Automated extraction of concepts from patient clinical records is an essential facilitator of clinical research. For this reason, the 2010 i2b2/VA Natural Language Processing Challenges for Clinical Records introduced a concept extraction task aimed at identifying and classifying concepts into predefined categories (i.e., treatments, tests and problems). State-of-the-art concept extraction approaches heavily rely on handcrafted features and domain-specific resources which are hard to collect and define. For this reason, this paper proposes an alternative, streamlined approach: a recurrent neural network (the bidirectional LSTM with CRF decoding) initialized with general-purpose, off-the-shelf word embeddings. The experimental results achieved on the 2010 i2b2/VA reference corpora using the proposed framework outperform all recent methods and ranks closely to the best submission from the original 2010 i2b2/VA challenge.

1 Introduction

Patient clinical records typically contain longitudinal data about patients’ health status, diseases, conducted tests and response to treatments. Analysing such information can prove of immense value not only for clinical practice, but also for the organisation and management of healthcare services. Concept extraction (CE) aims to identify mentions to medical concepts such as problems, test and treatments in clinical records (e.g., discharge summaries and progress reports) and classify them into predefined categories. The concepts in clinical records are often expressed with unstructured, “free” text, making their automatic extraction a challenging task for clinical Natural Language Processing (NLP) systems. Traditional approaches have extensively relied on rule-based systems and lexicons to recognise the concepts of interest. Typically, the concepts represent drug names, anatomical nomenclature and other specialized names and phrases which are not part of everyday vocabularies. For instance, “resp status” should be interpreted as “response status”. Such use of abbreviated phrases and acronyms is very common within the medical community, with many abbreviations having a specific meaning that differ from that of other lexicons. Dictionary-based systems perform concept extraction by looking up terms on medical ontologies such as the Unified Medical Language System (UMLS) (Kipper-Schuler et al., 2008). Intrinsically, dictionary- and rule-based systems are laborious to implement and inflexible to new cases and misspellings (Liu et al., 2015). Although these systems can achieve high precision, they tend to suffer from low recall (i.e., they may miss a significant number of concepts). To overcome these limitations, various machine learning approaches have been proposed (e.g., conditional random fields (CRFs), maximum-entropy classifiers and support vector machines) to simultaneously exploit the textual and contextual information while reducing the reliance on lexicon lookup (Lafferty et al., 2001; Berger et al., 1996; Joachims, 1998). State-of-the-art machine learning approaches usually follow a two-step process of feature engineering and classification. The feature engineering task is, in its own right, very laborious and demanding on expert knowledge, and it can become the bottleneck of the overall approach. For this reason, this paper proposes a highly streamlined alternative: to employ a contemporary neural network - the bidirectional LSTM-CRF - initialized with general-purpose, off-the-shelf word embeddings such
as GloVe (Pennington et al., 2014a) and Word2Vec (Mikolov et al., 2013b). The experimental results over the authoritative 2010 i2b2/VA benchmark show that the proposed approach outperforms all recent approaches and ranks closely to the best from the literature.

2 Related Work

Most of the research to date has framed CE as a specialized case of named-entity recognition (NER) and employed a number of supervised and semi-supervised machine learning algorithms with domain-dependent attributes and text features (Uzuner et al., 2011). Hybrid models obtained by cascading CRF and SVM classifiers along with several pattern-matching rules have shown to produce effective results (Boag et al., 2015). Moreover, (Fu and Ananiadou, 2014) have given evidence to the importance of including preprocessing steps such as truecasing and annotation combination. The system that has reported the highest accuracy on the 2010 i2b2/VA concept extraction benchmark is based on unsupervised feature representations obtained by Brown clustering and a hidden semi-Markov model as classifier (de-Bruijn et al., 2011). However, the use of a “hard” clustering technique such as Brown clustering is not suitable for capturing multiple relations between the words and the concepts. For this reason, Jonnalagadda et al. (Jonnalagadda et al., 2012) demonstrated that a random indexing model with distributed word representations can improve clinical concept extraction. Moreover, Wu et al. (Wu et al., 2015) have jointly used word embeddings derived from the entire English Wikipedia (Collobert et al., 2011) and binarized word embeddings derived from domain-specific corpora (e.g. the MIMIC-II corpus (Saeed et al., 2011)). In the broader field of machine learning, the recent years have witnessed a proliferation of deep neural networks, with outstanding results in tasks as diverse as visual, speech and named-entity recognition (Hinton et al., 2012; Krizhevsky et al., 2012; Lample et al., 2016). One of the main advantages of neural networks over traditional approaches is that they can learn the feature representations automatically from the data, thus avoiding the expensive feature-engineering stage. Given the promising performance of deep neural networks and the recent success of unsupervised word embeddings in general NLP tasks (Pennington et al., 2014a; Mikolov et al., 2013b; Lebret and Collobert, 2014), this paper sets to explore the use of a state-of-the-art deep sequential model initialized with general-purpose word embeddings for a task of clinical concept extraction.

