Faculteit Letteren en Wijsbegeerte
Masterscriptie
Master in de Taalkunde: Computationele Psycholinguïstiek

Extracting Drug, Reason, and Duration Mentions from Clinical Text Data

A Comparison of Approaches

Jens Lemmens

Promotor: Prof. Dr. Walter Daelemans
Copromotor: Dr. Simon Ŝuster
Assessor: Madhumita Sushil

Universiteit Antwerpen
Academiejaar 2018-2019

Ondergetekende, Jens Lemmens, student Taalkunde met profiel Computationele Psycholinguïstiek, verklaart dat deze scriptie volledig oorspronkelijk en uitsluitend door hemzelf geschreven is. Bij alle informatie en ideeën ontleend aan andere bronnen, heeft ondergetekende expliciet en in detail verwezen naar de vindplaatsen.

The undersigned, Jens Lemmens, student in Linguistics with profile Computational Psycholinguistics, declares that this thesis is entirely original and written exclusively by himself. For all information and ideas derived from other sources, the undersigned has explicitly and in detail referred to the locations.

2000 Antwerpen, 23/08/2019

Lemmens

Acknowledgements

I would like to take this opportunity to thank my supervisor, Prof. Dr. Walter Daelemans, for guiding me through the writing process, and proofreading my thesis. Secondly, I want to express my gratitude towards my second supervisor, Dr. Simon Suster, for the weekly sessions in the first semester, the programming help in the second semester, and for answering my many questions. Thirdly, I wish to thank Madhumita Sushil for being my third reader and assessor. Finally, I thank the Computational Linguistics and Psycholinguistics Research Center (CLiPS), for lending me a laptop, and letting me make use of their servers.

Extracting Drug, Reason, and Duration Mentions from Clinical Text Data

A Comparison of Approaches

Jens Lemmens
University of Antwerp,
Faculty of Arts,
Master of Linguistics,
Jens.Lemmens2@student.uantwerpen.be

The following study[1] compares different approaches to the extraction of drug, reason, and duration mentions from unstructured patient discharge summaries. The two latter drug-related pieces of information proved to be challenging to extract in the i2b2 2009 medication extraction challenge. Nevertheless, this information is vital for the patient's medical safety, especially since errors are frequently made when transferring this information from one doctor, hospital, or computer program to another. Further, it can be observed that more 'traditional' rule-based approaches are being replaced with neural approaches in recent studies. Therefore, it is the aim of this work to compare the performance of a rule-based system (MedEx-UIMA) with two retrained neural architectures (BLSTM-CNN and BLSTM-CRF) in order to shed light on how medication, reason, and duration mentions can be extracted more efficiently. The results show that, overall, the neural models outperform the rule-based model on all three named entities. The medication scores of all three systems were comparable to the medication scores of the winner of the i2b2 challenge. Duration mentions were extracted more accurately by the neural models than in the i2b2 challenge, because of the extra training data that was provided in this study, but remained more challenging than drug names. MedEx-UIMA did not exceed i2b2 scores for duration mentions, because of low recall. Finally, no model yielded F-scores above .50 for reason mentions, suggesting that they are intrinsically complex and require a form of world knowledge to be extracted accurately. However, further research must be conducted in order to determine whether large data sets can increase the performance of data-driven models on reason mentions.

1. Introduction

The goal of the present study is to compare three different systems that automatically extract drug, reason, and duration mentions from English patient discharge summaries. The first of these systems is MedEx-UIMA, which is a rule-based system used in professional environments. The two other systems were originally developed to extract information from news data, but are retrained on clinical data in this study. They both utilize bidirectional long-short-term memory (BLSTM): The first of those systems uses BLSTM in conjunction with convolutional neural networks (CNN), and the other system uses BLSTM in conjunction with conditional random fields (CRF). The reason why this

[1] https://github.com/JensLemmens/Thesis

comparative study is relevant is twofold. First, previous research discussed in Section 2 has indicated that the reason why a drug is prescribed ("reason"), and the period of time that this drug needs to be consumed ("duration") are the most challenging pieces of drug-related information to extract from unstructured clinical text data, whereas other medication-related information such as dosages (e.g. "1000 mg") and frequencies (e.g. "3x/day") can be extracted with relatively high accuracy (Uzuner, Solti, and Cadag 2010). Medication-related information is arguably one of the most important forms of clinical data, since it is critical for the patient's health care, medical safety, and overall well-being. Nevertheless, errors are frequently made when drug-related information is transferred from one doctor, hospital, or clinical computer program to another (Xu et al. 2010). Furthermore, clinical data is often written in narrative form, which makes it even more challenging to extract drug-related information automatically. Therefore, more research into how this information can be extracted accurately is needed.

A second reason why this comparative study is relevant is the fact that "traditional" rule-based and hybrid clinical information extraction models (the latter consisting of both knowledge-based and data-driven techniques) are being replaced with supervised and neural approaches in more recent studies (Wei et al. 2019). The reason for this paradigm shift is that it was believed that some form of knowledge structure was required to extract complex clinical information from unstructured text data, but further experimenting with data-driven techniques has shown that non-rule-based techniques can also contribute to the performance of clinical information extraction models (Dalianis 2018). Therefore, it is the aim of this work to compare the performance of one of these more "traditional" rule-based approaches with two more "modern" neural approaches on the automatic extraction of drug-related reason and duration mentions in order to shed light on how these valuable pieces of information can be extracted more accurately. Since all three systems achieved F-scores close to .90 in previous studies (Doan et al. 2010; Hofer et al.; Lample et al. 2016), it can be expected that they will perform (approximately) equally well in the experiments conducted later in this work.

