CROSS-DOMAIN EVALUATION OF A NON-NEURAL NAMED ENTITY TAGGER

Lewis Bails (leba@itu.dk)

IT University of Copenhagen
Department of Computer Science

ABSTRACT

Although there exists a plethora of annotated text from formal and semi-formal (structured) resources, the same cannot be said for informal, user-generated text. Because of the difference in distributions, language taggers trained solely on structured text are not guaranteed to generalise well to unstructured text [1]. Furthermore, it goes without saying, training a model from scratch with minimal data is undesirable. This paper aims to investigate and evaluate if mixing training data from abundant lower-variance and scarce high-variance distributions results in better performance in named entity recognition for user-generated text.

1. INTRODUCTION

Much of the data available for training language models come from arguably formal and more stringently documented sources such as newswire, broadcast conversation and weblogs. This is true of some of the most popular benchmarking datasets for named entity recognition (NER) [2, 3, 4]. However, there exists limited annotated data for unstructured, user-generated text such as that found on social media or instant messaging services. As such, models trained solely for NER in high-variance environments, such as Twitter or Facebook, can be more unreliable than those stationed within the newswire sphere. Moreover, models trained only on data from formal sources struggle to generalise to those from informal sources. This can be detrimental to, for example, the quality of content recommendations and search results for users of these services.
A common approach to solving this problem is to employ transfer learning. In our case, this may involve fine-tuning a model trained on formal text for classification on informal text. Data mixing, the approach we take to solving this problem, involves training a model with data from both distributions. This hypothesis is that, despite some differences in style, there is much to be learned from the abundant supplementary text. In our case, the majority of training examples come from formal sources; the minority being informal.

In the sections to follow we introduce the formal and informal datasets, outline the training process, and discuss the results on the test set. The test set is a personally annotated dataset consisting of Tweets randomly sampled in October, 2019. The relatively recent nature of the test set provides another challenge of generalising to the ever-evolving online language.

2. PREVIOUS WORK

Recent research regarding user-generated NER has primarily revolved around deep neural architectures with some mentioning fine-tuning large pre-trained models for new distributions. Daniken and Cieliebak [5], who took out second place for entity level and surface form annotations at the 2017 Workshop on Noisy User-generated Text (WNUT) [6] with a bidirectional LSTM (BiLSTM), fine-tuned a model trained on the WNUT 2015 dataset [7]. First place at this workshop went to Aguilar et al. [8] who also used a BiLSTM with a multi-task head which performed named entity segmentation and categorisation.

The emergence of the transformer architecture in 2017 [9], and their bi-directional variant's success on several NLP tasks since [10], has seemingly heralded a new era in language modelling which does away with recurrent models. These transformers act as a base with which to generate contextual embeddings. The embeddings are then used by a prediction layer which performs the actual task at hand. Despite their successes, however, the state-of-the-art (SOTA) models are exceedingly large and potentially unsuitable for production. Furthermore, training models of these sizes is expensive, time-consuming, and not particularly environmentally friendly [11]. Fine-tuning their trained weights, however, is much the opposite and worthy of the recent research focus.

Conditional random fields (CRF) are popular additions to many of the top performing language models, being used as the final classifier [12, 13, 14]. Early CRF implementations only used engineered binary features in their input. However, being essentially structured logistic regression, it is possible to learn weights for features within the continuous domain. For example one may use deep contextualised word/character embeddings or the output of a recurrent neural network [15].

3. DATA

3.1. Formal

The formal text comes from the English dataset proposed in the CoNLL-2003 shared task [3]. The text is from Reuters news stories between August 1996 and August 1997. The training set consists of 14,987 sentences for a total of 203,621 tokens. Tokens are tagged using the IOB-2 format where entities spanning multiple tokens have their first token prefixed by 'B' and subsequent tokens by 'I'. Tokens that are not deemed entities are tagged with 'O', meaning 'other'. Additional per-token features include POS and chunk tags. Possible entities are person (PER), organisation (ORG), location (LOC), and miscellaneous (MISC). From the original paper, 'MISC' "includes adjectives, like *Italian*, and events, like *1000 Lakes Rally*, making it a very diverse category." [3]

3.2. Informal

The informal text comes from the WNUT 2017 long-tail emerging entities dataset [6]. The annotated data in this set is comprised of 1000 tweets from 2011 [1] as well as more recent posts to Reddit, Twitter, YouTube, and StackExchange for a total of 3,395 sequences with 62,729 tokens. The tagging format used is also IOB-2. Possible entities are person, location, corporation, consumer good, creative work, and group. For the sake of agreement between CoNLL and WNUT annotations, the corporation entity is relabelled as organisation and those classes which are not found in the CoNLL annotations are relabelled under a common miscellaneous entity.

3.3. Test

The test set, on which the performance of the models are evaluated, is a personally annotated collection of 1203 (almost-) randomly sampled tweets from October, 2019. 'Almost' in that a small number of tweets were intentionally seeded to introduce cases with ambiguous entities. Entities are chosen from a modified set of labels derived from the OntoNotes 5.0 schema. This was to allow for maximum overlap with the training datasets whilst making full use of the spaCy Python package for pre-labelling (which uses the OntoNotes schema for the English model). All spaCy annotations were checked manually, with several corrections being required. For example, "Trump" is frequently misclassified by the "large" English model as being an organisation rather than a person.

