

Tumor Morphology Mentions Identification Using Deep Learning and Conditional Random Fields

Utpal Kumar Sikdar^a, Björn Gambäck^b and M Krishna Kumar^c

^aIBS Software Pvt. Ltd., Trivandrum, Techopark main gate, India-695581

^bDepartment of Computer Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway

^cIBS Software Pvt. Ltd., Trivandrum, Techopark main gate, India-695581

Abstract

The paper reports the application of several machine learning methods to the task of automatically finding tumor morphology mentions in Spanish clinical texts. Three setups based on Conditional Random Fields (CRF) techniques with different feature combinations were tested as well as a deep learning model (Bi-directional-LSTM-CNN). The best performance was achieved by combining two of the CRF-based learners and the neural network using a majority voting ensemble.

Keywords

named entity recognition, CRF, Bi-LSTM, CNN, GloVe

1. Introduction

To understand diseases, we need to extract certain key entities such as symptoms, duration, patient age and weight, etc. from unstructured textual medical data. This task, *clinical text mining*, is important to enable better clinical decision-making. It is, for example, very helpful if we can extract key entities from a pandemic situation (such as COVID-19, SARS, and locations) and take appropriate actions based on the disease symptoms and their attributes. Natural Language Processing fills an important role in extracting such key entities from different types of textual sources in various languages.

A myriad of medical texts are generated each day in various languages. Only in Spanish, almost a thousand electronic patient records are generated every minute. Hence automatically processing clinical texts in Spanish is a challenging task, but with a large potential for the medical user community as well as for the pharmaceutical industry and the patients.

Similar to Named Entity Recognition, tumor mention identification is a sequence labelling task. Following results published by several researchers in 2016 [1, 2, 3], state-of-the-art work on such sequence labelling tasks has focused on deep learning setups using a neural network structure, in particular Long Short-Term Memory Recurrent Neural Networks [LSTM; 4], followed by

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: utpal.sikdar@gmail.com (U.K. Sikdar); gamback@ntnu.no (B. Gambäck); krishna.kumar@ibsplc.com (M.K. Kumar)

URL: <https://www.linkedin.com/in/dr-utpal-kumar-sikdar-31a1779b/> (U.K. Sikdar);


<https://www.ntnu.edu/employees/gamback> (B. Gambäck);

<https://www.linkedin.com/in/m-krishna-kumar-56383220/> (M.K. Kumar)

ORCID: 0000-0002-5252-707X (B. Gambäck)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

a sequential Conditional Random Field [CRF; 5] layer. Hence Ma and Hovy [2] introduced a neural network architecture that benefits from both word and character level representations automatically, by using a combination of bidirectional LSTM (bi-LSTM), CNN (Convolutional Neural Network) and CRF. They tested on the CoNLL 2003 English NER dataset [6], obtaining 97.55% accuracy for part-of-speech tagging and 91.21% F_1 score for Named Entity Recognition. Chalapathy et al. [7] applied an LSTM-CRF classifier to the 2010 i2b2/VA Natural Language Processing Challenges for Clinical Records data [8], outperforming previous word on the data set. Habibi et al. [9] then compared the same setup to the best CRF-based results on 33 different evaluation sets in the biomedical domain, with the LSTM-CRF structure achieving F-scores on average 5% above the CRF baselines, mainly due to increased recall.

However, most recent work has looked into other ways to obtain similar results. Straková et al. [10] compared a LSTM-CRF model to a sequence-to-sequence (seq2seq) model, where the input sequence are the tokens encoded by a bi-LSTM and the matching output label sequence is predicted by an LSTM decoder, showing the seq2seq model to outperform the state-of-the-art models for recognising nested named entities, while the baseline LSTM-CRF model still was competitive on simpler, flat NER structures. Overall focus in NLP in 2017 turned to pre-training using transformer-based neural models such as BERT, Bidirectional Encoder Representations from Transformers [11], and Baevski et al. [12] hence report top results ($F_1 = 92.8$) on the CoNLL 2003 NER dataset with a bi-directional transformer model.