The present study is structured as follows: In Section 2, related research is discussed. This includes a selection of recent medication extraction studies (2.1), the i2b2 medication challenge (2.2), MedEx-UIMA (2.3), and the two neural models that will be used later in this study (2.4). Further, Section 3 outlines the methodology that was employed in this work, i.e. a description of the data (3.1), the experiments that were conducted (3.2), and the evaluation metrics that were used to measure the performance of the aforementioned systems (3.3). In Section 4, the results of MedEx-UIMA (4.1), the BLSTM-CNN model (4.2), and the BLSTM-CRF model (4.3) are published and discussed. Finally, Section 5 concludes the study with a summary of the most relevant findings and some further remarks and suggestions.

2. Related research

What follows is a discussion of related research that will function as a theoretical background for the remainder of this work. First, a number of studies are summarized in order to provide a brief overview of what subjects have been studied and what techniques have been utilized in the domain of medication extraction in recent years. Then, the i2b2 medication challenge is outlined, since this challenge indicated that duration and reason mentions are the most challenging medication-related pieces of information to extract. Further, the knowledge-based MedEx-UIMA system, the BLSTM-CNN model, and the BLSTM-CRF model are described, since these systems will be retrained and/or tested on the i2b2 2009 data later in this study.

2.1 Overview of medication extraction studies

The task of automatically extracting drug-related information belongs to the field of clinical information extraction (CIE). CIE in general has made substantial contributions to the clinical and (bio)medical field by studying, amongst other things, diseases and related symptoms, drugs and related information, and clinical workflow optimization (Wang et al. 2018). More specifically, most studies that have been carried out in the field of medication extraction in recent years have focused on determining the success rate of a new publicly available drug as a form of post-marketing surveillance, adverse drug reactions, dosage extraction, and drug-drug interactions, which is a term for the way different types of drugs interact when they are consumed simultaneously (Dalianis 2018).

The most frequently used machine-learning techniques in these studies were support vector machines (SVM) and conditional random fields (CRF), whereas the majority of the rule-based systems used regular expressions and dictionary look-up (Wei et al. 2019). Further, a third approach exists, which uses a combination of rule-based and data-driven techniques. This type of model is typically called a "hybrid" model. Early systems, e.g. (Xu et al. 2010), mainly utilized rule-based approaches for CIE tasks (Wei et al. 2019). One of the reasons for the usage of rule-based techniques is the presumption that a knowledge structure, such as an ontology or a set of rules developed by domain experts, is required for efficient automatic extraction of complex medical concepts (Dalianis 2018). However, further experimenting with hybrid approaches, e.g. (Patrick and Li 2010), and data-driven approaches, e.g. (Tao, Filannino, and Uzuner 2018), indicated that non-rule-based techniques can also contribute to the performance of CIE systems.

In recent years, great attention has been spent to neural approaches for CIE tasks (Wei et al. 2019). Studies devoted to the development of neural models have achieved state-of-the-art scores in their respective fields and have focused on, for instance, brain tumor segmentation, extracting clinical named entities, and extracting drug-related information (Wu et al. 2015; Li and Huang 2016; Hofer et al.; Hu et al. 2016). Two of these recent neural models (BLSTM-CNN and BLSTM-CRF) are discussed in Section 2.4, and retrained and tested in Section 3.2 order to evaluate their performance on medication, duration and reason extraction. They will then be compared to the rule-based MedEx-UIMA in Section 4.4.

2.2 i2b2 2009 challenge

As mentioned, this study focuses on duration and reason mentions in response to the results of the i2b2 2009 shared task challenge. The goal of this challenge was to develop a system that could extract all medication names and related information from deidentified patient discharge summaries (Uzuner, Solti, and Cadag 2010). All drugs of which the patient was, is or will be the receiver were considered as relevant. These included prescription substances (e.g. Lasix), over the counter medication (e.g. vitamin E), and biological substances required or suggested by doctors (e.g. blood for transfusion), but excluded entities referring to food, water, alcohol, tobacco or illegal drugs (Uzuner et al. 2010b).

For each medication mention, the following attributes had to be extracted (if present): dosage, mode of administration, frequency, duration, reason, event, temporal marker, certainty marker, and text type. The results of the participating teams suggested that especially the attributes "duration" and "reason" were challenging to extract

(Uzuner, Solti, and Cadag 2010). These attributes were defined as follows: "Duration" expresses the period of time during which the medication needs to be consumed. This includes precise temporal expressions (e.g. "for a month"), imprecise temporal expressions (e.g. "temporarily"), cumulative dosages (e.g. "1 pack"), and any expression with a meaning similar to "as needed". "Reason", on the other hand, denotes all problems, such as diseases (e.g. "diabetes") or symptoms (e.g. "headache"), for which the medication is the treatment. For an elaborate discussion of the other attributes, see the i2b2 annotation guidelines (Uzuner et al. 2010b). The best scores that were produced for durations and reasons were obtained by the University of Manchester (Spasic et al. 2010): Their rule-based model yielded system level phrase match F-scores (see below) of .53 and .46 respectively.