Neither the formal nor the informal dataset use the 7 number-like entities or the law entity found in OntoNotes 5.0; these tags were excluded. The distribution of the entities prior to selecting the most appropriate is shown in figure 1.

Due to time constraints, the test set has only been annotated by one person. It is common practice to, instead, combine annotations from multiple people to smooth out the biases any one annotator may impart on the test set. The combination may be achieved, for example, with majority voting for each tag.

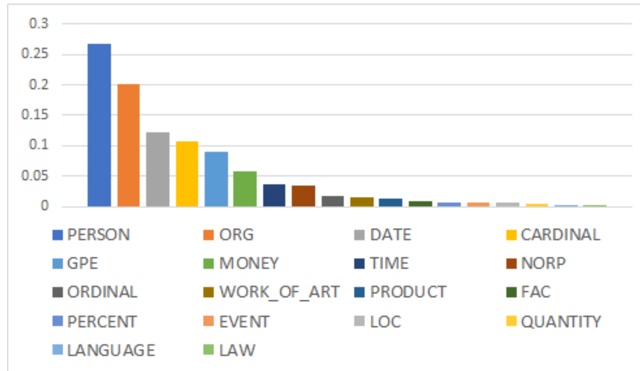

Fig. 1. Distribution of entities in tweets.

3.3.1. *Ambiguous Entities*

A small portion of the tweets are reserved for those with ambiguous entities. These entities are notorious for tripping up NER models and although including them will increase the difficulty of the test set, it is a good test of model quality. Examples of such entities are Alphabet (corporation or collection of letters), Apple (corportation or fruit), Amazon (corporation or location), Uber (corporation or German word meaning over or above), and Emirates (location, facility, or corporation).

Entity	Tweet
Apple	"The Apple never falls far from the tree."
Amazon	"...colloquially known as discus, is a genus of #cichlids native to the Amazon river basin."
Emirates	"#Ozil seeing UnaiEmery leaving the emirates for the last time #arsenal"
Uber	"I'm really not good at being away from my family I was uber sad yesterday"

Table 1. Ambiguous entities.

3.4. Tag Correspondence

Tables 2 and 3 show the how the entity tags are collapsed from the most finely-grained test set down to the most coarse CoNLL-2003 dataset. The correspondence is not perfect. Event entities are not recognised in WNUT, but they are included in CoNLL-2003 under the MISC tag. However, it is quite a rare entity and should not affect the results dramatically. In the OntoNotes schema, the NORP tag includes nationalities, religious, and political groups. However, in CoNLL-2003, political and religious groups are included under the ORG tag, whereas nationalities are included under the

MISC tag. As such, in annotating the test set, political and religious groups are included under the ORG tag instead of the NORP tag. Nationalities are kept under the NORP tag.

Entities	
WNUT-2017	CoNLL-2003
creative-work	MISC
product	MISC
person	PER
location	LOC
group	ORG
corporation	ORG
O	O

Table 2. WNUT-CoNLL tag correspondence.

Entities	
Test	CoNLL-2003
EVENT	MISC
WORK_OF_ART	MISC
PRODUCT	MISC
NORP	MISC
PERSON	PER
FAC	LOC
GPE	LOC
LOC	LOC
ORG	ORG
O	O

Table 3. Test-CoNLL tag correspondence.

4. MODEL

The classifier chosen for this study is the CRF. The CRF can be viewed as a generalisation of multinomial logistic regression to handle sequence classification by way of the same linear-chain framework used in the Hidden Markov Model. CRF is a discriminative classifier in that it directly models the conditional probability of the output given some observed features. The probability of each tag at a given time step is given by an exponentiated weighted combination of these input features, globally normalised.

A common method of parameter (weight) estimation is to select those that maximise the likelihood (in practice the log-likelihood) of the training data. This may be done via gradient based optimisation. In order to efficiently decode the most likely output sequence for each input sequence, the dynamic programming Viterbi algorithm is commonly used.

Three models are trained, each with different data combinations: formal only, informal only, and mixed.

5. FEATURES

Early non-neural approaches for language modelling composed feature vectors from various kinds of engineered features. We argue that the most common fall within the categories of lexical, shape, and grammatical/syntactic features. Lexical features are those that concern the word itself or its sub-parts. For instance: prefixes, suffixes, lemma, and stem. Many shape features are Boolean and also concern the characters themselves. These may include whether the word is uppercase, lowercase, or capitalised. Grammatical and syntactic features are more abstract in that they are a classification of a word based on how it is used within the language of interest or the document in which it appears. A part-of-speech (POS) tag and chunk tag are examples of each of these. The model may also condition the state at a particular time step on that of an arbitrary number of previous time steps. These transition probabilities are inferred during training, alongside the other parameters.

