Named Entity Analysis and Extraction – Discovering the Beauty of Power Law in Entity Length

Anonymous ACL submission

Abstract

We find from two benchmark datasets that the *non-common words*, which appear in named entities and hardly appear in common text, can differentiate named entities from common text and that the length of named entities follows a family of power law distributions. The findings motivate us to design a learning method named POM to extract named entities from free text.[1] POM defines a constituent-based tagging scheme with three tags, namely P, O, and M, indicating those non-common words as Predictors, Modifiers that modify predictors, and the words Outside named entity. In modeling, POM assigns a word with a POM tag under conditional random fields. Experiments demonstrate the effectiveness and efficiency of POM on two benchmark datasets compared with two representative state-of-the-art methods.

1 Introduction

Named entity extraction is a fundamental task for many further research and applications, such as named entity typing (Ling and Weld, 2012; Nakashole et al., 2013), entity linking (Ji and Grishman, 2011; Ling et al., 2015b), domain-specific entity recognition (Takeuchi and Collier, 2005; Krallinger et al., 2015), and aspect-based sentiment analysis (Pontiki et al., 2016).

Named entity extraction is defined by Grishman and Sundheim (1996); Chinchor (1997); Sang and Meulder (2003), aiming to extract named entities from free text. The task definition points to a central question: *what differentiates named entities from common words in free text?*

1) African/S	2) African/B Nations/I Cup/E
3) West/B African/E	4) West/B African/I States/E

(a) IOBES assignment: 'African' in different positions within labeled named entities is assigned with different tags of S, B, E, and I. The inconsistent tag assignment weakens the predictive power of 'African.'

1) African/P	2) African/P Nations/M Cup/M
3) West/M African/P	4) West/M African/P States/M

(b) POM assignment: 'African' in different positions within labeled named entities is consistently assigned with same tag of P. This protects 'African's predictive power.

Figure 1: IOBES and POM assignment in training

To answer the question we analyze two benchmark datasets (i.e., CoNLL03 (Sang and Meulder, 2003) and OntoNotes5 (Pradhan et al., 2013)) for the characteristics of named entities and have three findings. First, more than 92.2% of named entities contain *non-common words*, which hardly appear in common text. Second, named entities are mainly made up of proper nouns; in the whole text, more than 84.8% of proper nouns appear in named entities, and in named entities, more than 80.1% of words are proper nouns. Third, the length of named entities follows power law distributions, with well-defined means and finite variances.

The findings motivate us to design a learning method named POM to model named entities. Specifically, POM defines a constituent-based tagging scheme named POM scheme,[2] consisting of three tags: P, O, and M. P denotes the non-common words as Predictors, which are capable of predicting named entities, such as 'African' and 'Chinese.' M includes the words that Modify predictors in named entity; for example, 'West' modifies 'African' in 'West African.' O indicates the words appearing Outside named entity. POM models named entities under a framework of

[1] We will release the source code if the paper was accepted.

[2] We use 'POM' to denote our method and use 'POM scheme' to denote the tagging scheme that POM defines.

conditional random fields (CRFs) (Lafferty et al., 2001) by assigning a POM tag to a word.

POM scheme overcomes the problem of inconsistent tag assignment and makes full use of the predictive power of the non-common words. Zhong and Cambria (2018) illustrate that position-based tagging schemes like IOBES (Beginning-Inside-End-Single-Outside) suffer from the problem of inconsistent tag assignment and weakens the predictive power of time tokens when modeling time expressions. That problem widely exists when the position-based tagging schemes are used to model named entities. Under IOBES scheme, for example, the word 'African' in the four named entities 'African,' 'West African,' 'West African States,' and 'African Nations Cup' is assigned with different tags of S, E, I, and B. See Figure 1(a). By contrast, our POM scheme models named entities based on their constituents and assigns same constituent word with same tag. For example, under POM scheme, the above four 'African' are consistently assigned with the same tag of P. See Figure 1(b). The consistent tag assignment protects the predictive power of the non-common words. POM scheme is inspired by the constituent-based TOMN scheme (Zhong and Cambria, 2018); while TOMN scheme is defined to model only time expressions, POM scheme can model general entities, including time expressions.