For the evaluation of the participating systems, precision, recall and F-score were calculated for vertical and horizontal dimensions, patient and system level, and phrase and token matches (Uzuner, Solti, and Cadag 2010). "Vertical dimension" means that precision, recall, and F-score are calculated for each attribute separately. In contrast, "horizontal dimension" means that these scores are obtained by averaging over each attribute. "Patient level" means that precision, recall and F-score are measured for each individual patient discharge summary, and that the global scores are calculated by averaging over all results of each individual record. Conversely, "system level" scores are measured by counting the total number of true positives, false positives and false negatives in all patient discharge summaries. Afterwards, precision, recall and F-score are calculated on the basis of these counts. "Phrase match" means that the machine-annotated medication, duration, or reason mention is identical to the corresponding human-annotated medication, duration, or reason mention. In other words, one true positive, false positive, or false negative instance is equal to one complete drug, duration, or reason mention. On the other hand, "token match" means that a single token in a machine-annotated drug, duration, or reason mention occurs in the corresponding human-annotated entry. In other words, one true positive, false positive or false negative instance is equal to one specific token in a certain drug, duration or reason mention. The following example should clarify the difference between phrase matches and token matches:

1. m="Paracetamol caffeine" 10:0 10:1||du="3 days" 10:3 10:4||r="headache" 10:7 10:7
2. m="Paracetamol" 10:0 10:0||du="for 3 days" 10:2 10:4||r="nm"

When considering the first and second example as gold and machine annotation respectively, there are no phrase matches: None of the gold drug-related entities matches the corresponding human-annotated entity entirely. However, this does not mean that the system did not find the concerned clinical concept. In the examples above, the system found the medication and duration mentions, but did not recognize the same tokens as the human annotators as their respective boundaries. In contrast, when using the token match evaluation metric, one true positive and one false negative can be counted in the medication entries, two true positives and one false positive can be counted in the duration entries, and one false negative can be counted in the reason entries. In other words, calculating scores on token level provides a more representative view into the performance of a system than a phrase level evaluation. Therefore, the token match evaluation metric will be used in the experiments conducted later in this study (see Section 3.3).

From the twenty teams that participated in the i2b2 challenge, ten used a rule-based approach, four used a data-driven approach and six used a hybrid approach. Popular techniques among the rule-based systems were regular expressions, and look-up in medical ontologies (Yang 2010; Deléger, Grouin, and Zweigenbaum 2010; Mork et al. 2010; Tikk and Solt 2010; Xu et al. 2010). In comparison, the most frequently used data-driven techniques were SVM and CRF (Li et al. 2010; Patrick and Li 2010; Tikk and Solt 2010). Table 1 represents an overview of the top 10 best scoring systems, their approaches, and the final horizontal system level F-scores for phrase matches[2]. From this table, it can be derived that rule-based and hybrid systems were not only used more frequently, but were also more efficient than data-driven systems, since only one supervised system reached the top ten. It can be argued that the prime cause for these observations is the fact that i2b2 provided little annotated training data (see Section 3.1 for a detailed description of the data). Presumably, this is not only the reason why only few participants used data-driven systems, but also why these data-driven systems yielded lower scores than the rule-based and hybrid systems. Now that the i2b2 medication challenge has been outlined, and it has become clear why this study focuses on duration and reason mentions, the models that will be used later in this study can be described in the subsequent subsections.

Table 1
Overview of the top 10 systems in the i2b2 challenge, their approaches, and results.

Place	Team	Approach	F-score
1	University of Sydney (AUS)	Hybrid	.87
2	Vanderbilt University (USA)	Rule-based	.82
3	University of Manchester (UK)	Rule-based	.81
4	National Library of Medicine (USA)	Rule-based	.80
5	Humboldt University (GER)	Hybrid	.80
6	Open University (UK)	Rule-based	.80
7	University of Paris (FR)	Rule-based	.78
8	LIMSI (FR)	Rule-based	.77
9	University of Utah (USA)	Hybrid	.77
10	University of Wisconsin-Milwaukee (USA)	Supervised	.76

2.3 MedEx-UIMA

The original MedEx system was developed shortly before the start of the i2b2 medication challenge, but it was designed for the same task: extracting as much medication-related information from unstructured clinical text data as possible (Xu et al. 2010). The system consisted of a sentence splitter, a top-down chart parser accompanied by a context free grammar to parse the sentences, and a semantic tagger which labelled relevant pieces of information with a semantic category. In total, MedEx distinguished 11 semantic categories (including drug names and duration mentions). Surprisingly, "reason" or any other category with a similar meaning was not extracted by the original

[2] Phrase instead of token match evaluation metrics were used here, since token match scores were not publicly available for all systems.

system. However, the i2b2 MedEx model was designed to extract reason mentions in addition to the other semantic categories for the purpose of the challenge (Doan et al. 2010). It yielded system level token match F-scores of .89 for drug mentions, .37 for reason mentions, and .41 for duration mentions. With these scores, MedEx finished second in the challenge and was the best performing rule-based model in the challenge.

In 2014, MedEx-UIMA was developed. As its name suggests, MedEx-UIMA is a version of the original MedEx system implemented in an Unstructured Information Management Application (UIMA[3]) (Jiang et al. 2014). The goal of this system was not only to extract as much medication information as possible, but also to directly link all medication mentions with RxNorm[4] concepts for standardization purposes. To achieve this goal, a normalization module, which maps drug names to RxNorm concepts, was added to the system. In addition, a Drools[5] rule engine was included. In MedEx-UIMA, the user can decide themself whether to use the built-in rules or the Drools rules. Note that MedEx-UIMA was based on the original MedEx, which means that MedEx-UIMA does not extract reason mentions either. However, the reason scores that were obtained by the i2b2 model can be used as a baseline for the experiments that will be conducted with the other models later in this study.