The probability of an tag is usually conditioned not only on the feature vector for the token at the corresponding time step and previous tags, but also on neighbouring tokens. Combined, this vector defines a context window. A context window which spans the tokens from time t-1 to t+1, inclusive, is said to have a radius of 1, The proposed CRF, indeed, has a context window of radius 1.

As mentioned in section 2, modern SOTA systems do not (solely) rely on this kind of manual feature engineering. Instead favouring high-dimensional contextualised word and character embeddings to encode semantic information. But the gain in performance one may achieve is arguably marred by reduced interpretability.

As summary of the features used for this study is shown in table 4. The word shape replaces letters with ”x” or ”X” (capital), punctuation with ”p”, and digits with ”d”. For example, ”H3l!o” becomes ”Xdxpx”. Repeated character types are truncated at length 4.

6. TRAINING

The CRF implementation used is that provided by the python-crfsuite package. The sklearn-crfsuite wrapper module is used for scikit-learn compatability. Each of the models are trained using the L-BFGS algorithm. A randomised grid search is conducted to find reasonable coefficients for the L1 and L2 regularisation terms. The likelihood these parameters take on a particular value is described by exponential distributions parameterised by lambdas of 1 and 0.02, respectively. 20 parameter combinations are tested, with the best model selected based on average weighted flat F1 score from 5-fold cross validation.

Features		
Lexical	Shape	Gramm./Synt.
2-prefix	Is uppercase?	POS
3-prefix	Is capitalised?	Dependency
2-suffix	Is digit?	
3-suffix	Is alpha?	
Normalised	Is lower?	
Lowercase	Is ASCII?	
Lemma	Is punct?	
Word shape	Is stop word?	
	Like URL?	
	Is currency?	

Table 4. Features used per token.

7. RESULTS

Tables 5-7 show test statistics based on a flat, per-token measure. For each of the statistics, the person and location entities are consistently higher than the miscellaneous and organisation entities. The WNUT model recorded the highest overall precision, however the mixed-data model had the highest overall recall and F1 score. Tables 8-10 show test statistics based on several, per-entity (span-based) measures, of varying degrees of strictness. These measures are type, partial, exact, and strict, as proposed by the International Workshop on Semantic Evaluation (SemEval), 2013 [16]. The implementation of the span-based measures is courtesy of Batista [17]. The mixed-data model recorded the highest F1 score for each measure. The type measure saw some particularly low scores given the relatively lenient nature of this measure. This measure considers classifications as correct if there is any overlap between the gold standard and prediction and the entity type is the same. The exact measure, which considers perfect overlap regardless of entity type as correct, is noticeably higher. One would argue this suggests the classifier is better at detecting the entity boundaries than the entity type.

Label	Precision	Recall	F1-score	Support
B-LOC	0.335	0.513	0.405	115
I-LOC	0.261	0.261	0.261	46
B-MISC	0.305	0.269	0.286	171
I-MISC	0.094	0.101	0.097	129
B-ORG	0.334	0.357	0.351	207
I-ORG	0.146	0.371	0.210	62
B-PER	0.429	0.500	0.462	312
I-PER	0.301	0.527	0.383	112
Weighted Ave.	0.314	0.383	0.340	1154

Table 5. Flat classification report for CoNLL model.

Label	Precision	Recall	F1-score	Support
B-LOC	0.468	0.443	0.455	115
I-LOC	0.375	0.391	0.383	46
B-MISC	0.237	0.129	0.167	171
I-MISC	0.218	0.147	0.176	129
B-ORG	0.867	0.126	0.219	207
I-ORG	1.000	0.016	0.032	62
B-PER	0.652	0.551	0.597	312
I-PER	0.699	0.518	0.595	112
Weighted Ave.	0.574	0.318	0.365	1154

Table 6. Flat classification report for WNUT model.

Label	Precision	Recall	F1-score	Support
B-LOC	0.493	0.609	0.545	115
I-LOC	0.389	0.304	0.341	46
B-MISC	0.431	0.275	0.336	171
I-MISC	0.316	0.186	0.234	129
B-ORG	0.500	0.295	0.371	207
I-ORG	0.355	0.355	0.355	62
B-PER	0.695	0.583	0.634	312
I-PER	0.659	0.536	0.591	112
Weighted Ave.	0.524	0.416	0.458	1154

Table 7. Flat classification report for mixed-data model.

Label	Type	Partial	Exact	Strict
LOC	0.460	0.593	0.556	0.419
MISC	0.249	0.447	0.408	0.225
ORG	0.319	0.552	0.527	0.282
PER	0.428	0.590	0.581	0.425
Overall	0.369	0.552	0.530	0.348

Table 8. Span-based F1 scores for CoNLL model.

Label	Type	Partial	Exact	Strict
LOC	0.485	0.653	0.634	0.455
MISC	0.172	0.352	0.336	0.148
ORG	0.169	0.553	0.529	0.169
PER	0.604	0.693	0.596	0.689
Overall	0.406	0.591	0.578	0.394

Table 9. Span-based F1 scores for WNUT model.

Label	Type	Partial	Exact	Strict
LOC	0.618	0.727	0.700	0.582
MISC	0.333	0.478	0.449	0.304
ORG	0.343	0.620	0.593	0.305
PER	0.604	0.736	0.731	0.600
Overall	0.489	0.656	0.637	0.467

Table 10. Span-based F1 scores for mixed-data model.

