TAGWAR: An Annotated Corpus for Sequence Tagging of War Incidents

Nancy Sawaya, Shady Elbassuoni, Fatima K. Abu Salem and Roaa Al Feel
Computer Science Department, American University of Beirut, Lebanon

Keywords: Information Extraction, Event Extraction, Sequence Tagging, War Incidents, Deep Learning.

Abstract: Sequence tagging of free text constitutes an important task in natural language processing (NLP). In this work, we focus on the problem of automatic sequence tagging of news articles reporting on wars. In this context, tags correspond to details surrounding war incidents where a large number of casualties is observed, such as the location of the incident, its date, the cause of death, the actor responsible for the incident, and the number of casualties of different types (civilians, non-civilians, women and children). To this end, we begin by building TAGWAR, a manually sequence tagged dataset consisting of 804 news articles around the Syrian war, and use this dataset to train and test three state-of-the-art, deep learning based, sequence tagging models: BERT, BiLSTM, and a plain Conditional Random Field (CRF) model, with BERT delivering the best performance. Our approach incorporates an element of input sensitivity analysis where we attempt modeling exclusively at the level of articles' titles, versus titles and first paragraph, and finally versus full text. TAGWAR is publicly available at: https://doi.org/10.5281/zenodo.3766682.

1 INTRODUCTION

Extracting events from news articles has become a popular NLP task. A common approach to extract such events is to utilize sequence tagging. Sequence tagging is the process of assigning a tag to each word in a sequence of words. In this paper, we aim to sequence tag news articles, the sort of which appear around military wars, with tags corresponding to war incidents such as the location of the incident, its date, the cause of death, the actor responsible for the incident, and the number of casualties of different types (civilians, non-civilians, women and children). The output is a dataset representing fully tagged articles. Such a dataset can help contribute to further data analysis that aids in the process of fact checking, such as finding the total number of casualties given a particular actor over multiple news articles, finding the number of war incidents of a particular type (e.g. chemical attack, air bombardment, etc.) in a given location within a specified date interval, and so on and so forth.

To this end, we build TAGWAR, a dataset consisting of 804 manually sequence tagged news articles retrieved from FA-KES, a fake news dataset around the Syrian war (Abu Salem et al., 2019). We utilize the BIOE sequence tagging approach where 'B' stands for beginning, 'I' for inside, 'O' for outside, and 'E' for end of an attribute. Thus, in TAGWAR, each word in each news article is tagged with one of the following tags: 'B-LOC', 'I-LOC', 'E-LOC', 'B-CIV', 'I-CIV', 'E-CIV', 'B-NCV', 'I-NCV', 'E-NCV', 'B-WMN', 'I-WMN', 'E-WMN', 'B-CHD', 'I-CHD', 'E-CHD', 'B-ACT', 'I-ACT','E-ACT', 'B-COD', 'I-COD', 'E-COD', 'B-DAT', 'I-DAT', 'E-DAT', or 'O', where each of the acronyms represent the following:

- 'O' designating words outside the scope
- 'LOC' designating the location of incident
- 'CIV' designating the number of civilians dead
- 'NCV' designating the number of non-civilians dead
- 'WMN' designating the number of women dead
- 'CHD' designating the number of children dead
- 'ACT' designating the actor responsible for the incident
- 'COD' designating the cause of death
- 'DAT' designating the date of incident

Table 1 shows a snippet of one annotated news article in TAGWAR. To the best of our knowledge, TAGWAR is the only dataset that contains news articles around the Syrian war that are sequence tagged as described above.

Sawaya, N., Elbassuoni, S., Salem, F. and Al Feel, R.
TAGWAR: An Annotated Corpus for Sequence Tagging of War Incidents.
DOI: 10.5220/0010135202430250
In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 1: KDIR*, pages 243-250
ISBN: 978-989-758-474-9

Table 1: Example Annotated TAGWAR Article.

Word	Tag
Friday	B-DAT
15	I-DAT
Jul	I-DAT
2016	E-DAT
11	B-CIV
civilians	E-CIV
four	B-WMN
women	E-WMN
four	B-CHD
children	E-CHD
Syrian	B-ACT
or	I-ACT
Russian	E-ACT
air	B-COD
raids	E-COD
Deir	B-LOC
Ezzor	E-LOC

To attain the goal of automatic sequence tagging, we then use TAGWAR to train and test three state-of-the-art, deep learning based, sequence tagging models: BERT, BiLSTM, and a plain Conditional Random Field (CRF) model, with BERT delivering the best performance. Our approach incorporates an element of input sensitivity analysis where we attempt modeling exclusively at the level of articles' titles, versus titles and first paragraph, and finally versus full text.