The performance of MedEx-UIMA was measured by processing 826 clinical notes with the original MedEx system and the MedEx-UIMA system. When using the MedEx-annotated notes as gold standards, MedEx-UIMA yielded an F-score of .97 (averaged across all semantic categories). After the calculation of this score, 100 randomly chosen mismatching drug mentions were manually evaluated, and 58% of those mentions were processed more accurately by MedEx-UIMA, suggesting that it performed slightly better than the original MedEx system (Jiang et al. 2014). The MedEx-UIMA version that will be used later in this study is the 2015 version, which can be downloaded for free after online registration[6].

2.4 Neural approaches

As mentioned in Section 2.1, various recent CIE studies have focused on neural approaches. In a 2016 study, a model consisting of bidirectional long-short term memory (BLSTM) and convolutional neural networks (CNNs) was developed to extract person names, organization names, location names, and miscellaneous entities from news data (Chiu and Nichols 2016). The model was trained on CoNLL 2003[7] and OntoNotes 5.0[8] data, and used character and word embeddings as core representations. Further, capitalization features and BIOES-tagging (Beginning, Inside, Outside, End, Single) were also included in the model. The latter is a tagging scheme in the form of X-Y,

[3] Software written in Java that is capable of analyzing large volumes of unstructured data. https://uima.apache.org/

[4] RxNorm is a tool created by the Unified Medical Language System (UMLS) which normalizes generic and branded drug names and related terminology. The purpose of RxNorm is to optimize clinical workflow and minimize communication errors caused by ambiguous terminology. https://www.nlm.nih.gov/research/umls/rxnorm/index.html

[5] Drools is an application in which medical experts can represent domain knowledge in the form of rules. These rules are then implemented into a model in order to make tagging decisions. https://www.drools.org

[6] https://sbmi.uth.edu/ccb/resources/medex.htm

[7] Conference of Natural Language Learning shared task which included automatic extraction of named entities from news data.

[8] Corpora containing English, Arabic and Chinese text data retrieved from a variety of sources, such as telephone conversations, broadcast news, web-blogs, etc.

where X represents the position of the token relative to the phrase it belongs to, and Y represents the type of named entity the token belongs to (i.e. "m" for "medication", "du" for "duration", or "r" for "reason" in this study). X can take the form of any letter in "BIOES": "B" means that the relevant token is the first token in the phrase and that the phrase is at least two tokens long. "I" means that the token is not the first or last token in the phrase and that the phrase is at least three tokens long. "E" means that the token is the last token in the phrase and that the phrase is at least two tokens long. "S" means that the relevant token is the only token in the phrase. Finally, "O" means that the token is not part of a named entity that the model is expected to extract. What follows is an example sentence that was tagged with BIOES tags on drugs, durations, and reasons to clarify how the tagging system works:

> "Patient" (O) "was" (O) "prescribed" (O) Paracetamol" (S-m) "for" (B-du) "three" (I-du) "days," (E-du) "because" (O) "of" (O) "severe" (B-r) "headache." (E-r)

Finally, mini-batch stochastic gradient descent was used for learning. The BLSTM-CNN model was tested on the CoNLL 2003 and OntoNotes 5.0 data sets, and yielded F-scores of .92 and .86 respectively. This makes the system competitive with the best scoring system on the CoNLL 2003 data set and more efficient than the best scoring system on the OntoNotes data set that was previously reported (Chiu and Nichols 2016).

In recent research[9], the BLSTM-CNN model was retrained on the i2b2 2009 data set for the extraction of the same information that the models in the i2b2 challenge had to extract (Hofer et al.). First, the i2b2 data was parsed to CoNLL 2003 format. For this task, the data was labelled with part-of-speech (POS) tags, and with BIO tags, where "E" is replaced with "I", and "S" is replaced with "B" compared to BIOES tagging. Finally, the data was split into a training, development, and test set consisting of 70%, 15%, and 15% of the data respectively. The retrained model yielded an horizontal system level F-score for token matches of .86 on the test set, which is comparable to the results of the winning team of the challenge. The author of this study challenged his readers to calculate the scores for each attribute, and optimize the model for the attributes that are most difficult to extract, i.e. "reason" and "duration", which is exactly what will be done later in this work.

In a study similar to the work above, a BLSTM-CRF model was designed for the extraction of information from news data (Lample et al. 2016). The purpose of the study was to develop a model that could perform accurate information extraction from news data without the use of large amounts of annotated data sets, which are often expensive and time-consuming to create (Velupillai et al. 2015), without language specific knowledge resources, and without the need for intensive feature engineering. The model was trained on CoNLL 2003 news data, and used both token and character representations (Lample et al. 2016). Each word in each sentence was tagged with the BIOES tagging scheme, and training was done with stochastic gradient descent (SGD). Conditional random fields were used to model dependant tagging decisions in the BIOES tagging system (e.g. "B-du" cannot precede "I-r"). The model achieved a state-of-the-art F-score of .91 on the CoNLL 2003 test set, which makes it competitive with the BLSTM-CNN model. However, unlike the BLSTM-CNN model, the BLSTM-CRF model has not yet been retrained on clinical data before the present study.

[9] https://towardsdatascience.com/deep-learning-for-named-entity-recognition-3-reusing-a-bidirectional-lstm-cnn-on-clinical-text-e84bd28052df

3. Methodology

In this section, the methodology that was employed to achieve insight into the performance of MedEx-UIMA, the BLSTM-CNN model, and the BLSTM-CRF model on the extraction of drug, duration, and reason mentions, is outlined. Section 3.1 describes the data that was used for training, validation, and testing, Section 3.2 explains what experiments are conducted in this study, and finally, Section 3.3 describes the evaluation metrics that are used in these experiments.