7.1. WNUT-2017 Shared Task

For the sake of curiosity, a second WNUT-only model was trained, this time with the original labels. Weights on the regularisation terms were the same as those used in the re-labelled training process. The 'entity' and 'surface' F1 scores for the model are found in table 11, alongside the results from the other participating teams [18].

Team	F1 (entity)	F1 (surface)
Drexel-CCI	26.30	25.26
Bails	**28.90**	**26.26**
MIC-CIS	37.06	34.25
FLYTXT	38.35	36.31
Arcada	39.98	37.77
SJTU-Adapt	40.42	37.62
SpinningBytes	40.78	39.33
UH-RiTUAL	41.86	40.24

Table 11. Emerging entities extraction scores (F1 scores out of 100).

8. ANALYSIS

8.1. Confusion Matrices

One cannot immediately draw useful conclusions from the confusion matrices below. We can see that the models generally get the prefix of the tag correct. If pressed, we might say that entity predictions are slightly favouring the PER tag, this is to be expected because of the disproportionate amount of PER tags in the training data. Apart from this, there isn't a particularly strong diagonal, indicating widespread misclassification, which is evident from the underwhelming classification reports in the previous section.

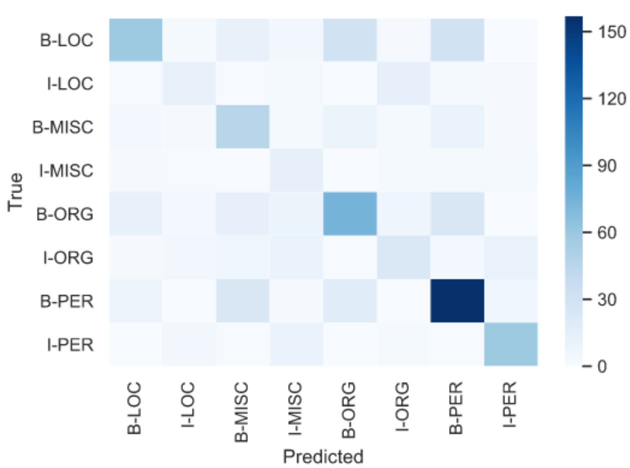

Fig. 2. Confusion matrix for CoNLL model.

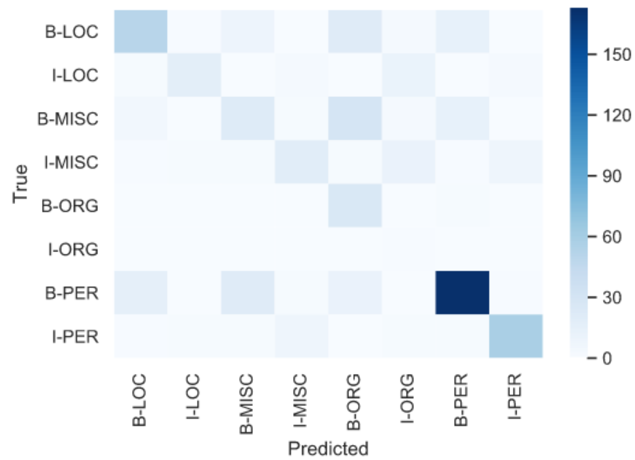

Fig. 3. Confusion matrix for WNUT model.

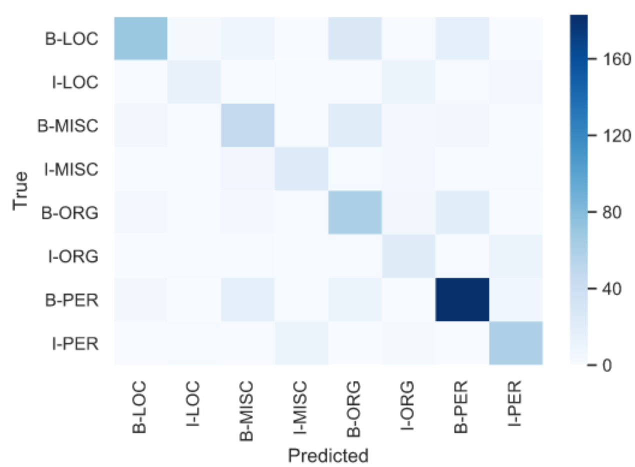

Fig. 4. Confusion matrix for mixed-data model.

8.2. Transition Weights

The transition weights are relatively comparable between models. Large negative weights are seen for transitions from 'O' to any inner tag, as well as from any inner tag to an inner tag of another entity type. This is to be expected. The WNUT model's transition matrix is more sparse than the CoNLL and mixed models due to the fact the training data is smaller and several of the possible transitions are not seen at all during training.