The paper is organized as follows. In Section 2 we review related work in the area of sequence tagging and event extraction mechanisms. In Section 3 we describe how TAGWAR was constructed and sequence tagged. In Section 4 we describe the deep learning models that were trained and tested using TAGWAR, yielding an automatic process for sequence tagging of news articles around the Syrian war. Finally, we conclude and present future directions in Section 5.

2 RELATED WORK

Hamborg et al. (Hamborg et al., 2018) propose Giveme5W, the first open-source, syntax-based 5W extraction system for news articles. Answers to the five journalistic W questions (5Ws) describe the main event of a news article, i.e., who did what, when, where, and why. The system retrieves an article's main event by extracting phrases that answer the journalistic 5Ws. In an evaluation with three assessors and 60 articles, the authors find that the extraction precision of 5W phrases is $p = 0.7$.

Tanev et al. (Tanev et al., 2008) present a real-time news event extraction system that is capable of accurately and efficiently extracting violent and disaster events from online news without using much linguistic sophistication. Lee et al. (Lee et al., 2003) propose an Ontology-based Fuzzy Event Extraction (OFEE) agent for Chinese e-news summarization. The OFEE agent contains Retrieval Agent (RA), Document Processing Agent (DPA) and Fuzzy Inference Agent (FIA) to perform the event extraction for Chinese e-news summarization.

Naughton et al. (Naughton et al., 2006) focus on merging descriptions of news events from multiple sources, to provide a concise description that combines the information from each source. The authors describe and evaluate methods for grouping sentences in news articles that refer to the same event. The key idea is to cluster the sentences, using two novel distance metrics.

Faiz (Faiz, 2006) develop a Natural Language Processing method for extracting temporal information of events from news articles. The extraction process is based on the result of a morpho-syntactic analysis. The obtained results, which are a translation of morpho-syntactic sentences, are scrutinised for temporal markers. Piskorski and Atkinson (Piskorski and Atkinson, 2011) give an overview of the fully operational Real-time News Event Extraction Framework developed for Frontex, the EU Border Agency, to facilitate the process of extracting structured information on border security-related events from online news. In particular, a hybrid event extraction system is constructed, which is then applied to the stream of news articles continuously gathered and pre-processed by the Europe Media Monitor - a large-scale multilingual news aggregation engine.

Piskorski et al. (Piskorski et al., 2011) present a real-time and multilingual news event extraction system developed at the Joint Research Centre of the European Commission. It is capable of accurately and efficiently extracting violent and natural disaster events from online news. In particular, a linguistically relatively lightweight approach is deployed, in which clustered news are heavily exploited at all stages of processing. The technique applied for event extraction assumes the inverted-pyramid style of writing news articles, i.e., a scheme where the most important parts of the story are placed in the beginning and the least important facts are left toward the end.

Wang (Wang, 2012) propose a novel approach of 5W1H event semantic elements extraction (who, what, whom, when, where, how) for Chinese news event knowledge base construction. The approach comprises a key event identification step, an event semantic elements extraction step, and an event ontology population step. The authors first use a machine learning method to identify the key events from Chinese news stories. Then, they extract event 5W1H elements by employing the combination of SRL, NER technique and rule-based method.

Reichart and Barzilay (Reichart and Barzilay, 2012) addressed the extraction of event records from documents that describe multiple events. Specifically, they aimed to identify the fields of information contained in a document and aggregate together those fields that describe the same event. To exploit the inherent connections between field extraction and event identification, they proposed to model the two connections jointly. They experimented with two datasets that consist of newspaper articles describing multiple terrorism events, and showed that their model substantially outperforms traditional pipeline models.