The present work looks at various ways to compare and combine neural network learners (bi-LSTM and CNN) with conditional field classifiers, but rather than utilising the CRF directly as a layer in the deep learning setup, it is used in parallel to the network in an ensemble strategy, inspired by previous work on named entity recognition in social media data [13] and for under-resourced languages [14, 15].

Experiments were carried out on the Spanish CANTEMIST (CANcer TExt Mining Shared Task – tumor named entity recognition) data [16], which is introduced in the next section together with methods needed to further process the data in order to use it in the system setups described in Section 3. Experimental results are reported in Section 4 and analysed in Section 5. Finally, Section 6 sums up and points to ways the work could be extended and potentially improved.

2. Datasets and Preprocessing

The CANTEMIST shared task organisers provided training, development-1, development-2, and test data [16]. The data had been annotated using the ‘brat’ format [17]. Statistics of the datasets are reported in Table 1.

The training and development sets include data in a plain text file (.txt) together with a file containing the ‘brat’ annotation (.ann). The test data include the text file only. All tumor morphology mentions are annotated according to their corresponding character offsets in UTF-8 encoded plain text medical documents. The organisers provided 5,232 documents in the test data set, but out of those only 300 were actually utilised for evaluation purposes.

Since CRFs cannot handle the ‘brat’ stand-off annotation format directly, Begin-Inside-Outside (BIO) tags had to be converted from the ‘brat’ notation by aligning their character offsets to the

Table 1
Shared Task training and development dataset statistics

Data set	Number of Documents	Number of Mentions
Training	501	6,272
Development-1	250	3,258
Development-2	250	2,607
Total	1,001	12,137

character offsets of the tokens: after tumor mentions identification, the BIO tags are converted to the 'brat' stand-off annotation format with the offsets of tumor mentions with respect to the plain text file provided by the shared task organisers. The NLTK [18] tool was used to tokenise the plain text. Finally, for evaluation purposes, the given gold label annotation is compared to the annotated 'brat' stand-off format given by the predicted BIO tagging assigned by the experimental models described below.

3. CANTEMIST NER Identification

Within medical text processing, the task of named entity recognition is to identify medical entities from the unstructured clinical data. To identify CANTEMIST NERs, several methods were tested on the BIO tagging converted data, including Conditional Random Fields and combining a Bi-directional Long Short Term Memory network with a Convolutional Neural Network [19]. A majority voting ensemble approach was also applied to combine the outputs of the different methods.

3.1. Conditional Random Fields

Conditional Random Field classifiers were chosen as baseline indicators since they have produced state-of-art results on sequence labelling tasks such as named entity recognition in different domains. Three systems were developed for identification of tumor morphology mentions from the unstructured text, using different feature sets. The classifiers were trained using the following three sets of features, with the first two sets being based on the focus word itself (textual features types resp. binary flag types) and the third feature set being based on information extracted from the word's context.

- **Textual features**

- focus word (current word)
- word-lower (lower case version of the focus word)
- word-normalised (all upper case characters, lower case, digits and other characters are replaced by 'A', 'a', '0' and '_', respectively)
- word stem
- suffix n-grams (last one, two or three characters)
- prefix n-grams (last one, two or three characters)

- **Binary features**

Table 2
CRF model parameters and basic settings

Parameter name	Values
algorithm	lbfgs [21]
c1	0.05
c2	0.01
max_iterations	60
min_freq	0
all_possible_transitions	True

- is-all-upper-case
- starts-with-upper-case
- is-upper-case-middle
- is-any-digit
- is-single-digit
- is-double-digit
- is-any-punctuation-character
- is-any-under_score
- is-any-special-character (based on a list of special characters extracted from the data provided by the organisers, e.g., ‘%’, ‘±’, ‘,’, and ‘’)

- **Contextual features**

- local context (with a -m to +n window, i.e., from m preceding to n following tokens)
- beginning-of-sentence
- end-of-sentence

For the experiments, crfsuite [20] was utilised as implemented in the sklearn package.¹ The parameters and values given in Table 1 were used as the default model during the tumor mentions extraction process.

