the corresponding entity annotations. To address this, we split those tokens into their components in a postprocessing step. Table 1 shows statistics about corpora sizes.

Hyperparameters. The embeddings have 300 (byte-pair-encoding), 300 (fastText) and 4,048 (FLAIR) dimensions. For English, we use clinical BERT embeddings with 768 dimensions which are constructed by averaging the last four layers with the scalar mix operation proposed by Liu et al. (2019). We concatenate all embeddings to one input vector, resulting in a total input dimensionality of 5,416. Analogously, we use multilingual BERT (768 dim.) and domain-specific fastText embed-

Models	English	Spanish
Yang and Garibaldi (2015)	96.0	
Zhao et al. (2018)	94.0	
Alsentzer et al. (2019)	93.0	
Lange et al. (2019b)		**97.0**
Hassan et al. (2019)		96.3
Perez et al. (2019)		96.0
Our PIPELINE (ANON only)	**96.1**	96.8
Our STACKED	95.9	96.8
Our MULTITASK	95.2*	96.7

Table 2: F1 results for de-identification. * highlights our models with statistically significant differences compared to PIPELINE (ANON only).

dings (100 dim.) for Spanish, resulting in 5,516 input dimensions. For the LSTM, we use 256 hidden units per direction. The task-specific hidden layer of the multitask model has 128 units.

Training. For training, we use stochastic gradient descent with a learning rate of 0.1 and a batch size of 32 sentences. The learning rate is halved after 3 consecutive epochs without improvements on the development set. For the joint models, we pretrain the anonymization part for 3 epochs and, then, use a higher initial learning rate of 0.2 for the concept extraction part. We perform early stopping on the development set. If no development set was provided by the corpus (i2b2 2010 corpus), we held out 10% of the training set as development set. Note that we use the same hyperparameters for all our models and all tasks, which were tuned on the Spanish concept extraction data.

Evaluation. We train each model with three random initializations and report F1 for exact matching for the best model in all experiments. We perform statistical significance testing to check if our joint models are better than the PIPELINE model. We use paired permutation testing with 2^{20} permutations and a significance level of 0.05.

4.2 Results

Table 2 shows that the de-identification component of our PIPELINE model which was trained on the single task of de-identification sets the new state of the art on English and performs comparable on Spanish. The performance difference to our prior work (Lange et al., 2019b) is due to a slightly different set of input embeddings. However, we found no statistically significant differences to that model. The de-identification performance of STACKED is comparable to the PIPELINE model, the de-identification performance of MULTITASK

Models	No PHI	English	Spanish
de Bruijn et al. (2010)	no	85.2	
Xu et al. (2012)	no	84.9	
Alsentzer et al. (2019)	no	87.7	
Sun and Yang (2019)	no		89.2
Lange et al. (2019a)	no		88.6
Our PIPELINE	**yes**	88.0	89.6
Our STACKED	**yes**	88.7*	90.0
Our MULTITASK	no	**88.9***	**90.3***
Xiong et al. (2019)†	no		91.1
Stoeckel et al. (2019)†	no		90.5
Our MULTITASK†	no		**91.4***

Table 3: F1 results for concept extraction. We indicate for each model whether anonymized data is used during the extraction training. † indicates models which are trained on a combination of training and development set. * highlights our models with statistically significant differences compared to our PIPELINE.

Embedding	English	Spanish
fastText	81.5	78.7
byte-pair-encoding	83.4	83.9
FLAIR	83.0	82.4
Multilingual BERT	84.4	85.9
Clinical BERT (English)	87.2	-
Clinical fastText (Spanish)	-	79.7
Concatenation of all	88.1	89.7

Table 4: Effects of different embeddings on the concept extraction task (without anonymization).

is slightly lower, however, we only found statistically significant differences for the MULTITASK model for English.

The results for our concept extraction models in comparison to state of the art are shown in Table 3. We set the new state of the art on both languages. While in the PIPELINE setting, the CE performance is slightly lower (as it has been trained on the non-anonymized texts but is evaluated on the de-identification output), training de-identification and CE jointly leads to considerable improvements for both STACKED and MULTITASK with statistically significant differences for both models in English and for MULTITASK also in Spanish. Especially the results of STACKED in comparison to PIPELINE shows that end-to-end training of the two steps is promising, while still preserving privacy aspects during model training by restricting internal access to PHI tokens. The performance of each embedding used in our experiments is shown in Table 4. As mentioned before, we did not include multilingual BERT embeddings for English, but show their results for completeness.

	train on	test on	dev	test
(1)	non-anon.	non-anon.	89.2	89.7
(2)	non-anon.	anon-predicted	89.1	89.6
(3)	anon-predicted	non-anon.	89.2	89.6
(4)	anon-predicted	anon-predicted	89.6	90.0
(5)	anon-gold	anon-predicted	89.5	90.0
(6)	anon-gold	anon-gold	89.6	90.1

Table 5: Pipeline analysis results on Spanish concept extraction. "non-anon." indicates the original text without anonymization; "anon-gold" and "anon-predicted" refer to texts with replacements for gold/predicted de-identification labels, respectively.