Imran et al. in (Imran et al., 2013a) describe an automatic system for extracting information nuggets from microblog posts such as tweets during disaster times. Their proposed system utilizes machine learning techniques to filter out non-informative tweets and classify the rest into a set of fine-grained classes such as caution and advice, donation, information source, and casualties and damage. They also employ machine learning to extract short self-contained structured information such as location references, time references, number of casualties, type of casualty/damage, type of caution and so on. Their system was trained and tested on a real-world disaster-related dataset consisting of hundreds of thousands of tweets about the Joplin tornado which took place in 2011. The training data for their machine learning techniques was generated using crowdsourcing. The results of their experiments show that indeed machine learning can be utilized to extract structured information nuggets from unstructured text-based microblogging messages with good precision and recall.

In a follow-up work also by Imran et al. (Imran et al., 2013b), the authors utilize conditional random fields to train machine learning models to extract the information nuggets defined in their earlier work (Imran et al., 2013a) from disaster-related tweets. They evaluate their techniques on two disaster-related datasets, the first containing tweets generated during the Joplin tornado in 2011 and the second consisting of tweets generated during the Sandy hurricane in 2012. They report promising results in terms of extraction accuracy. They also test their models on a non-disaster dataset containing tweets related to a sport event and show that their extraction models are useful for extracting information from socially-generated content in general.

Piotrkowicz et al. (Piotrkowicz et al., 2017) propose extracting deductions from headline text as opposed to information nuggets. They build the headline corpus using the Guardian content API, downloading all headlines published during April 2014. Preprocessing of headlines takes place by part of speech tagging (POS) using the Stanford POS Tagger and parsing using the Stanford Parser. The authors also link keywords in text to relevant Wikipedia pages in order to identify entities in the text. This is performed using a tool geared towards short text (TagMe API), typically suited for headlines. In this work, annotation is automatic, however the authors also built a manually annotated gold standard subset of the dataset of 120 headlines, annotated by PhD students in linguistics and calculated inter-annotation agreement using Fleiss Kappa.

Nguyen et al. (Nguyen et al., 2016) describe the MUC-4 corpus. The corpus contains 1700 news articles around terrorist incidents taking place in Latin America. The authors perform document annotation to extract events related to life (injury or death), attacks, charge-indict, arrest-jail, release-parole, etc. from the articles. They also annotate entity coreference chains in documents. However, only the entities appearing at least once as an event argument are annotated with coreference chains. Following this step, the authors perform relevant document retrieval. They retrieve unannotated data from the Web using search engines for each annotated document, and select documents from the Web about the exact same topic, thereby extracting 1724 articles in addition to the 100 annotated documents they originally had. Unannotated documents retrieved from the Web are then used for model learning while manually annotated data is used as a development dataset. In tandem with this, the authors also build a separate test dataset. The authors propose that for better results, the dataset should contain redundancy (several documents about the same events).

Chen et al. in (Chen et al., 2015) describe an event as a specific occurrence involving participants, an event mention as a sentence where the event is mentioned, an event trigger as a verb or noun that implies the event occurred, an event argument as an entity involved in the event, and argument role as the relationship between the argument and the event. The authors use the ACE 2005 dataset (Doddington et al., 2004), which contains text documents from sources such as newswire reports, weblogs, and discussion forums. They then employ a dynamic multi-pooling convolutional neural network with automatically learned features for multiclass classification.

Yang et al. (Yang and Mitchell, 2016) extracted event triggers and the mentions of entities and spatiotemporal information in the ACE 2005 dataset, the same dataset used in (Chen et al., 2015). As a result, they also rely on the same notions defined in (Chen et al., 2015) such as mention, trigger, and argument. The authors propose to extract context in order to be able to extract events from the text. They hypothesize that the interpretation of events is highly contextually dependent and that to make correct predictions, a model needs to concomitantly account for mentions of events and entities together with the discourse of context. They propose a structured model for learning within event structures that can effectively capture the dependencies between an event and its arguments, and between the semantic roles and entity types for the argument.

Gashteovski et al. (Gashteovski et al., 2019) describe an Open information extraction (OIE) corpus called OPIEC, which was extracted from the text of English Wikipedia. OPIEC contains valuable metadata such as provenance information, confidence scores, linguistic annotations, and semantic annotations including spatial and temporal information. The authors analyze the OPIEC corpus by comparing its content with knowledge bases such as DBpedia (Auer et al., 2007) or YAGO (Suchanek et al., 2008), which are also based on Wikipedia. They found that most of the facts between entities present in OPIEC cannot be found in DBpedia and/or YAGO, that OIE facts often differ in the level of specificity compared to knowledge base facts, and that OIE open relations are generally highly polysemous.