3.2. Bi-directional-LSTM-CNN

A bi-directional Long Short-Term Memory network was also applied to identify tumor mentions, with a Convolutional Neural Network used to induce character-level inputs as features to the model. Word features along with the character-level information map each word in the input string to potential tumor mention scores for the different categories. The following features were used as model inputs for the tumor mention identification network:

Word Embeddings

A publicly available GloVe [22] word embedding for Spanish.² For each word, a vector of size 300 was extracted from the pre-trained word embedding model.

Character Embeddings

A uniform distribution with range [0.5, 0.5] and size 52 was used as character embedding

¹<https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html>

²<https://github.com/dccuchile/spanish-word-embeddings>

for the CNN layer. The character sets include all characters in the training and test data together with PADDING values.

Word Label Features

To guide the network, some binary features were extracted explicitly from the input word, namely starts-with-upper-case, is-all-upper-case, is-any-digit, is-all-digit, and 'other' (if the word does not match any of the previous features). These were fed to the model as word label features.

Word Probability Features

Using a Naïve Bayes approach, a probability was assigned to each word as to whether it belonged to either the tumor mentions category or the non-tumor mentions category. The probability $\hat{P}(w_i, c)$ of a word w_i belonging to a category c (c is either the mentions category m or the non-mentions n) was calculated based on its conditional probability $\hat{P}(w_i|c)$ given the category c and the category's prior probability, $\hat{P}(c)$:

$$\hat{P}(w_i, c) = \frac{\hat{P}(w_i|c) * \hat{P}(c)}{\sum_{c_k \in \{m, n\}} \hat{P}(w_i|c_k) * \hat{P}(c_k)} \quad , \quad c \in \{m, n\} \quad (1)$$

where the conditional probability was calculated using add-one (Laplace) smoothing as:

$$\hat{P}(w_i|c) = \frac{freq(w_i, c) + 1}{(\sum_{w \in D} freq(w, c)) + |D|} \quad , \quad c \in \{m, n\} \quad (2)$$

with $|D|$ being the size of the dictionary (all word unique unigrams in the data) and $freq(w_i, c)$ the frequency of word w_i in category c , while the prior probabilities for the categories were calculated based on the total number of mentions and non-mentions (again using Laplace smoothing):

$$\hat{P}(c) = \frac{freq(c) + 1}{(\sum_{c_k \in \{m, n\}} freq(c_k)) + 1} \quad , \quad c \in \{m, n\} \quad (3)$$

If the probability $\hat{P}(w_i, c)$ was higher than 0.5, the word was considered to belong to that mentions category.

3.3. Majority Voting Ensemble

Based on the feature combinations, three models were generated using Conditional Random Fields. Another model was developed using the Bi-directional-LSTM-CNN network introduced in Section 3.2. A fifth ensemble-based model was created by taking the majority vote of the outputs of two of the CRF models and the Bi-LSTM-CNN model.

4. Results

Five test runs were submitted to the shared task using different models trained on the development set data. The different models used in the five runs were created as described in the previous section and characterised as follows:

Table 3
Development-1 data results

	Precision	Recall	F-score
Run-1	0.768	0.728	0.748
Run-2	0.774	0.735	0.754
Run-3	0.768	0.727	0.747
Run-4	0.691	0.710	0.700
Run-5	0.779	0.735	0.757

Run-1: A Conditional Random Fields classifier along with the features mentioned in Section 3.1 using a context window size of two preceding to two following words.

Run-2: A CRF classifier utilising fewer features than the one in Run-1, namely only: current word, word stem, prefix of two and three characters, suffix of two and three characters, starts-with-upper-case, is-upper-case-middle, is-any-digit, is-single-digit, and end-of-sentence, together with a context of two preceding and *one* following words.

Run-3: A third CRF model, again with a context of two preceding and one following words, combined with probability of the current word belonging to a tumor mention as assigned by the Naïve Bayes classifier, as well as the probability given by the Naïve Bayes classifier of the current word belonging to the non-tumor mention category, and utilising the following additional features: current word, word-lower-case, word stem, prefix of one and two characters, suffix of two characters, starts-with-upper-case, is-all-upper-case, is-single-digit, is-double-digit, is-any-under_score, and end-of-sentence.