3.1 Data

The data that was used in this study was identical to the data used in the i2b2 2009 challenge and was obtained from the i2b2 website[10]. The original data set comprised 1243 deidentified and unstructured patient discharge summaries (Uzuner, Solti, and Cadag 2010). In the 2009 challenge, 696 of these records were released during the development period. Only 17 of those records were annotated. The remaining 547 records, of which 251 were annotated collectively by the participating teams, were held out for testing. In the annotation files, a new entry was created on a separate line for each medication mention. Attributes related to this drug followed on the same line. Consider the following example:

> m="heparin" 66:8 66:8||do="nm"||mo="nm"||f="nm"||du="nm"||r="nm"||
> ln="narrative"

In the example above, "m" stands for "medication name", "do" for "dosage", "mo" for "mode", "f" for "frequency", "du" for "duration", "r" for "reason", and "ln" for "list/narrative". If an attribute is not mentioned, this is indicated with "nm". Attributes are separated by two pipe symbols (||). For each attribute, entry onset (first character of the phrase) is expressed by the line index (1-based), followed by a colon, and the token index (0-based). Entry offset (last character of the phrase) is expressed identically, and follows after entry onset.

Note that i2b2 only provides 10 of the 17 annotated discharge summaries that comprised the training set in the 2009 challenge, which makes it impossible to recreate the original training set. Also note that, as explained above, the original training set provided in the 2009 challenge was relatively small, since it contained only 17 patient discharge summaries. Hence, the data was split as follows: 26 of the 261 documents functioned as test set, 26 documents functioned as validation set[11], and the rest of the documents functioned as training set. The advantage of this new split is that more data is used for training, which could have a positive influence on the scores of reason and duration mentions, since these occur relatively seldom. The disadvantage, however, is that it is harder to compare the results of the systems that participated in the i2b2 challenge with the performance of the systems in this study. In Table 2, the number of medication, reason and duration mentions in the train, development, and test set can be found.

[10] https://www.i2b2.org/NLP/DataSets/Main.php

[11] Although no parameters are optimized in the experiments conducted later in this study, a validation set is required when (re)training the BLSTM-CNN ands BLSTM-CRF model. For the same reason, no cross-validation is performed.

Table 2
Number of medication, reason, and duration mentions in train, development, and test set.

Type	Train	Dev	Test	Total
Medication	7593	893	832	9318
Duration	464	65	42	571
Reason	1368	156	170	1694

For the purpose of this study, code was written to modify the data. What follows is a description of what modifications were made: First, all irrelevant attributes (i.e. all attributes but "medication", "duration", and "reason") were discarded from each entry in each annotated file. This resulted in entries of the following format:

m="heparin" 66:8 66:8||du="nm"||r="nm"

For the experiments conducted with MedEx-UIMA, the reason mentions were also discared from the annotation files, since this model does not extract reason mentions. For the experiments conducted with the neural models, the content of all annotated files was parsed to CoNLL 2003 format with the help of the script[12] that was used to retrain the BLSTM-CNN model in previous research[13]. To convert the data to CoNLL format, three files were created: one for the training set, one for the development set, and one for the test set. For each token in each unstructured patient discharge summary in a particular set (train, development, or test) a new row containing this token was created. A second column contained the token's corresponding IOB-tag, which was retrieved by parsing the human-annotated discharge summaries. These two columns were then concatenated with the help of a Pandas data frame. What follows is an example of a sentence in CoNLL 2003 format in which drugs, durations, and reasons have been tagged with BIO tags:

Patient (O)
took (O)
Paracetamol (B-m)
for (B-du)
5 (I-du)
days (I-du)
because (O)
of (O)
headache. (B-r)

For the BLSTM-CRF model, an empty line was inserted after each line break in the unstructured discharge summaries, since this was required for training. Other columns could be inserted between the token and its IOB-tag such as a column that contains part-of-speech tags, but this was not mandatory and did not influence the results.

[12] https://github.com/mxhofer/i2b2_2009–to–CoNLL
[13] https://towardsdatascience.com/deep-learning-for-named-entity-recognition-3-reusing-a-bidirectional-lstm-cnn-on-clinical-text-e84bd28052df

3.2 Experiments

In the first experiment of this study, the performance of MedEx-UIMA was tested on the test set described above. This experiment was not only conducted with the built-in rule set, but also with the Drools rule engine.

In the second experiment, the BLSTM-CNN model was retrained, validated and tested on the train, validation, and test sets that were created above with the i2b2 2009 data. The parameter settings that were recommended in previous experiments[14] were utilized in this experiment: the BIO tagging scheme, 30 epochs, a dropout rate of .5, an LSTM state size of 200, a learning rate of .0105, a convolutional width of 3, and a Nadam optimizer for learning. Training was done with the help of Keras and Tensorflow.

In the third experiment, the BLSTM-CRF model was retrained and tested identically to the BLSTM-CNN model with its default parameter settings: 100 epochs, a character embedding dimension of 25, a token embedding dimension of 100, a dropout rate of .5, and stochastic gradient descent (SGD) for learning. For the sake of comparability, the same training data was used in this experiment as in Experiment 2. Therefore, BIO instead of the default BIOES tagging scheme was used. Training was done with the help of Theano.