3 DATASET CONSTRUCTION

TAGWAR is constructed using the 804 English news articles around the Syrian war that were obtained from the FA-KES dataset (Abu Salem et al., 2019). The total number of sentences in FA-KES is 10,759 sentences (title + content). The average length of the articles is 317.1 words and the total number of unique terms is 14,566. The articles in FA-KES were retrieved from a "variety of media outlets representing mobilisation press, loyalist press, and diverse print media", thus generating a representative corpus of these various types of media outlets. The 804 English news articles from FA-KES report on war incidents taking place in Syria in the years 2011 to 2018.

Each article in FA-KES was subjected to manual annotation by a pair of native Arabic speaking annotators with excellent command of the English language. The annotators were guided by the following questions:

1. How many civilians died in the incident?
2. How many children were targeted in the incident?
3. How many adult women were targeted in the incident?
4. How many non-civilians died in the incident?
5. What is the cause of death?
6. Who does the article blame for the casualties?
7. Where does the article claim the deaths happened?
8. When did the incident happen (Day/Month/Year)?

The annotators were further asked to copy and paste the portion of the article where the answer for each of the above questions appeared. We then used the BIOE sequence tagging scheme with the aim of getting every word in every news article to be associated with a tag. A tag consists of one of these letters: 'B', 'I', 'O', or 'E', designating respectively 'beginning', 'inside', 'outside', or 'end' of an attribute, followed by a '-' sign, and then followed by three letters that represent the type of information that was initially extracted by the annotator. For example, if the information extracted for the location of the incident happens to be 'inside the pizza shop', the first word of the location ('inside') would be tagged as the beginning of location ('B-LOC'), the last word ('shop') would be tagged 'E-LOC' (representing the end of location), and any word in between those two words would be marked as 'I-LOC', representing the inside of location. Following up on this particular example, 'the' and 'pizza' are both tagged as 'I-LOC'. For the remaining words of the articles, which were not retrieved by the annotators, each token would be tagged as 'O', designating the outside of an attribute (words that are outside the scope of information that we are interested in).

Overall, words in the news articles were labeled with one of the following tags: 'B-LOC', 'I-LOC', 'E-LOC', 'B-CIV', 'I-CIV', 'E-CIV', 'B-NCV', 'I-NCV', 'E-NCV', 'B-WMN', 'I-WMN', 'E-WMN', 'B-CHD', 'I-CHD', 'E-CHD', 'B-ACT', 'I-ACT','E-ACT', 'B-COD', 'I-COD', 'E-COD', 'B-DAT', 'I-DAT', 'E-DAT', or 'O' (where 'O' stands for words outside the scope, 'LOC' for location of incident, 'CIV' for number of civilians dead, 'NCV' for number of non-civilians dead, 'WMN' for number of women targeted, 'CHD' for number of children killed, 'ACT' for actor responsible of incident, 'COD' for cause of death, and 'DAT' for date of incident). The total number of labeled tokens in TAGWAR is 13,515 tokens – excluding the words tagged with 'O' – and 256,567 – including the words tagged with 'O'.

The Fleiss-Kappa inter-annotator agreement was calculated between the labels extracted from the two annotators and reported in table 2. The Fleiss-Kappa

Table 2: Fleiss-Kappa Agreement Between Annotators.

Label	Agreement
Location	0.85
Cause-Of-Death	0.72
Actor	0.77
Civilians	0.78
Children	0.83
Women	0.76
NonCivilians	0.76
Date	0.74

Table 3: Detailed Disagreement Between Annotators.

Label	Substring	Blank	Disagreement	Total
Location	29	14	71	114
COD	89	29	98	216
Actor	52	63	39	154
Civilians	23	0	141	164
Children	2	26	17	45
Women	2	14	1	17
NonCivilians	20	0	71	91
Date	36	114	51	201

agreement used is a strict measure of agreement, requiring the answers of the two annotators to be exactly the same for it to count towards an agreement. Table 3 displays the number of articles in our dataset that exhibited disagreement. The *Substring* column corresponds to the number of articles where disagreement was detected because the answer of one annotator turned out to be a substring of the answer of the other annotator. This means that the annotators agreed on the answer but one of them included a longer part of the text in the answer. However, we still count this as a disagreement in our Fleiss-Kappa calculation. The *Blank* column corresponds to the number of articles where disagreement happened when one of the annotators left an answer as a blank (missed the answer in the text) whereas the other annotator extracted an answer from the text. The *Disagreement* corresponds to the number of articles where both annotators gave an answer to the question with neither of them a substring of the other and both are completely different. The *Total* column corresponds to the total number of disagreement for each label (sum of all the aforementioned columns). Note that for in the case of disagreement, we resolved it by picking one of the annotations by the two annotators arbitrarily.