3 The Proposed Approach

CE can be formulated as a joint segmentation and classification task over a predefined set of classes. As an example, consider the input sentence provided in Table 1. The notation follows the widely adopted in/out/begin (IOB) entity representation with, in this instance, HCT as the test and 2U PRBC as the treatment. In this paper, we approach the CE task by the bidirectional LSTM-CRF framework where each word in the input sentence is first mapped to either a random vector or a vector from a word embedding. We therefore provide a brief description of both word embeddings and the model hereafter.

Word embeddings are vector representations of natural language words that aim to preserve the semantic and syntactic similarities between them. The vector representations can be generated by either count-based approaches such as Hellinger-PCA (Lebret and Collobert, 2014) or trained models such as Word2Vec (including skip-grams and continuous-bag-of-words) and GloVe trained over large, unsupervised corpora of general-nature documents. In its embedded representation, each word in a text is represented by a real-valued vector, $x$, of arbitrary dimensionality, $d$.

Recurrent neural networks (RNNs) are a family of neural networks that operate on sequential data. They take as input a sequence of vectors $(x_1, x_2, ..., x_n)$ and output a sequence of class posterior probabilities, $(y_1, y_2, ..., y_n)$. An intermediate layer of hidden nodes, $(h_1, h_2, ..., h_n)$, is also part of the model.
In an RNN, the value of the hidden node at time $t$, $h_t$, depends on both the current input, $x_t$, and the previous hidden node, $h_{t-1}$. This recurrent connection from the past timeframe enables a form of short-term memory and makes the RNNs suitable for the prediction of sequences. Formally, the value of a hidden node is described as:

$$ h_t = f(U \cdot x_t + V \cdot h_{t-1}) $$

where $U$ and $V$ are trained weight matrices between the input and the hidden layer, and between the past and current hidden layers, respectively. Function $f(\cdot)$ is the sigmoid function, $f(x) = 1/1 + e^{-x}$, that adds non-linearity to the layer. Eventually, $h(t)$ is input into the output layer and convolved with the output weight matrix, $W$:

$$ y_t = g(W \cdot h_t), \text{ with } g(z_m) = \frac{e^{z_m}}{\sum_{k=1}^{K} e^{z_k}} $$

Eventually, the output is normalized by a multi-class logistic function, $g(\cdot)$, to become a proper probability over the class set. Therefore, the output dimensionality is equal to the number of concept classes. Although an RNN can, in theory, learn long-term dependencies, in practice it tends to be biased towards its most recent inputs. For this reason, the Long Short-Term Memory (LSTM) network incorporates an additional “gated” memory cell that can store long-range dependencies (Hochreiter and Schmidhuber, 1997). In its bidirectional version, the LSTM computes both a forward, $\overrightarrow{h_t}$, and a backward, $\overleftarrow{h_t}$, hidden representation at each timeframe $t$. The final representation is created by concatenating them as $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. In all these networks, the hidden layer can be regarded as an implicit, learned feature that enables concept prediction. A further improvement to this model is provided by performing joint decoding of the entire input sequence in a Viterbi-style manner using a CRF (Lafferty et al., 2001) as the final output layer. The resulting network is commonly referred to as the bidirectional LSTM-CRF (Lample et al., 2016).

### 4 Experiments

#### 4.1 Dataset

The 2010 i2b2/VA Natural Language Processing Challenges for Clinical Records include a concept extraction task focused on the extraction of medical concepts from patient reports. For the challenge, a total of 394 concept-annotated reports for training, 477 for testing, and 877 unannotated reports were de-identified and released to the participants alongside a data use agreement (Uzuner et al., 2011). However, part of this data set is no longer being distributed due to restrictions later introduced by the Institutional Review Board (IRB). Thus, Table 2 summarizes the basic statistics of the training and test data sets which are currently publicly available and that we have used in our experiments.