3.3 Evaluation Metrics

To determine the performance of MedEx-UIMA, a script that calculates system level token match F-scores for drug and duration mentions was written. The script also calculated the micro-average F-score on token level. As mentioned in Section 2.2, token level evaluation was chosen as evaluation metric, because this metric provides a more representative view of the performance of a model than phrase level evaluation. In the evaluation script, the MedEx-UIMA output is first converted to i2b2 format. Then, for each drug/duration mention in each machine-annotated file, it is checked whether there is a drug/duration mention in the human-annotated file that occurs on the same line, and that has at least one token index in common. If these conditions are true, the number of true positives, false positives, and false negatives are counted as described above. If no corresponding drug/duration mention was found, the number of tokens in the machine-annotated entity are counted as false positives. Finally, for each drug/duration mention in each human-annotated file, it is checked whether this particular phrase was (partially) found by MedEx-UIMA. If not, the number of tokens in that phrase were counted as false negatives.

For the neural models, the default scoring methods that were included in the code were used for evaluation. Since both systems require input data in CoNLL 2003 format, this means that true positives, false positives and false negative were counted on system level for token matches: If a predicted label matched the true label, this was seen as a true positive. Conversely, if a predicted label (e.g. "B-m") did not match the true label (e.g. "B-du"), this was seen as a false positive case for the predicted label ("m") and as a false negative case for the true label ("du"). Both the scores for each individual attribute (vertical dimension) and the micro average scores (horizontal dimension) were included. In other words, this scoring method matches the MedEx-UIMA scoring method.

[14] https://towardsdatascience.com/deep-learning-for-named-entity-recognition-3-reusing-a-bidirectional-lstm-cnn-on-clinical-text-e84bd28052df

4. Results and Discussion

What follows is a quantitative and qualitative discussion of the results of the experiments described in Section 3.2. Quantitative means that the performance of the models will be expressed in the scores described in Section 3.3, whereas qualitative means the type of errors that were made by the models are discussed. For MedEx-UIMA, this means that a selection of errors that were made will be discussed in more detail. For the BLSTM-CNN and BLSTM-CRF models, this means that their respective confusion matrices will be studied.

4.1 Experiment 1

In Table 3 and 4, the results of MedEx-UIMA with the built-in rules and the Drools rule engine can be found respectively. First, it can be observed that MedEx-UIMA performed better with the Drools rule engine than with the built-in rule set for both medication and duration mentions. Secondly, the i2b2 MedEx model performed marginally better on medication mentions than MedEx-UIMA (.89 vs. .87), due to the recall scores that were significantly lower than the precision scores for both the built-in and the Drools rule set. The reason for this is that MedEx-UIMA systematically ignores substances required by doctors such as "insulin", "blood", and "fluid bolus", which were seen as medication mentions in the i2b2 challenge. Moreover, MedEx-UIMA tends to omit tokens in medication entries that contain multiple tokens, e.g. "vitamin" instead of "vitamin C", "cyclosporine" instead of "cyclosporine micromeral", and "KCL" instead of "KCL slow release".

Further, the scores for duration mentions were, as expected, significantly lower than the scores for medication mentions. Interestingly, the performance of MedEx-UIMA on duration mentions was significantly lower in this study than in the i2b2 challenge. More specifically, the precision score for duration mentions was relatively high compared to other systems in the i2b2 challenge, but recall remained low, which is why the overall F-score was lower than in the i2b2 challenge. The reason for this low recall is that some phrases that i2b2 would categorize as duration mentions were often categorized as frequencies or dosage amounts by MedEx-UIMA (e.g. cumulative dosages such as "x 30 tablets" or "1 pack"). Further, in sentences such as "The patient took antibiotics for 3 weeks", i2b2 would annotate "for 3 weeks" as a duration mention, whereas MedEx-UIMA would annotate "3 weeks" as duration mention, ignoring the preposition "for". In sum, duration mentions remain challenging to extract for MedEx-UIMA, whereas medication mentions can be extracted without much problems.

Table 3
Results of MedEx-UIMA (built-in rules)

Type	Precision	Recall	F-score
Micro average	.9239	.7579	.8327
Medication	.9317	.8112	.8673
Duration	.5714	.1290	.2105

Table 4
Results of MedEx-UIMA (Drools rule engine)

Type	Precision	Recall	F-score
Micro average	.9200	.7642	.8349
Medication	.9448	.8180	.8768
Duration	.4706	.2258	.3048

4.2 Experiment 2

In Table 5, the quantitative results of Experiment 2 can be found. Precision scores were higher than recall scores for drugs and durations, but not for reasons. Similar to the results of Experiment 1, the scores achieved for duration mentions were significantly lower than the scores achieved for medication mentions. The same can be said about the scores for reason mentions. Nevertheless, the BLSTM-CNN model performed relatively well on the two latter attributes compared to the i2b2 models, especially on duration mentions. The cause for this higher performance could be the fact that more training data was used to train the model and/or the possibility that, regardless of the quantity of training data, the model is intrinsically more robust than the models that participated in the i2b2 challenge. Further, in Table 5 it can be observed that the results for reasons still remained below 50% (.45), whereas the the results for the duration mentions exceeded 60% (.64).

In other words, the model benefited from the extra training data when it comes to durations, whereas reasons remain challenging to extract in spite of the extra training data. The cause for this is presumably the structural and semantic complexity of reason mentions. This means that they can be long in comparison to drug and duration mentions, e.g. "a bacterial urinary infection" vs. "Paracetamol" or "5 days", and that some form of world knowledge is required to recognize the reason why a drug is prescribed. For models that do not implement a certain form of knowledge structure, its performance relies completely on the linguistic features that it was trained on. Therefore, further research has to be carried out in order to determine whether this complex drug-related piece of information can be extracted more accurately by a data-driven model that has been trained on even more data than in this study.