From \ To	O	B-LOC	I-LOC	B-MISC	I-MISC	B-ORG	I-ORG	B-PER	I-PER
O	2.654	1.986	-4.378	1.761	-5.198	2.097	-5.659	4.151	-4.821
B-LOC	-0.002	0.098	8.171	0.748	-1.629	-0.047	-2.3	-1.312	-1.213
I-LOC	-0.707	0	6.817	0	-0.389	-0.646	-1.313	-0.727	-0.687
B-MISC	-0.002	-0.078	-1.735	-1.004	8.375	0.933	-2.737	1.321	-0.919
I-MISC	-1.168	-0.442	-0.634	0.154	7.843	0.361	-1.177	-0.048	-0.687
B-ORG	0.056	-1.157	-1.302	-0.317	-0.769	-0.817	8.875	-0.787	-1.022
I-ORG	-0.731	-1.685	-0.969	-0.227	-0.824	-1.628	8.049	-0.361	-1.489
B-PER	-0.126	0	-0.057	-1.058	-0.375	-1.343	-0.739	-2.356	9.014
I-PER	0	-0.859	-0.022	-0.377	-0.206	-0.162	-0.666	-1.273	6.801

Fig. 5. Weights on transitions between tags for CoNLL model.

From \ To	O	B-LOC	I-LOC	B-MISC	I-MISC	B-ORG	I-ORG	B-PER	I-PER
O	3.758	0.058	-6.058	1.562	-7.335	1.362	-3.858	1.978	-5.977
B-LOC	0	0.146	8.08	-0.257	-0.828	-0.852	-0.253	-1.233	-1.101
I-LOC	-0.938	0	6.87	0.09	-0.656	0	-0.18	0	-0.337
B-MISC	-0.554	-1.173	-0.45	0	8.944	-0.22	0	-0.743	0
I-MISC	-0.645	-1.322	-0.79	-0.899	8.026	-0.149	0	-0.767	-0.776
B-ORG	-0.48	0.283	-0.525	-0.369	-0.69	0.756	6.314	-1.218	0
I-ORG	-0.646	0.198	0	-0.47	0	0	5.944	0	0
B-PER	-0.027	-1.54	-0.017	-0.892	-1.214	0	-0.139	0	9.583
I-PER	-0.399	-0.662	0	-0.359	-0.817	0	0	0.402	6.304

Fig. 6. Weights on transitions between tags for WNUT model.

From \ To	O	B-LOC	I-LOC	B-MISC	I-MISC	B-ORG	I-ORG	B-PER	I-PER
O	2.92	0.739	-6.478	1.218	-7.249	1.058	-7.336	2.721	-7.159
B-LOC	-0.076	0.376	9.297	0.613	-2.078	-0.098	-2.855	-1.867	-1.365
I-LOC	-0.465	0.434	8.089	0.221	-1.117	-1.268	-1.495	-0.295	-0.553
B-MISC	0.053	-0.412	-1.981	-0.518	9.18	0.918	-3.286	1.275	-1.376
I-MISC	-0.73	-1.46	-0.825	0	8.11	-0.535	-1.531	-0.504	-0.783
B-ORG	0.005	-0.884	-0.984	-0.284	-1.984	0	9.938	-1.465	-0.639
I-ORG	-0.696	-1.419	-0.685	-0.074	-1.98	-1.831	9.188	-0.322	-1.107
B-PER	-0.248	-0.324	0	-0.975	-1.129	-1.467	-0.869	-1.502	9.864
I-PER	0.02	-1.337	-0.158	-0.604	-0.896	-0.119	-0.789	-0.101	7.417

Fig. 7. Weights on transitions between tags for mixed-data model.

8.3. Feature Weights

Figures 8-10 show the 5 greatest and least weighted features for each tag. Across all models, there is a relatively strong positive correlation between the beginning or end of a sentence and the observed token not being an entity (O). Likewise, having a 'day' suffix is a strong indicator of a non-entity token. For the WNUT model, the B-ORG tag has large weights on the *lower* and *norm* dummy variables for 'Twitter', 'Facebook', and 'Walmart'. This indicates some overfitting as the model is simply memorising entities. As similar argument can be made for the I-MISC tag of the CoNLL model. As one may expect, with entities commonly being capitalised words, *is_lower* and *is_punct* are regularly among the strong negative weights and never among the strongest positive weights for any entity tag.

8.4. Feature Importance

In section 8.3, we analysed the relationships between individual features and each tag. In this section we analyse how important each feature is to the overall performance of the model. Retraining models with subsets of features to evaluate their importance is computationally expensive. One may avoid this by instead evaluating permutation importance for each feature [19]. With this approach we remove any useful information a particular feature offers the trained model by replacing it with noise in each of the test examples. Shuffling, or permuting, the feature 'column' among test examples has this effect.

Tables 12-14 show 10 features of greatest importance to each of the models. The numbers inside the parenthesis beside the feature names are the mean and standard deviation of the reduction in flat F1-score when that feature is not available to the model. The greater the performance reduction, the greater the perceived importance. The central token's part-of-speech appears to be of significant importance to each of the models, followed by its shape and several 'shape'-like binary features. Apart from a few affixes, there is a distinct lack of lexical features. It is possible that a strong correlation between lexical features is the cause of this. By permuting one of the correlated features, the model still has access to the information via the other correlated feature. This would result in a lower reported importance than the true value. For example, lexical features 'norm' and 'lower' are the same for 99% of the tokens. Spearman's rank correlation coefficient for these features is 0.96. In an attempt to address this, we could cluster the features based on how correlated we believe them to be, and shuffle all columns in the cluster. The importance across the models for the aforementioned cluster and separate features is shown in table 15. For each model, the mean cluster importance is higher than the sum of the individual mean feature importances. We leave further cluster analysis for future studies.