Run-4: The bi-directional-LSTM-CNN model described in Section 3.2.

Run-5: A combination of the outputs of Run-1, Run-2 and Run4 using majority voting. If the outputs of all three models differed, the mention category was chosen randomly among the outputs.

The shared task organisers provided an evaluation script to measure system performance based on micro-averaged precision, recall and F_1 -score. The five models described above were tested on the two data development sets, development-1 and development-2. As can be seen in Table 3, Run-5 performed best on the development-1 data, with micro-average precision, recall and F-score values of 77.9%, 73.5% and 75.7%, respectively.

However, on the development-2 data set, Run-2 performed best of the models, with micro-averaged precision, recall and F_1 values of 75.8%, 74.7% and 75.3%, respectively. The performance of all five models in terms of micro-averaged precision, recall and F-score on development-2 are reported in Table 4.

All the models were applied to the blind test data which was provided by the shared task organisers. During testing, the shared task training and development data sets were merged and used as training set to build the models that were then applied to the test data.

Table 4
Development-2 data results

	Precision	Recall	F-score
Run-1	0.753	0.742	0.747
Run-2	0.758	0.747	0.753
Run-3	0.743	0.739	0.741
Run-4	0.675	0.739	0.701
Run-5	0.757	0.746	0.752

Table 5
Blind test data results

	Precision	Recall	F-score
Run-1	0.758	0.746	0.752
Run-2	0.746	0.745	0.746
Run-3	0.756	0.747	0.751
Run-4	0.697	0.751	0.723
Run-5	0.765	0.764	0.764

On the unseen test data, the Run-5 ensemble model outperformed all the other models, with the micro-averaged precision, recall and F-score values of 76.5%, 76.4% and 76.4%, respectively. The test results of all models are reported in Table 5.

5. Error Analysis and Discussion

As can be seen in the tables in the previous section, the variations between the five models are small, in particular in terms of recall. However, the deep learner of Run-4 in general performed poorer in terms of precision than the CRF-based models. On the other hand, the deep learner actually showed slightly better recall than the CRF-based models on the unseen test data, indicating that it is better at generalising.

Closer analysing the outputs on the two development data sets, Table 6 shows the confusion matrices for Run-1. It is clear that many of the tumor mentions were not identified by the system, with many mentions miss-classified into other categories. In particular the I(nside) and O(utside) mention tags often got confused, while the system in general was better at pin-pointing the B(egin) mention category.

A common kind of error was found on multi-word tumor mentions due to incorrect boundary identification, such as finding only *carcinoma ductal* instead of the full multi-word mention *carcinoma ductal de páncreas extendido al mesenterio*. An inverse kind of boundary detection issue is early start of the predicted BIO tagging such as *Infiltración del peritoneo parietal por adenocarcinoma* being tagged instead of the actual tumor mention *adenocarcinoma*.

Furthermore, in some cases two multi-word tumor mentions got grouped together (often along with a few non-mention words) and were tagged as one single multi-word mention by the prediction.

Table 6
Development data (Run-1) Confusion Matrices

Predicted \ Actual	Development-1			Development-2		
	B	I	O	B	I	O
B	2,813	98	252	2,328	78	217
I	74	2,942	760	77	2,560	635
O	371	1,282	208,309	202	892	168,441

Hence proper phrase identification must be in focus to avoid these kinds of incorrect entity boundaries. It is also possible that there is a need to add more features to the CRF-based models in order to extract more mentions and try to minimise category miss-classification.

6. Conclusion and Future Work

Three Conditional Random Fields classifiers and a Deep Learning approach (a bi-LSTM-CNN combination) were trained and tested on the task of identifying tumor mentions in Spanish medical texts. The best performance was obtained with an ensemble model using majority voting to combine two of the CRF learners and the bi-LSTM-CNN model.

Overall the differences between the models were small in terms of recall, while the deep learner struggled somewhat compared to the CRF-based models in terms of precision. On the unseen test data, however, the bi-LSTM-CNN network showed slightly better recall than the other individual models, although still being outperformed by the voting ensemble.