We evaluate POM on two datasets against two representative state-of-the-art methods and experiment results show the effectiveness and efficiency of POM compared with the state-of-the-art methods. Moreover, our experiments show that simple handcrafted feature methods achieve comparable results but are much more efficient than neural network based methods in named entity extraction. This suggests that the named entity extraction and classification should be addressed separately.

In summary, our contributions are as follows.

- We recognize from two benchmark datasets the capability of non-common words to differentiate named entities from common text and discover the beauty of power law in entity length.
- We model named entities by a constituent-based tagging scheme, which can protest the predictive power of the non-common words.
- We conduct experiments on two datasets, and the results verify our method's effectiveness and efficiency and suggest us to address the named entity extraction and classification separately.

2 Related Work

Our work is related to named entity recognition and power law in language.

2.1 Named Entity Recognition (NER)

Named entity recognition has a long history. Nadeau and Sekine (2007) summarize early years' development in terms of languages (e.g., English, German, and Chinese) (Wang et al., 1992; Chen and Lee, 1996; Grishman and Sundheim, 1996; Chinchor, 1997; Sang and Meulder, 2003), text types (e.g., scientific and journalistic) and domains (e.g., sports and business) (Maynard et al., 2001; Poibeau and Kosseim, 2001; Minkov et al., 2005), statistical learning techniques (e.g., hidden Markov models, maximum entropy models, and conditional random fields) (Bikel et al., 1997; Sekine, 1998; Borthwick et al., 1998; Asahara and Matsumoto, 2003; McCallum and Li, 2003), and engineering features (e.g., word-level features and dictionary features) (Bikel et al., 1997; Ravin and Wacholder, 1997; Yu et al., 1998; Collins and Singer, 1999; Collins, 2002; Silva et al., 2004).

Recently, many researchers use neural networks and word embeddings to develop models on the CoNLL03 dataset (Collobert et al., 2011; Passos et al., 2014; Huang et al., 2015; Ling et al., 2015a; Santos and Guimaraes, 2015; Luo et al., 2015; Lample et al., 2016; Ma and Hovy, 2016; Chiu and Nichols, 2016; Xu et al., 2017; Liu et al., 2018).

POM benefits some features from the traditional methods, and refines the significant features and filter out the irrelevant features according to a deeper analysis of named entities. Unlike the neural network based methods that compute semantic similarities among words, POM focus on the distinction between named entities and common text. And unlike most NER methods that treat the named entity extraction and classification as an end-to-end task, our analysis suggests that the two subtasks should be addressed separately.

2.2 Power Law in Language

Zipf's law reveals that the words' rank-frequency in corpora follows power law distributions (Zipf, 1936, 1949; Mandelbrot, 1961). Sigurd et al. (2004) observe that the word length and sentence length can be fitted by a variant of gamma distributions but not power law distributions. We now discover that the length of named entities in corpora follows a family of power law distributions.

Table 1: Statistics of the datasets. '#T' denotes the number of entity types.

Dataset		#Docs	#Words	#Entities	#T
CoNLL03	Train	946	203,621	23,499	4
	Dev.	216	51,362	5,942	
	Test	231	46,435	5,648	
	Total	**1,393**	**301,418**	**35,089**	
OntoNotes5	Train	2,729	1,578,195	124,057	18
	Dev.	406	246,009	19,960	
	Test	235	155,330	11,396	
	Total	**3,370**	**1,979,534**	**155,413**	
OntoNotes*	Train	2,729	1,578,195	81,222	11
	Dev.	406	246,009	12,721	
	Test	235	155,330	7,537	
	Total	**3,370**	**1,979,534**	**101,480**	

3 Named Entity Analysis

3.1 Datasets

The datasets we use for named entity analysis are CoNLL03, OntoNotes5, and OntoNotes*.[3]

CoNLL03 is a benchmark dataset derived from Reuters RCV1 Corpus, with 1,393 news articles between August 1996 and August 1997; it contains 4 entity types[4] (Sang and Meulder, 2003).