Rank	Feature
1	POS (0.086, 0.005)
2	is_lower (0.063, 0.005)
3	shape (0.038, 0.005)
4	is_title (0.028, 0.003)
5	is_alpha (0.023, 0.002)
6	-1shape (0.018, 0.004)
7	-1POS (0.018, 0.003)
8	is_stop (0.017, 0.005)
9	suffix3 (0.016, 0.002)
10	EOS (0.016, 0.005)
Base F1 score	0.340

Table 12. Feature permutation importance for CoNLL model.

O		B-LOC		I-LOC		B-MISC		I-MISC		B-ORG		I-ORG		B-PER		I-PER	
W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature
5.814	suffix3:day	3.142	lemma:golf	1.972	-1shape:XXX	3.392	shape:X$	1.899	-1dep:dep	2.703	suffix2:EC	2.704	-1shape:X	3.408	+1shape:ddxx	2.64	-1shape:X.
4.903	is_lower	2.997	suffix2:ia	1.701	suffix3:ast	3.152	suffix3:ans	1.847	norm:masters	2.611	shape:XXX	2.661	+1shape:d	2.786	shape:X.	2.61	is_title
3.376	shape:X	2.982	suffix2:IA	1.601	suffix2:ka	2.712	suffix3:ian	1.847	lower:masters	2.577	suffix2:RC	1.816	+1shape:ddd.d d	2.668	+1shape:d-d-dd-d	1.741	suffix2:ez
3.315	BOS	2.699	+1shape:dddd-dd-dd	1.575	+1shape:dddd-dd-dd	2.646	suffix2:SH	1.823	norm:open	2.509	suffix3:ire	1.707	suffix2:rs	2.514	shape:XxXxxxx	1.655	-1suffix3:nda
3.204	EOS	2.422	norm:germany	1.478	is_digit	2.644	suffix3:ese	1.823	lower:open	2.486	BOS	1.566	-1norm:boatmen	2.478	suffix2:ER	1.569	-1dep:compound
-2.25	-1shape:X	-1.51	suffix3:ura	-1.1	is_oov	-1.22	suffix2:al	-1.19	is_upper	-1.72	suffix3:ngo	-1.35	-1suffix2:om	-1.78	suffix3:men	-1.34	-1suffix3:son
-2.4	suffix2:AN	-1.59	is_lower	-1.41	+1suffix2:er	-1.27	suffix2:ia	-1.23	-1suffix2:SH	-1.8	-1shape:dddd	-1.61	EOS	-1.92	suffix3:ion	-1.57	-1shape:XXXX
-2.46	-1shape:Xx	-1.6	suffix3:ena	-1.49	+1suffix2:th	-1.43	is_punct	-1.27	-1suffix2:ed	-2	suffix3:ana	-1.65	+1pos:SYM	-1.96	-1is_ascii	-1.68	-1suffix2:ch
-2.65	token:division	-1.7	suffix2:es	-2.31	+1is_digit	-1.53	is_alpha	-1.32	+1suffix2:ne	-2.26	shape:X	-1.78	suffix2:er	-2.01	suffix3:ans	-1.87	is_ascii
-2.76	pos:PROPN	-2.02	dep:det	-2.91	-1suffix2:ia	-3.78	is_lower	-1.46	-1suffix2:se	-3.1	is_lower	-2.29	+1shape:ddd	-2.08	suffix3:ire	-1.87	bias

Fig. 8. Largest and smallest weights for each tag - CoNLL model.