To improve on the results, it would make potentially be good to incorporate other features such as part-of-speech tags and to utilise tools for noun phrase identification and chunking, at least in the CRF-based models. The deep learners could benefit from having access to word embeddings specifically pre-trained for the clinical domain. The machine learning models could also be improved by applying feature selection and hyper-parameter optimisation based on an evolutionary approach, such as Genetic Algorithms. Finally, more and other types of models could be generated using other classification algorithms, alternative neural network setups, and ensemble models with weighted voting approaches.

References

- [1] Z. Yang, R. Salakhutdinov, W. W. Cohen, Multi-task cross-lingual sequence tagging from scratch, *CoRR abs/1603.06270* (2016). URL: <http://arxiv.org/abs/1603.06270>.
- [2] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1064–1074. URL: <https://www.aclweb.org/anthology/P16-1101>. doi:10.18653/v1/P16-1101.
- [3] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, 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*, Association for Computational Linguistics, San Diego, CA, USA, 2016, pp. 260–270.
- [4] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural computation* 9 (1997) 1735–1780.
- [5] J. Lafferty, A. McCallum, F. C. Pereira, Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data, in: *Proceedings of the 18th International Conference on Machine Learning*, Morgan Kaufmann, Williamstown, MA, USA, 2001, pp. 282–289.
- [6] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147. URL: <https://www.aclweb.org/anthology/W03-0419>.
- [7] R. Chalapathy, E. Zare Borzeshi, M. Piccardi, Bidirectional LSTM-CRF for clinical concept extraction, in: *Proceedings of the Clinical Natural Language Processing Workshop (Clinical-NLP)*, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 7–12. URL: <https://www.aclweb.org/anthology/W16-4202>.
- [8] O. Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, *Journal of the American Medical Informatics Association* 18 (2011) 552–556. URL: <https://doi.org/10.1136/amiajnl-2011-000203>. doi:10.1136/amiajnl-2011-000203.
- [9] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition, *Bioinformatics* 33 (2017) i37–i48. URL: <https://doi.org/10.1093/bioinformatics/btx228>. doi:10.1093/bioinformatics/btx228.
- [10] J. Straková, M. Straka, J. Hajič, Neural architectures for nested NER through linearization, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5326–5331.
- [11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *CoRR abs/1810.04805* (2018). URL: <http://arxiv.org/abs/1810.04805>.
- [12] A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, M. Auli, Cloze-driven pretraining of self-attention networks, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5363–5372.
- [13] U. K. Sikdar, B. Gambäck, Named entity recognition for Amharic using stack-based deep learning, in: International Conference on Computational Linguistics and Intelligent Text Processing, Springer, Budapest, Hungary, 2017, pp. 276–287.
 - [14] B. Gambäck, U. K. Sikdar, Named entity recognition for Amharic using deep learning, in: 2017 IST-Africa Week Conference (IST-Africa), IEEE, Windhoek, Namibia, 2017, pp. 1–8.
 - [15] U. K. Sikdar, B. Gambäck, A feature-based ensemble approach to recognition of emerging and rare named entities, in: Proceedings of the 3rd Workshop on Noisy User-generated Text, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 177–181. URL: <https://www.aclweb.org/anthology/W17-4424>. doi:10.18653/v1/W17-4424.
 - [16] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the CANTEMIST track for cancer text mining in Spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, Spanish Society for Natural Language Processing, Málaga, Spain, 2020.
 - [17] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, brat: a web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 2012, pp. 102–107.
 - [18] S. Bird, E. Loper, NLTK: The natural language toolkit, in: Proceedings of the ACL Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 214–217.
 - [19] J. P. C. Chiu, E. Nichols, Named entity recognition with bidirectional LSTM-CNNs, CoRR abs/1511.08308 (2015). URL: <https://arxiv.org/abs/1511.08308>.
 - [20] N. Okazaki, CRFsuite: a fast implementation of conditional random fields (CRFs), 2007. URL: <http://www.chokkan.org/software/crfsuite>.
 - [21] D. C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, *Mathematical Programming* 45 (1989) 503–528.
 - [22] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543.