OntoNotes5 is a portion of OntoNotes 5.0 dataset for named entity analysis and consists of 3,370 articles collected from different sources: newswire, web data, broadcast news, and broadcast conversation; it has 18 entity types[5] (Pradhan et al., 2013).

OntoNotes*. We find that OntoNotes5 is far from perfect in annotation. For example, its named entity guideline states that ORDINAL includes all the ordinal numbers and CARDINAL includes the whole numbers, fractions, and decimals, but we find many ordinals and numbers in common text. To get a high quality dataset for named entity analysis, we derive a dataset named OntoNotes* from OntoNotes5 by removing the entity types whose named entities are mainly composed of numerals.[6] In addition, some sequences are annotated inconsistently; for the 'the Cold War,' for example, in some cases the whole sequence is treated as an entity while in some cases only the 'Cold War' is an entity. In OntoNotes*, all the 'the' at the beginning of entities and all the ''s' at the end of entities are moved outside the entities.

Table 2: Percentage of named entities each has at least one word that hardly appears in common text.

	Train	Dev.	Test
CoNLL03	98.77	99.19	98.62
OntoNotes*	92.20	95.22	95.61

Table 1 summarizes the statistics of the datasets. In setting the training, development, and test sets, we follow the setting by CoNLL03 shared task (Sang and Meulder, 2003) for CoNLL03 and follow the public setting[7] by OntoNotes 5.0 dataset's author for OntoNotes5 and OntoNotes*. OntoNotes5 is only used in entity length analysis.

3.2 Findings

Although the datasets vary in source, corpus size, text type, generated time, and concern different entity types, we find that their named entities demonstrate some common characteristics.

Finding 1 *Named entity contains non-common word(s); more than 92.2% of named entities each has at least one word that hardly appear in the common text*

Table 2 reports the percentage of named entities each has at least one word that hardly appears in common text (case sensitive). 'Common text' here means the whole text with named entities excluded. For a word, if the rate of its occurrences in named entities over the ones in the whole text reaches a threshold, then we treat it as hardly appearing in common text. For CoNLL03, the threshold is set by 1; which means the word does not appear in common text. For OntoNotes*, the threshold is set by 0.95; because the common text contains some words that should be treated as named entities, such as 'Google' and 'American.'

We can see that for each of the training, development, and test sets, more than 92.2% of named entities have words hardly appearing in common text. Such kind of words can be used to predict named entities and we call them as *non-common words*. The non-common words in the development and test sets also hardly appear in the common text of training set.

Finding 2 *Named entities are mainly made up of proper nouns. In the whole text, more than 84.8% of proper nouns appear in named entities; and within named entities, more than 80.1% of the words are proper nouns.*