#### 4.2 Evaluation Methodology

Our models have been blindly evaluated on the 2010 i2b2/VA CE test data using a strict evaluation criterion requiring the predicted concepts to exactly match the annotated concepts in terms of both boundary
Table 3: Performance comparison between the bidirectional LSTM-CRF (bottom three lines) and state-of-the-art systems (top five lines) over the 2010 i2b2/VA concept extraction task.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hidden semi-Markov Model (deBruijn et al., 2011)</td>
<td>86.88</td>
<td>83.64</td>
<td>85.23</td>
</tr>
<tr>
<td>Distributional Semantics CRF (Jonnalagadda et al., 2012)</td>
<td>85.60</td>
<td>82.00</td>
<td>83.70</td>
</tr>
<tr>
<td>Binarized Neural Embedding CRF (Wu et al., 2015)</td>
<td>85.10</td>
<td>80.60</td>
<td>82.80</td>
</tr>
<tr>
<td>CliNER (Boag et al., 2015)</td>
<td>79.50</td>
<td>81.20</td>
<td>80.00</td>
</tr>
<tr>
<td>Truecasing CRFSuite (Fu and Ananiadou, 2014)</td>
<td>80.83</td>
<td>71.47</td>
<td>75.86</td>
</tr>
<tr>
<td>Random - Bidirectional LSTM-CRF</td>
<td>81.06</td>
<td>75.40</td>
<td>78.13</td>
</tr>
<tr>
<td>Word2Vec - Bidirectional LSTM-CRF</td>
<td>82.61</td>
<td>80.03</td>
<td>81.30</td>
</tr>
<tr>
<td>GloVe - Bidirectional LSTM-CRF</td>
<td>84.36</td>
<td>83.41</td>
<td>83.88</td>
</tr>
</tbody>
</table>

Table 3 shows the performance comparison between state-of-the-art CE systems and the proposed bidirectional LSTM-CRF with different initialization strategies. As a first note, the bidirectional LSTM-CRF initialized with GloVe outperforms all recent approaches (2012-2015). On the other hand, the best submission from the 2010 i2b2/VA challenge (deBruijn et al., 2011) still outperforms our approach. However, based on the description provided in (Uzuner et al., 2011), these results are not directly comparable since the experiments in (deBruijn et al., 2011; Jonnalagadda et al., 2012) had used the original dataset which has a significantly larger number of training samples. Using general-purpose, pre-trained embeddings improves the F1 score by over 5 percentage points over a random initialization. In general, the results achieved with the proposed approach are close and in many cases above the results achieved by systems based on hand-engineered features.

### Results and Analysis

4.3 Results and Analysis

This paper has explored the effectiveness of the contemporary bidirectional LSTM-CRF for clinical concept extraction. The most appealing feature of this approach is its ability to provide end-to-end recognition using general-purpose, off-the-shelf word embeddings, thus sparing effort from time-consuming feature construction. The experimental results over the authoritative 2010 i2b2/VA reference corpora look promising, with the bidirectional LSTM-CRF outperforming all recent approaches and ranking closely to the best submission from the original 2010 i2b2/VA challenge. A potential way to further improve its performance would be to explore the use of unsupervised word embeddings trained from domain-specific resources such as the MIMIC-III corpora (Johnson et al., 2016).

### Conclusion

This paper has explored the effectiveness of the contemporary bidirectional LSTM-CRF for clinical concept extraction. The most appealing feature of this approach is its ability to provide end-to-end recognition using general-purpose, off-the-shelf word embeddings, thus sparing effort from time-consuming feature construction. The experimental results over the authoritative 2010 i2b2/VA reference corpora look promising, with the bidirectional LSTM-CRF outperforming all recent approaches and ranking closely to the best submission from the original 2010 i2b2/VA challenge. A potential way to further improve its performance would be to explore the use of unsupervised word embeddings trained from domain-specific resources such as the MIMIC-III corpora (Johnson et al., 2016).

1https://github.com/raghavchalapathy/Bidirectional-LSTM-CRF-for-Clinical-Concept-Extraction
References


Xiao Fu and Sophia Ananiadou. 2014. Improving the extraction of clinical concepts from clinical records. *Proceedings of the 4th Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTExtM04)*.


Rémi Lebret and Ronan Collobert. 2014. Word embeddings through hellinger PCA. In *European Chapter of the Association for Computational Linguistics (EACL)*.