4 AUTOMATIC TAGGING OF NEWS ARTICLES

Next, we trained various deep learning based models to automatically tag a news article using the tagging approach we described earlier. The first such model is a BiLSTM based sequence tagging model, inspired by the OpenTag (Zheng et al., 2018) approach. The model comprises a word embeddings layer as its first layer. The output of the embeddings layer represents the input to a BiLSTM layer. The LSTM cell's job is to generate a hidden vector h_t for each token x_t represented by its embedding e_t, which is passed as input to the LSTM cell. That way the generated vector would be passed as input to the next layer. As our approach is sequence tagging based, which requires a lookup of previous and future contexts, we employed a Bidirectional-LSTM instead of a plain LSTM, which consists of two hidden vectors: one for backward direction, and another for forward direction, where the two vectors are concatenated to form the final output as a new hidden vector h_t defined as:

$$h_t = \sigma([\overrightarrow{h_t}, \overleftarrow{h_t}]) \tag{1}$$

However, a BiLSTM has limitations in the context of our sequence tagging approach. To illustrate, a BiLSTM cannot measure the coherency of tags of a sequence of words. For instance, consider the following sequence of words "Starbucks coffee shop". Here, there is a possibility we might retrieve as labels all of the following: 'B-LOC', 'E-LOC', and 'I-LOC', respectively, despite that the 'E-LOC' label should not precede 'I-LOC'. This could be fixed by adding a Conditional Random Field (CRF) layer which focuses on predicting the labels of a sequence jointly. In this example, the CRF layer would predict 'B-LOC I-LOC E-LOC' for "Starbucks coffee shop". The CRF function can be written as follows:

$$Pr(\mathbf{y}|\mathbf{x};\Psi) \propto \prod_{t=1}^{T} \exp\left(\sum_{k=1}^{K} \Psi_k f_k(y_{t-1}, y_t, x)\right) \tag{2}$$

where $x = \{x_1, x_2, ...x_n\}$ is the input sequence, $y = \{y_1, y_2, ...y_n\}$ is the corresponding label sequence, f_k(y,x) is the feature function; Ψ_k is the corresponding weight to be learnt; K is the number of features; and y_t and y_{t-1} are the neighboring tags at timesteps t and $t-1$, respectively.

Moreover, we also added an attention layer on top of our model which would highlight important concepts, rather than focus on all the information. The attention-focused hidden state representation l_t of a token at timestep t is given by the summation of the hidden state representations $h_{t'}$ of all other tokens at timesteps t', each weighted by their similarity $\alpha_{t,t'}$ to the hidden state representation h_t of the current token:

$$l_t = \sum_{t'=1}^{n} \alpha_{t,t'} . h_{t'} \tag{3}$$

Finally, to avoid over-fitting, we applied the L2 regularization technique to our BiLSTM model.

Table 4: Performance of the BiLSTM model, CRF model, and BERT model on TAGWAR test data.

	CRF Model			BiLSTM Model			BERT Model		
Tag	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
ACT	0.37	0.08	0.14	0.30	0.12	0.17	0.87	0.92	0.85
COD	0.37	0.11	0.17	0.22	0.01	0.02	0.43	1.00	0.59
LOC	0.66	0.74	0.70	0.63	0.66	0.65	0.00	0.00	0.00
CIV	0.52	0.24	0.33	0.40	0.12	0.18	0.42	1.00	0.58
NCV	0.54	0.36	0.43	0.00	0.00	0.00	0.97	1.00	0.98
CHD	0.55	0.25	0.34	0.00	0.00	0.00	0.38	1.00	0.54
WMN	0.00	0.00	0.00	0.00	0.00	0.00	0.44	1.00	0.61
DAT	0.68	0.48	0.57	0.48	0.37	0.42	0.76	1.00	0.86
Average	0.53	0.41	0.44	0.42	0.33	0.34	0.53	0.86	0.63

Table 5: Performance of the BiLSTM model, CRF model, and BERT model on TAGWAR test data (titles only).