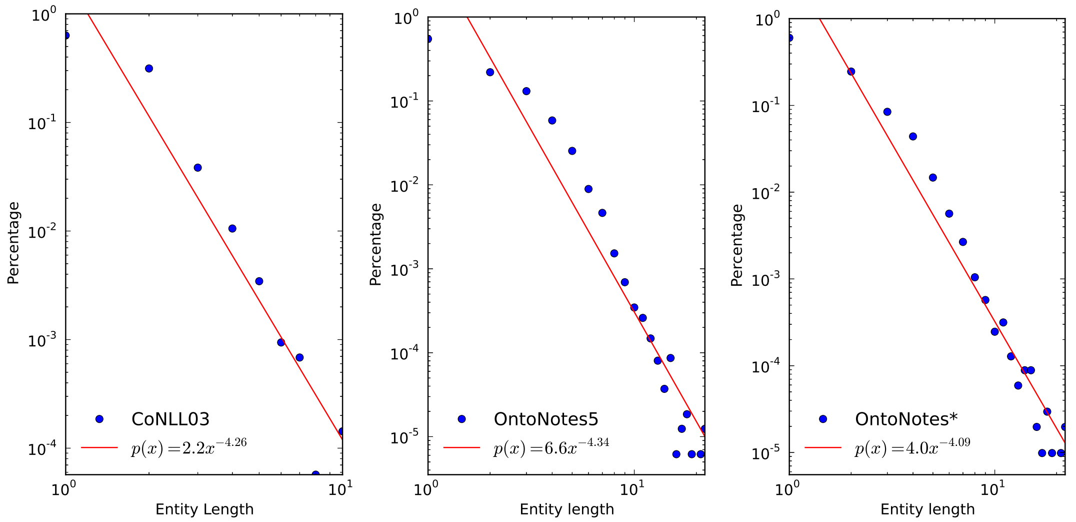
[3] The original CoNLL03 and OntoNotes5 datasets include data in English and other languages, here we focus on the English data.

[4] CoNLL03's 4 entity types are PER, LOC, ORG, and MISC.

[5] OntoNotes5's 18 entity types are CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, and WORK_OF_ART.

[6] The removed entity types include CARDINAL, DATE, MONEY, ORDINAL, PERCENT, QUANTITY, and TIME.

[7] https://github.com/ontonotes/conll-formatted-ontonotes-5.0



(a) CoNLL03 (b) OntoNotes5 (c) OntoNotes*

Figure 2: Length distribution of named entities. Entity length denotes the number of words in entities.

Table 3: Top 4 POS tags in named entities and their percentage over the whole tags in named entities (P_{entity}) and over the corresponding tags in the whole text (P_{text})

CoNLL03			OntoNotes*		
POS	P_{entity}	P_{text}	POS	P_{entity}	P_{text}
NNP	83.81	84.82	NNP	77.67	85.88
JJ	5.82	17.57	JJ	4.60	6.77
NN	4.89	6.46	NN	4.57	2.91
NNPS	1.55	94.12	NNPS	2.50	93.04

We find that named entities are mainly made up of proper nouns.[8] Table 3 lists the top 4 part-of-speech (POS) tags in named entities and their percentage over the whole tags in named entities (P_{entity}) and over the corresponding tags in the whole text (P_{text}). We can see that the top 4 POS tags in both of CoNLL03 and OntoNotes* are the same and they are NNP, JJ, NN, and NNPS. The P_{entity} of proper nouns (including NNP and NNPS) reaches more than 80.1%, and this indicates that named entities are mainly made up of proper nouns. The P_{text} of proper nouns reaches more than 84.8%, and this indicates that in the whole text, the proper nouns mainly appear in named entities.[9] Within named entities, a large portion of the JJ words are the words that indicate languages or nationalities (e.g., 'American,' 'British,' and 'Chinese'). The nationality words also hardly appear in common text.

Table 4: Percentage of named entities in length

Length	CoNLL03	CotoNotes5	OntoNotes*
1	63.19	54.86	60.08
2	31.40	21.99	24.53
3	3.83	13.08	8.31
4	1.05	5.85	4.39
≥5	0.53	4.22	2.57

Finding 3 *Entity length follows a family of power law distributions, with well-defined means and finite variances.*

Figure 2 plots the length distribution of named entities in log-log scale. We can see that the length of named entities from different datasets can be fitted by a family of power law distributions $p(x) = Cx^{-\alpha}$, in which the constant C is unimportant and the exponent α is of interest. The power law fits the length distribution of CoNLL03's named entities with $\alpha = 4.26$ (see Figure 2(a)), fits the one of OntoNotes5's with $\alpha = 4.34$ (see Figure 2(b)), and fits OntoNotes*'s with $\alpha = 4.09$ (see Figure 2(c)). The three α are all greater than 3, indicating that the three power law distributions have well-defined means and finite variances (if $\alpha > 3$, then the power law's first and second moments converge)[10] (Newman, 2005).