In qualitative terms, the confusion matrix in Table 6 shows that one type of named entity was rarely confused with another type of named entity (e.g. "B-m" was labeled as "B-r"), especially for duration mentions (0 times in Table 6). In other words, if a token that was part of a relevant named entity was tagged with the wrong label, this label was virtually always "O". These type of errors include both cases where the entire named entity was not recognized, but also cases where one token (first or last token of the phrase) was not recognized. A second type of error were cases where the model labeled tokens that were not part of a relevant named entity ("O") as part of a named entity (e.g. "B-m"). Since recall scores were higher than precision scores for both medication and duration mentions, this second type of error occurred more frequently than the first type of error.

Table 5
Results of BLSTM-CNN model

Type	Precision	Recall	F-score
Micro average	.7619	.8477	.8062
Medication	.8187	.9340	.8725
Duration	.6034	.5731	.6364
Reason	.4599	.4433	.4514

Table 6
Confusion matrix BLSTM-CNN model (y-axis=gold, x-axis=machine)

Tag	B-m	I-m	B-r	I-r	B-du	I-du	O
B-m	986	2	0	0	0	0	27
I-m	11	633	0	0	0	0	31
B-r	4	0	108	12	0	0	70
I-r	1	0	10	69	0	0	65
B-du	0	0	0	0	41	0	11
I-du	0	0	0	0	0	127	22
O	156	42	69	58	17	34	31481

4.3 Experiment 3

In Table 7, the results of Experiment 3 can be found. In this table, the same trends can be observed as in the results of Experiment 2. First, the BLSTM-CRF model performed better on medication mentions than on duration and reason mentions. Secondly, scores for duration mentions were significantly higher than scores for reason mentions, and they were also significantly higher than the scores that were achieved for duration mentions in the i2b2 challenge. This supports the hypothesis that reason mentions are intrinsically more challenging to extract than duration mentions. Thirdly, although the results for reason mentions were the lowest, they were still higher than most results obtained in the i2b2 challenge for reason mentions. Therefore, it is still unclear whether large data sets can increase the performance of data-driven models on reason mentions significantly. Finally and in contrast with the BLSTM-CNN model, recall scores were higher than precision scores for drug and duration mentions.

In terms of quality (see confusion matrix in Table 8), the bulk of the errors the BLSTM-CRF model made were cases were a token was part of a named entity, but was not labeled as part of a named entity (e.g. "B-m" was labeled as "O"). A second type of errors were cases where a token that was not part of a named entity was labeled as part of a named entity (e.g. "O" was labeled as "B-m"). Both types of errors include cases where either an entire phrase was not recognized or where only one token (usually the first or the last token of a phrase) was not recognized as part of a named entity (e.g. "3 days" instead of "for 3 days" or "3 times" instead of "3 times daily") or vice versa. As the quantitative analysis suggests, both types of errors were made more frequently for reason than for duration mentions, and more frequently for duration than for medication mentions. Finally, cases where the model mistook a type of named entity for another (e.g. "B-m" labeled as "B-r", 1 instance in Table 8) were relatively rare.

Table 7
Results of the BLSTM-CRF model

Type	Precision	Recall	F-score
Micro average	.8785	.7684	.8198
Medication	.9201	.8621	.8901
Duration	.8000	.5385	.6437
Reason	.5641	.3402	.4244

Table 8
Confusion matrix BLSTM-CRF model (y-axis=gold, x-axis=machine)

Tag	B-m	I-m	B-r	I-r	B-du	I-du	O
B-m	900	8	1	1	0	0	105
I-m	8	630	0	0	0	0	37
B-r	1	0	73	7	0	0	113
I-r	0	1	5	50	0	0	89
B-du	0	0	0	0	29	0	23
I-du	0	0	0	0	2	70	77
O	40	8	38	22	4	10	31678

4.4 General Discussion of the Results

From the results in Table 3, 4, 5, and 7, it can be concluded that the medication scores for all systems were on par with the winning team of the i2b2 2009 challenge (+- .88 F-score). Further, the BLSTM-CRF model extracted the medication (.8901) and duration (.6437) mentions with the highest accuracy of the three models. The BLSTM-CNN model, however, achieved results for duration (.6364) mentions that were only marginally lower, and extracted reason mentions with the highest accuracy of all models (.4514). MedEx-UIMA yielded the best scores with the Drools rule engine, but scored lower on duration mentions (.3048) than the two neural models and obtained virtually the same results for medication (.8768) mentions as the BLSTM-CNN model (.8725). Overall, the BLSTM-CRF model scored the best, but in terms of recall, the BLSTM-CNN model performed better.

In terms of quality, the error analysis of MedEx-UIMA showed that most errors were cases where MedEx-UIMA systematically ignored certain terms, such as "blood" or "insulin", and that the model also omitted tokens (usually the first or last) of a phrase, e.g. "vitamin" instead of "vitamin C" or "three days" instead of "for three days". Similarly, the confusion matrices in Table 6 and Table 8 show that the bulk of errors that were made by the neural models were instances where they omitted (tokens in) phrases. Named entities were generally not mistaken for another type of named entity.