O		B-LOC		I-LOC		B-MISC		I-MISC		B-ORG		I-ORG		B-PER		I-PER	
W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature
7.506	suffix3:day	4.004	shape:X.X.	2.281	norm:center	3.267	BOS	2.999	shape:x	5.255	norm:twitter	2.017	-1is_title	3.829	shape:#XXXX	2.262	+1shape:XX
5.406	BOS	3.058	suffix3:lhi	2.251	+1suffix3:ery	3.094	shape:xXxxxx	2.269	-1suffix2:ux	5.255	lower:twitter	1.621	prefix2:In	3.438	lower:pope	2.233	suffix3:yer
4.323	EOS	3.028	shape:#Xxxxx	2.248	dep:pobj	3.079	-1lemma:watch	2.253	dep:advcl	4.495	lower:facebook	1.621	prefix3:In	3.438	norm:pope	2.186	-1suffix2:il
3.261	is_punct	2.865	suffix3:nia	2.05	suffix2:ay	3.015	suffix2:ua	2.207	-1suffix3:ndy	4.495	norm:facebook	1.453	suffix2:ut	3.273	suffix3:oss	2.007	+1prefix3:Ni
3.006	shape:x	2.824	suffix2:ca	2.031	+1suffix2:ne	2.857	lemma:Youtube	2.043	shape:d	2.875	norm:walmart	1.394	suffix3:ner	3.251	+1suffix3:kes	2.007	+1prefix2:Ni
-2.61	suffix2:EN	-1.25	-1pos:PART	-1.5	-1pos:PROPN	-1.27	-1suffix2:ll	-2.06	+1dep:nsubj	-1.28	+1suffix2:en	-0.84	is_alpha	-1.94	suffix3:ica	-1.62	-1suffix2:om
-2.66	suffix2:ex	-1.67	shape:X	-1.62	dep:nsubj	-1.37	is_lower	-2.07	shape:dd	-1.63	shape:Xxxx	-0.86	suffix2:on	-2.08	suffix3:den	-1.68	suffix2:ce
-2.78	-1shape:Xxx-	-1.99	suffix2:es	-1.88	+1pos:VERB	-1.42	+1dep:aux	-2.14	suffix2:se	-1.67	shape:Xxx	-0.89	-1is_alpha	-2.27	is_stop	-1.7	+1dep:ccomp
-3.09	suffix3:ods	-2.12	-1is_ascii	-1.91	-1dep:nmod	-1.61	shape:Xxx	-2.32	shape:xxxx	-1.87	shape:XX	-1.59	pos:NOUN	-2.38	+1dep:aux	-2.36	pos:VERB
-3.76	suffix2:si	-2.66	shape:Xxx	-2.48	-1shape:XX	-1.96	is_punct	-2.4	+1dep:dep	-2.15	-1is_ascii	-2.6	+1dep:punct	-2.94	-1is_ascii	-2.41	bias

Fig. 9. Largest and smallest weights for each tag - WNUT model.

O		B-LOC		I-LOC		B-MISC		I-MISC		B-ORG		I-ORG		B-PER		I-PER	
W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature	W	Feature
6.56	suffix3:day	3.761	lemma:golf	1.982	norm:center	4.848	shape:X$	2.632	shape:x	4.506	lower:twitter	3.145	-1shape:X	4.188	suffix3:dys	2.761	-1shape:X.
5.527	BOS	3.546	+1shape:dddd-dd-dd	1.834	-1shape:XXX	4.541	shape:xxxx-Xxxxx	2.483	norm:open	4.506	norm:twitter	3.122	+1shape:d	3.695	+1shape:ddxx	2.714	suffix2:ez
3.171	is_punct	3.521	suffix2:IA	1.828	-1suffix2:ew	3.746	shape:Xxxxx-xxx	2.483	lower:open	3.867	BOS	2.646	+1suffix3:ker	3.647	shape:XxXxxxx	2.586	is_title
3.166	shape:X	3.368	+1shape:dd,ddd	1.822	shape:dd	3.701	BOS	2.435	-1dep:dep	3.634	shape:XxxXxxx	2.331	+1shape:ddd.dd	3.297	+1shape:d-d-dd-d	2.322	-1shape:xxx
3.15	EOS	3.1	suffix2:ia	1.805	+1shape:dddd-dd-dd	3.588	suffix3:ese	2.118	shape:Xxxxx-xxxx	3.358	norm:facebook	2.137	-1suffix2:ap	3.27	-1shape:d.	1.896	prefix3:BE
-2.75	lemma:french	-1.75	suffix2:ez	-1.69	+1shape:dd	-1.66	+1lemma:do	-1.91	-1pos:PUNCT	-1.76	suffix3:ana	-1.64	-1shape:xxxx	-1.72	+1suffix2:ls	-1.64	-1suffix3:een
-2.78	+1shape:dddd-dd-dd	-1.84	suffix3:ish	-1.83	+1suffix2:er	-1.75	dep:prep	-1.93	+1suffix3:ach	-1.82	suffix2:PI	-1.86	-1suffix2:om	-1.76	-1is_ascii	-2.04	-1suffix2:ch
-2.87	suffix2:LO	-1.84	suffix2:es	-2.17	-1suffix2:go	-1.78	-1suffix3:can	-1.99	pos:PUNCT	-1.84	+1shape:dddd-dd-dd	-2.15	suffix2:er	-1.78	+1shape:d-d	-2.08	-1suffix3:son
-2.94	shape:Xxxxx-xxxx	-1.86	+1is_digit	-2.4	+1is_digit	-2.63	is_punct	-2.2	-1suffix2:at	-1.99	shape:X	-2.15	shape:d	-2.03	suffix3:ban	-2.35	+1lemma:win
-3.3	token:division	-2.46	dep:det	-3.08	-1suffix2:ia	-2.7	is_lower	-2.31	shape:xxxx	-2.49	is_lower	-2.48	+1shape:ddd	-3.11	suffix3:ans	-2.61	bias

Fig. 10. Largest and smallest weights for each tag - mixed-data model.

Rank	Feature
1	POS (0.167, 0.004)
2	is_stop (0.072, 0.005)
3	shape (0.064, 0.004)
4	is_lower (0.049, 0.002)
5	is_title (0.047, 0.004)
6	dep (0.039, 0.006)
7	-1dep (0.030, 0.009)
8	is_alpha (0.027, 0.007)
9	suffix2 (0.023, 0.009)
10	-1is_title (0.021, 0.001)
Base F1 score	0.365

Table 13. Feature permutation importance for WNUT model.