Table 4 presents the distribution of entity length in percentage. The percentage of one-word entities in the three datasets is 63.19%, 54.68%, and 60.08%, respectively; and named entities on average contain respective 1.45, 1.85, and 1.67 words.

[8] If consider the whole OntoNotes5 dataset, then named entities are mainly made up of proper nouns and cardinal numbers.

[9] The P_{text} of proper nouns does not reach 100% probably because individual dataset just concerns certain types of named entities.

[10] Moments $E(x^n) = \int_{x_0}^{\infty} x^n p(x) dx = \frac{C}{n+1-\alpha}\left[x^{-\alpha+n+1}\right]_{x_0}^{\infty}$

We find that one specific type of named entities might not follow a power law distribution, but when considering the whole types, their entity length tend to exhibit the characteristic of power law. For example, the PER in CoNLL03 contains 34.35% one-word entities and 62.34% two-word entities; this length distribution cannot be fitted by a power law. But for the whole entity types, the entity length follows a power law distribution.

The phenomena of non-common words and proper nouns in named entities can be explained by the process of annotation. An annotator annotates a sequence of word(s) as a named entity primarily because the annotator encounters in that sequence certain word(s) that should be treated as (part of) a named entity; and those words are mainly the proper nouns.

The phenomenon of power law in entity length can be explained by the principle of least effort (Zipf, 1949). Whenever we need an entity to communicate, on the premise of being able to make our ideas understood, we prefer to use a short one. In this sense, the distribution of entity length indicates the probability of the number of words we prefer to use in entity; Table 4 suggests that the probability of our preference for a one-word entity is about 60%. Similar to the preferential mechanism that results in scale-free power law property in random network (Barabasi and Albert, 1999), the preference for short entity also results in scale-free power law property in entity length.

To summarize, an average entity contains about two words, with at least one non-common word that is mainly a proper noun. To model the entity, we focus on modeling the non-common word.

4 POM: A Constituent-based Tagging Scheme for Named Entity Extraction

Figure 3 shows the overview of POM. POM mainly consists of three parts: POM scheme, word lexicons, and named entity modeling.

4.1 POM Scheme

POM scheme is a constituent-based tagging scheme and includes three tags: P, O, and M; they indicate the constituents of named entity, namely Predictor, Modifier, and the words Outside named entity. Finding 1 shows that non-common words are capable of predicting named entities and POM models the non-common-words as predictors, such as 'Africa,' 'Chicago,' and 'Chinese.' Modifiers modify predictors in named entities; for example, 'West' modifies 'Africa' in 'West Africa' and 'University' modifies 'Chicago' in 'Chicago University.' Predictor and modifier are defined on tokens and based on the constituent role that they play in named entities; their modification relation is not as strict as grammatical relation. In 'Chicago University,' for example, 'Chicago' is a predictor because itself alone indicates a named entity while 'University' is a modifier because itself alone does not indicate a named entity.

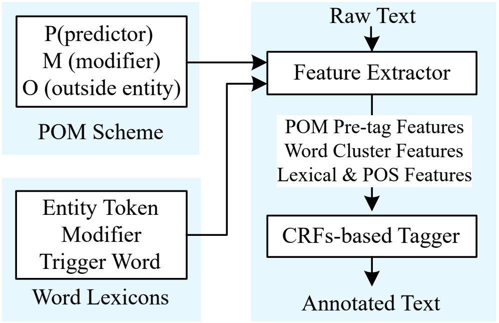


Figure 3: Overview of POM. Top-left shows the three-tag POM scheme. Bottom-left is the word lexicons, a set of entity-related words. Right-hand side shows the named entity modeling, with the help of word lexicons and POM scheme.