In sum, these results suggest that the scores of duration mentions can be increased drastically by gathering sufficient training data, but also that reason mentions remain challenging to extract, regardless of the amount of training data that is provided. Further, it can be concluded that CIE models that do not incorporate knowledge-based or rule-based techniques can be competitive with systems that rely completely on rules and/or knowledge structures.

5. Conclusion

This study compared the performance of a rule-based system (MedEx-UIMA) and two neural models (BLSTM-CNN and BLSTM-CRF) on the automatic extraction of drug, duration, and reason mentions from English clinical text data. The goal of this study was to shed new light on which approach was the most efficient to extract medication-related information that has proven to be challenging to extract in the i2b2 2009 challenge. This study is relevant because of two reasons: First, accurately extracting medication-related information from patient discharge summaries is essential for the patient's medical safety and overall well-being, especially because communication errors are frequently made when transferring medication-related information from one doctor, hospital, or clinical computer program to another. Secondly, this study sheds light on the strengths and weaknesses of neural clinical information extraction systems, which have been replacing the more traditional rule-based models in recent years.

The results of the conducted experiments indicate that models that do not utilize rule-based or knowledge-based techniques can be competitive with models that rely (completely or partially) on rules and/or knowledge structures. From the results of Experiment 2 and Experiment 3 can also be concluded that the accuracy with which duration mentions are automatically extracted can drastically be improved with collecting sufficient (annotated) training data, whereas reason mentions remain intrinsically challenging to extract.

In other words, after this study, it remains unclear how to extract reason mentions with relatively high accuracy. Therefore, further research must be conducted in order to provide more insights into this problem. More specifically, larger unstructured clinical data sets must be created in order to determine whether the performance of models that automatically extract reason mentions can be improved by providing more training data or not. Further, creating a hybrid model that uses both data-driven and rule-based techniques, and which has been trained on large annotated clinical data sets may provide further insights into the problem of the automatic extraction of the reason why a specific drug needs to be consumed.

References

Chiu, J. P. C. and E. Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Dalianis, H. 2018. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer.

Deléger, L., C. Grouin, and P. Zweigenbaum. 2010. Extracting medical information from narrative patient records: The case of medication-related information. *Journal of the American Medical Informatics Association*, 17(5):555–558.

Doan, S., L. Bastarache, S. Klimkowski, J. C. Denny, and H. Xu. 2010. Integrating existing natural language processing tools for medication extraction from discharge summaries. *Journal of the American Medical Informatics Association*, 17(1):528–531.

Hofer, M., A. Kormilitzin, P. Goldberg, and A. Nevado-Holgado. Few-shot learning for named entity recognition in medical text. *arXiv*, available at: https://arxiv.org/abs/1811.05468.

Hu, K., Q. Gan, Y. Zhang, S. Deng, F. Xiao, W. Huang, C. Cao, and X. Gao. 2016. Brain tumor segmentation using multi-cascaded convolutional neural networks and conditional random field. *IEEE Access*, 4.

Jiang, M., Y. Wu, A. Shah, P. Priyanka, J. C. Denny, and H. Xu. 2014. Extracting and standardizing medication information in clinical text - the MedEx-UIMA system. *Journal of the American Medical Informatics Association*, 17(5):37–42.

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL 2016*, San Diego, California, USA.

Li, P. and H. Huang. 2016. Clinical information extraction via convolutional neural network. *arXiv*, available at:

https://arxiv.org/abs/1603.09381.

Li, Z., F. Liu, L. Antieau, Y. Cao, and H. Yu. 2010. Lancet: A high precision medication event extraction system for clinical text. *Journal of the American Medical Informatics Association*, 17(5):563–567.

Mork, J. G., O. Bodenreider, D. Demner-Fushman, R. I. Dogan, F.-M. Lang, Z. Lu, A. Névéol, L. Peters, S. E. Shooshan, and A. R. Aronson. 2010. Extracting Rx information from clinical narrative. *Journal of the American Medical Informatics Association*, 17(5):536–539.

Patrick, J. and M. Li. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527.

Spasic, I., F. Sarafraz, J. A. Keane, and G. Nenadic. 2010. Medication information extraction with linguistic pattern matching and semantic rules. *Journal of the American Medical Informatics Association*, 17(1):532–535.

Tao, C., M. Filannino, and Ö Uzuner. 2018. FABLE: A semi-supervised prescription information extraction system. In *AMIA Annual Symposium Proceedings Archive*, pages 1534–1543.

Tikk, D. and I. Solt. 2010. Improving textual medication extraction using combined conditional random fields and rule-based systems. *Journal of the American Medical Informatics Association*, 17(5):540–544.

Uzuner, Ö., I. Solti, and E. Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Uzuner, Ö., I. Solti, F. Xia, and E. Cadag. 2010b. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

Velupillai, S., D. Mowery, B. R. South, M. Kvist, and H. Dalianis. 2015. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of Medical Informatics*, 10(1):183–193.

Wang, Y., L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Wei, Q., Z. Ji, Z. Li, J. Du, J. Wang, J. Xu, Y. Xiang, F. Tiryaki, S. Wu, Y. Zhang, C. Tao, and H. Xu. 2019. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, 26(5).

Wu, Y., M. Jiang, J. Lei, and H. Xu. 2015. Named entity recognition in Chinese clinical text using deep neural network. *Studies in Health Technology and Informatics*, 216:624–628.

Xu, H., S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny. 2010. MedEx: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

Yang, H. 2010. Automatic extraction of medication informaton from medical discharge summaries. *Journal of the American Medical Informatics Association*, 17(5):545–548.