4.3 Analysis of Pipeline Setting

Finally, we analyze the impact of de-identification on CE. The results for training and testing our CE model on different inputs (non-anonymized vs. anonymized) are shown in Table 5. We restrict our analysis to Spanish since the data is labeled with both de-identification and concept information (see Section 4.1). Thus, we can also investigate the difference between gold and predicted de-identification labels. The CE model benefits from being trained and evaluated on anonymized data (lines 4-6). However, it hurts to train on non-anonymized data and evaluate on predicted de-identification labels (line 1 vs. 2) and vice versa (line 1 vs. 3). This supports our motivation that it is necessary to investigate anonymization and downstream applications together. The difference of training on gold vs. predicted de-identification labels (lines 4-6) is only marginal, suggesting that state-of-the-art de-identification systems are good enough to be used in such settings.

5 Conclusion

In this paper, we close the gap and consider de-identification of clinical text together with concept extraction, a possible downstream application. We investigate the effects of de-identification on concept extraction and show that it positively influences the concept extraction performance. We propose two models to learn both tasks jointly, a multitask model and a stacked model, and set the new state of the art on medical concept extraction benchmark datasets for English and Spanish.

Acknowledgments

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2010. NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning*, pages 160–167, Helsinki, Finland. ACM, ACM.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. 2018. Learning anonymized representations with adversarial neural networks. *arXiv preprint arXiv:1802.09386*.

Jenny Rose Finkel and Christopher D. Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 326–334, Boulder, Colorado. Association for Computational Linguistics.

Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2019. Adversarial learning of privacy-preserving text representations for de-identification of medical records. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5829–5839, Florence, Italy. Association for Computational Linguistics.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Fadi Hassan, Mohammed Jabreel, Najlaa Maaroof, David Sánchez, Josep Domingo-Ferrer, and Antonio Moreno. 2019. Recrf: Spanish medical document anonymization using automatically-crafted rules and CRF. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing*, pages 727–734, Bilbao, Spain. CEUR Workshop Proceedings.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation)*, Miyazaki, Japan. European Language Resources Association.

Wahed Hemati and Alexander Mehler. 2019. Lstmvoter: chemical named entity recognition using a conglomerate of sequence labeling tools. *Journal of Cheminformatics*, 11(1):3.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computing*, 9(8):1735–1780.

Ander Intxaurrondo. 2019. SPACCC (spanish clinical case corpus) tokenizer.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations*, Toulon, France.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, 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*, 3:160035.

Kohei Kajiyama, Hiromasa Horiguchi, Takashi Okumura, Mizuki Morita, and Yoshinobu Kano. 2018. De-identifying free text of Japanese dummy electronic health records. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 65–70, Brussels, Belgium. Association for Computational Linguistics.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Ilia Korvigo, Maxim Holmatov, Anatolii Zaikovskii, and Mikhail Skoblov. 2018. Putting hands to rest: efficient deep cnn-rnn architecture for chemical named entity recognition with no hand-crafted rules. *Journal of Cheminformatics*, 10(1):28.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. Chemdner: The drugs and chemical names extraction challenge. *Journal of Cheminformatics*, 7(1):S1.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Lukas Lange, Heike Adel, and Jannik Strötgen. 2019a. NLNDE: Enhancing neural sequence taggers with attention and noisy channel for robust pharmacological entity detection. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 26–32, Hong Kong, China. Association for Computational Linguistics.

Lukas Lange, Heike Adel, and Jannik Strötgen. 2019b. NLNDE: the neither-language-nor-domain-experts' way of spanish medical document de-identification. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing*, pages 671–678, Bilbao, Spain. CEUR Workshop Proceedings.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmchem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7(1):S3.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing*, pages 618–638, Bilbao, Spain. CEUR Workshop Proceedings.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869, Doha, Qatar. Association for Computational Linguistics.

Denis Newman-Griffis and Ayah Zirikly. 2018. Embedding transfer for low-resource medical named entity recognition: A case study on patient mobility. In *Proceedings of the BioNLP 2018 workshop*, pages 1–11, Melbourne, Australia. Association for Computational Linguistics.

Naiara Perez, Laura García-Sardiña, Manex Serras, and Arantza del Pozo. 2019. Vicomtech at MEDDOCAN: medical document anonymization. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing*, pages 696–703, Bilbao, Spain. CEUR Workshop Proceedings.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. 2019. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Manuel Stoeckel, Wahed Hemati, and Alexander Mehler. 2019. When specialization helps: Using pooled contextualized embeddings to detect chemical and biomedical entities in Spanish. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 11–15, Hong Kong, China. Association for Computational Linguistics.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of Biomedical Informatics*, 58:S20–S29.

Cong Sun and Zhihao Yang. 2019. Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, Hong Kong, China. Association for Computational Linguistics.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Julien Tourille, Matthieu Doutreligne, Olivier Ferret, Aurélie Névéol, Nicolas Paris, and Xavier Tannier. 2018. Evaluation of a sequence tagging tool for biomedical texts. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 193–203, Brussels, Belgium. Association for Computational Linguistics.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Ying Xiong, Yedan Shen, Yuanhang Huang, Shuai Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Jun Yan, and Yi Zhou. 2019. A deep learning-based system for PharmaCoNER. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 33–37, Hong Kong, China. Association for Computational Linguistics.

Yan Xu, Kai Hong, Junichi Tsujii, and Eric I-Chao Chang. 2012. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5):824–832.

Hui Yang and Jonathan M Garibaldi. 2015. Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, 58:S30–S38.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Yue-Shu Zhao, Kun-Li Zhang, Hong-Chao Ma, and Kun Li. 2018. Leveraging text skeleton for de-identification of electronic medical records. *BMC medical informatics and decision making*, 18(1):18.