Table 5: Number of entity tokens and trigger words collected in POM

Category	#Entity Tokens	#Trigger words
PER	8,424	27
LOC	835	6
ORG	-	78
MISC	207	68

4.2 Word Lexicons

We collect word lexicons from the list provided by CoNLL03 shared task for PER and LOC, and from Wikipedia for LOC and MISC; the MISC word lexicons include only 207 nationality words. Such kind of word lexicons are called entity tokens. We also collect two kinds of modifiers: generic modifiers and trigger words. Generic modifiers can modify several types of named entities, such as 'of,' and 'south'; while trigger words modify certain type of named entities, such as 'Mr.' for PER, 'City' for LOC, 'Inc.' for ORG, and 'Cup' for MISC. The trigger words have the capability to predict named entities and determine their boundaries; for example, 'Mr.' appears outside named entities while 'City,' 'Inc.,' and 'Cup' inside. Unlike previous works (Kazama and Torisawa, 2007;

Ratinov and Roth, 2009) that use lexicons in word sequences, we use lexicons in words. For example, we do not use 'Chicago University' but use 'Chicago' and 'University.' Table 5 summarizes the number of the collected entity tokens and trigger words; the generic modifiers include 18 words. In experiments, POM uses the same word lexicons for CoNLL03 and OntoNotes* datasets.

4.3 Named Entity Modeling

Named entity modeling consists of two parts: feature extraction and model learning and tagging.

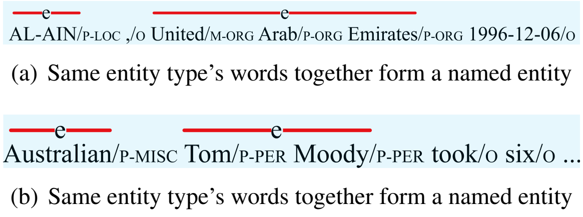
4.3.1 Feature Extraction

The features we extract include three kinds: POM pre-tag features, word cluster features, and lexical & POS features. During feature extraction, the ith word in text is denoted by w_i.

POM Pre-tag Features. POM pre-tag features is designed to leverage the information of the non-common words under POM scheme. Specifically, a word is pre-tagged by P if it satisfies two conditions: (1) the word hardly appears in the common text of training set and the threshold of hardly appearing for each dataset is same to the one set in Section 3.2; (2) the word has a POS tag of NNP* (i.e., NNP or NNPS) or is matched by the entity tokens or is hyphenized by at least one entity token (e.g., 'U.S.-based' and 'English-oriented'). The words that are matched by modifiers are pre-tagged by M. Other words are pre-tagged by O.

Word Cluster Features. Word clusters have shown their effectiveness in improving the coverage in many information extraction tasks (Liang, 2005; Ratinov and Roth, 2009; Ritter et al., 2011; Owoputi et al., 2013). We follow previous works to derive the prefix paths of 4, 8, and 12 bits from a hierarchical word clusters as features for a word. In implementation, we use publicly available bllip-clusters[11] for CoNLL03 and use Liang (2005)'s implementation[12] of Brown clustering algorithm (Brown et al., 1992) to train the word clusters on OntoNotes 5.0 dataset (Pradhan et al., 2013) for OntoNotes*.

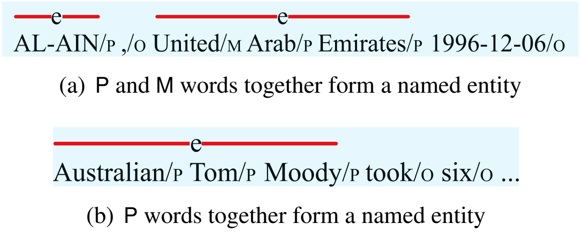
Lexical & POS Features. The lexical & POS features include (1) the word w_i itself, its lowercase, and its lemma; (2) if it is initial capitalized and if it is the beginning of a sentence; and (3) POS tags.


e
e
AL-AIN/P-LOC ,/O United/M-ORG Arab/P-ORG Emirates/P-ORG 1996-12-06/O
e
e
Australian/P-MISC Tom/P-PER Moody/P-PER took/O six/O ...