Rank	Feature
1	POS (0.174, 0.006)
2	is_lower (0.076, 0.006)
3	is_title (0.056, 0.004)
4	shape (0.050, 0.005)
5	is_stop (0.043, 0.009)
6	suffix3 (0.042, 0.006)
7	is_alpha (0.036, 0.002)
8	-1POS (0.032, 0.005)
9	suffix2 (0.029, 0.004)
10	-1shape (0.028, 0.002)
Base F1 score	0.458

Table 14. Feature permutation importance for mixed-data model.

Feature	CoNLL	WNUT	Mixed
norm	0.006 (0.003)	0.016 (0.003)	0.015 (0.001)
lower	0.006 (0.003)	0.016 (0.004)	0.018 (0.004)
norm+lower	0.012 (0.004)	0.032 (0.005)	0.033 (0.004)
(norm,lower)	0.014 (0.005)	0.047 (0.003)	0.043 (0.003)

Table 15. Example cluster importance.

9. CONCLUSIONS

From the results observed, there is reason to believe that, for tagging entities in tweets, using formal text to supplement informal text can be advantageous. For both flat and span-based F1 scores, the mixed-data approach to training a CRF outperformed both the formal-only and informal-only approaches. Each of the models performed best on person and location entities and worst on the miscellaneous entities. This is possibly due to the imperfect collapsing scheme used to re-label the informal and test datasets. According to the non-parametric permutation technique for determining feature importance, the POS tag and several shape features exhibit the strongest positive correlation with the overall flat F1 score. However, correlated features may be masking each other's importance, as exemplified in table 15.

10. FUTURE WORK

This study only makes use of engineered features which possibly fail to fully capture the semantic meaning of words. Conversely, recent approaches to tagging make use of word embeddings, whether contextual or otherwise. These embeddings are able to accurately capture semantic meaning. In the future, one may repeat this study using continuous work embeddings instead of engineered feature vectors, or, perhaps, a combination of both.

Naturally, the user-generated text, being informal and largely unedited, has a higher rate of grammatical errors and spelling mistakes. The abundant formal training data is quite the opposite. This leads one to hypothesise that even better results would be seen for the mixed-data approach if we were to synthesise training data via augmentation of the formal data. For example, we could drop random letters, switch letters, or replace entire words with synonyms.

11. REFERENCES

[1] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni, "Named entity recognition in tweets: An experimental study," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., July 2011, pp. 1524–1534, Association for Computational Linguistics.

[2] Erik F. Tjong Kim Sang, "Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition," in *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[3] Erik F. Tjong Kim Sang and Fien De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, Stroudsburg, PA, USA, 2003, CONLL '03, pp. 142–147, Association for Computational Linguistics.

[4] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel, "Ontonotes: The 90in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, Stroudsburg, PA, USA, 2006, NAACL-Short '06, pp. 57–60, Association for Computational Linguistics.

[5] Pius von Däniken and Mark Cieliebak, "Transfer learning and sentence level features for named entity recognition on tweets," in *Proceedings of the 3rd Workshop*

on Noisy User-generated Text, Copenhagen, Denmark, Sept. 2017, pp. 166–171, Association for Computational Linguistics.

[6] *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[7] *Proceedings of the Workshop on Noisy User-generated Text*, Beijing, China, July 2015. Association for Computational Linguistics.

[8] Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio, "A multi-task approach for named entity recognition in social media data," in *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Copenhagen, Denmark, Sept. 2017, pp. 148–153, Association for Computational Linguistics.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[11] Emma Strubell, Ananya Ganesh, and Andrew McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 3645–3650, Association for Computational Linguistics.

[12] Yufan Jiang, Chi Hu, Tong Xiao, Chunliang Zhang, and Jingbo Zhu, "Improved differentiable architecture search for language modeling and named entity recognition," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3576–3581, Association for Computational Linguistics.

[13] Jana Straková, Milan Straka, and Jan Hajic, "Neural architectures for nested NER through linearization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 5326–5331, Association for Computational Linguistics.

[14] Abbas Ghaddar and Phillippe Langlais, "Robust lexical features for improved neural network named-entity recognition," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, Aug. 2018, pp. 1896–1907, Association for Computational Linguistics.

[15] Zhiheng Huang, Wei Xu, and Kai Yu, "Bidirectional LSTM-CRF models for sequence tagging," *CoRR*, vol. abs/1508.01991, 2015.

[16] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo, "SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, June 2013, pp. 341–350, Association for Computational Linguistics.

[17] D. Batiasta, "Named entity evaluation as in semeval 2013 task 9.1," https://github.com/davidsbatista/NER-Evaluation, 2019.

[18] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham, "Results of the WNUT2017 shared task on novel and emerging entity recognition," in *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Copenhagen, Denmark, Sept. 2017, pp. 140–147, Association for Computational Linguistics.

[19] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001.