	CRF Model			BiLSTM Model			BERT Model		
Tag	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
ACT	0.53	0.57	0.55	0.55	0.52	0.54	0.89	0.72	0.79
COD	0.47	0.56	0.51	0.49	0.56	0.52	0.47	1.00	0.64
LOC	0.77	0.77	0.77	0.73	0.70	0.72	0.25	0.22	0.22
CIV	0.44	0.49	0.46	0.44	0.60	0.50	1.00	1.00	1.00
NCV	0.63	0.49	0.55	0.38	0.43	0.41	1.00	1.00	1.00
CHD	1.00	0.50	0.67	0.80	0.67	0.73	0.83	1.00	0.9
WMN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DAT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average	0.60	0.61	0.60	0.57	0.59	0.58	0.55	0.61	0.56

Table 6: Performance of the BiLSTM model and the CRF model on TAGWAR test data (titles and first paragraph).

	CRF Model			BiLSTM Model			BERT Model		
Tag	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
ACT	0.52	0.34	0.41	0.41	0.38	0.40	0.39	1.00	0.56
COD	0.44	0.43	0.44	0.39	0.39	0.39	0.48	1.00	0.65
LOC	0.73	0.78	0.75	0.62	0.76	0.68	0.00	0.00	0.00
CIV	0.58	0.47	0.52	0.62	0.48	0.54	1.00	0.92	0.96
NCV	0.39	0.38	0.39	0.24	0.23	0.23	0.95	1.00	0.98
CHD	0.59	0.70	0.64	0.50	0.26	0.34	0.52	1.00	0.68
WMN	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.33	0.40
DAT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average	0.57	0.54	0.55	0.50	0.51	0.50	0.48	0.75	0.60

Our second model is a BERT (Bidirectional Encoder Representations from Transformers) model using Google AI Language's BERT (Devlin et al., 2018). The model's architecture is similar to that of the BiLSTM model, but using a BERT layer instead of a word embeddings layer. It consists of an input layer, a masking layer that replaces words with a mask sequence, a BERT layer, which applies the bidirectional training of Transformer, a ReLu layer, and a Sigmoid output layer.

Finally, we also trained a third model consisting of a word embeddings layer followed by a CRF layer. We trained all the models on 80% of the data and tested them on the remaining 20%. Besides, we validated the models on 20% of the training data, which was used to tune the hyperparameters of the different models.

Table 4 shows the results of the different models on the test set of TAGWAR. As can be seen from the table, the BERT model outperforms the BiLSTM model and the CRF model in terms of average precision, recall and f-measure overall the tag classes.

Noting that news articles might many a time involve an element of sensationalism where the bulk of the information is largely revealed through their title or perhaps the title and the first paragraph, we embark to explore the sensitivity of our models in relation to this observation. Table 5 shows the performance of our three models when trained and tested on titles only as opposed to the full content of the news articles. In this case, the CRF model outperforms the other two models in terms of average precision, recall and f-measure over all tag classes. Table 6 presents performance results when the models are trained and tested on both the title as well as the first paragraph of each news article. As can be observed from the table, the BERT model outperforms the BiLSTM and the CRF models in terms of both average recall and average f-measure.

Looking at the three tables together, the overall performance for the CRF model and the BiLSTM one is best when extracting the information from the titles only. This is intuitive given that titles are short and concise compared to the full text of the article, which typically contains a large body of information that does not directly relate to the incident. In contrast, the best performance for the BERT model is attained when the information is extracted from the full text, except when it comes to precision. It is also evident that when testing on the title and first paragraph, which typically reveal most of the relevant information for an incident earlier on, the performance is better than in the case of testing on the full content of the article, except for the BERT model.

When investigating the individual tags, in the case of using the full text (Table 4), the highest performance is obtained for the LOC (location of the incident) and DAT (date of the incident) tags for both the BiLSTM and the CRF models. As for the BERT model, the highest performance is obtained for the NCV tag (number of non-civilians involved in the incident), followed by the DAT tag. On the other hand, when using the titles only or the titles and the first paragraph (Table 5 and Table 6), the ACT (actor) and the LOC tags are the ones that achieved the highest performance for the CRF model. For the BiLSTM model, the tags that performed the best are the CHD (number of children casualties) and LOC. Finally, in the case of the BERT model, the tags CIV (number of civilian casualties) and NCV are the ones that achieved the highest performance.

5 CONCLUSION

In this paper, we described TAGWAR, a dataset consisting of 804 news articles about the Syrian war that were manually sequence tagged. We then used TAGWAR to train and test three deep learning based sequence tagging models to automatically tag news articles, which included a BiLSTM model, a BERT model and a CRF model. Overall, the BERT model performed best when trained and tested on TAGWAR. Moreover, all models with the exception of BERT, performed better when trained on the titles only, as well as on the titles and first paragraph, as opposed to the full content of the news articles. BERT, in contrast, was not significantly sensitive to this aspect of selective training.

We perceive our work to be able to pave the way towards automatic fake news detection around the Syrian conflict. In particular, we plan to deploy, extend, and hone our information extraction models on a large dataset of curated news articles around the Syrian war in order to automatically extract various pieces of information around war incidents. This can in turn form the basis of a robust, end-to-end fact checking pipeline that can allow validating sequence tagged news articles against witness databases such as the Violations Data Center (VDC)[1], one of the leading repositories documenting the human burden of the Syrian war.

REFERENCES

Abu Salem, F. K., Al Feel, R., Elbassuoni, S., Jaber, M., and Farah, M. (2019). Fa-kes: A fake news dataset around the syrian war. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):573–582.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Chen, Y., Xu, L., Liu, K., Zeng, D., and Zhao, J. (2015). Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Faiz, R. (2006). Identifying relevant sentences in news articles for event information extraction. *International Journal of Computer Processing of Oriental Languages*, 19(01):1–19.

Gashteovski, K., Wanner, S., Hertling, S., Broscheit, S., and Gemulla, R. (2019). OPIEC: an open information extraction corpus. *CoRR*, abs/1904.12324.

Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., and Gipp, B. (2018). Giveme5w: main event retrieval from news

[1] https://vdc-sy.net/en/

articles by extraction of the five journalistic w questions. In *International Conference on Information*, pages 356–366. Springer.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013a). Extracting information nuggets from disaster- related messages in social media.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013b). Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1021–1024.

Lee, C.-S., Chen, Y.-J., and Jian, Z.-W. (2003). Ontology-based fuzzy event extraction agent for chinese e-news summarization. *Expert Systems with Applications*, 25(3):431–447.

Naughton, M., Kushmerick, N., and Carthy, J. (2006). Event extraction from heterogeneous news sources. In *proceedings of the AAAI workshop event extraction and synthesis*, pages 1–6.

Nguyen, K.-H., Tannier, X., Ferret, O., and Besançon, R. (2016). A dataset for open event extraction in English. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1939–1943, Portorož, Slovenia. European Language Resources Association (ELRA).

Piotrkowicz, A., Dimitrova, V., and Markert, K. (2017). Automatic extraction of news values from headline text. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 64–74, Valencia, Spain. Association for Computational Linguistics.

Piskorski, J. and Atkinson, M. (2011). Frontex real-time news event extraction framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 749–752. ACM.

Piskorski, J., Tanev, H., Atkinson, M., Van Der Goot, E., and Zavarella, V. (2011). Online news event extraction for global crisis surveillance. In *Transactions on computational collective intelligence V*, pages 182–212. Springer.

Reichart, R. and Barzilay, R. (2012). Multi event extraction guided by global constraints. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 70–79. Association for Computational Linguistics.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203–217.

Tanev, H., Piskorski, J., and Atkinson, M. (2008). Real-time news event extraction for global crisis monitoring. In *International Conference on Application of Natural Language to Information Systems*, pages 207–218. Springer.

Wang, W. (2012). Chinese news event 5w1h semantic elements extraction for event ontology population. In *Proceedings of the 21st International Conference on World Wide Web*, pages 197–202. ACM.

Yang, B. and Mitchell, T. M. (2016). Joint extraction of events and entities within a document context. *CoRR*, abs/1609.03632.

Zheng, G., Mukherjee, S., Dong, X. L., and Li, F. (2018). Opentag: Open attribute value extraction from product profiles. *CoRR*, abs/1806.01264.