(a) Same entity type's words together form a named entity

(b) Same entity type's words together form a named entity

Figure 4: Examples of named entity extraction using entity types. Label e indicates named entity.


e
e
AL-AIN/P ,/O United/M Arab/P Emirates/P 1996-12-06/O
e
Australian/P Tom/P Moody/P took/O six/O ...

(a) P and M words together form a named entity

(b) P words together form a named entity

Figure 5: Examples of named entity extraction without using entity types.

For the POM pre-tag and lexical & POS features, we extract them for w_i in a 5-word window, namely the features of w_{i-2}, w_{i-1}, w_i, w_{i+1}, and w_{i+2}. For the word cluster features, we consider them for only the current w_i.

Labeling Tag Assignment. We use POM scheme as our labeling tags. The words outside named entity are assigned with O. Within named entities, a word that has a POS tag of NNP* or is matched by entity tokens (including those with hyphens) is assigned with P; otherwise it is assigned with M.

In feature extraction, w_i is represented by a feature vector ($\mathbf{x_i}$, y_i), in which the vector $\mathbf{x_i}$ is derived from the features and the label y_i is from the labeling tags. POM scheme is used as a kind of pre-tag features in feature extraction and as labeling tags for sequence tagging.

4.3.2 Model Learning and Tagging

POM models named entities under a framework of CRFs (Lafferty et al., 2001) with the feature vectors. In implementation, we use Stanford Tagger[13] to get the word lemma and POS tags and use CRFSuite[14] with default parameters to learn the model and tag the sequence. In sequence tagging, POM assigns each word with one of POM tags.

Although we focus on named entity extraction, the baseline methods (see Section 5.1) treat named entity extraction and classification as an end-to-end task. For fair comparison, we include the en-

[11] http://people.csail.mit.edu/maestro/papers/bllip-clusters.gz

[12] https://github.com/percyliang/brown-cluster

[13] http://nlp.stanford.edu/software/tagger.shtml

[14] http://www.chokkan.org/software/crfsuite/

Table 6: Performance of POM and baseline methods on CoNLL03 and OntoNotes* datasets

Dataset	Method	Dev.			Test		
		$Pr.$	$Re.$	F_1	$Pr.$	$Re.$	F_1
CoNLL03	StanfordNER	**97.15**	95.88	96.51	95.05	93.41	94.22
	LSTM-CRF	96.77	96.92	96.84	**95.38**	94.44	94.91
	POM	96.71	**97.13**	**96.91**	95.27	**94.72**	**94.99**
OntoNotes*	StanfordNER	93.19	91.57	92.38	**93.58**	91.45	92.50
	LSTM-CRF	-	-	-	-	-	-
	POM	**93.25**	**91.96**	**92.60**	93.47	**92.79**	**93.13**

tity types in model learning and tagging but report only the extraction results and leave the model without using entity types to the factor analysis.

Named Entity Extraction. After tagging, we extract named entities from the tagged sequences. For the model using entity types, those words that appear together and are tagged with same entity type form a named entity. See Figure 4 for two examples. For the model without entity types, those P and M words that appear together are extracted as a named entity. See Figure 5 for examples.

5 Experiments

We evaluate POM on CoNLL03 and OntoNotes* datasets against two representative state-of-the-art methods, StanfordNER and LSTM-CRF.

5.1 Experiment Setting

Datasets. The datasets used in our experiments include CoNLL03 and OntoNotes* and they are detailed in Section 3.1.

Baseline Methods. Our state-of-the-art baseline methods include StanfordNER (Finkel et al., 2005) and LSTM-CRF (Lample et al., 2016). StanfordNER derives handcrafted features under CRFs with IO scheme (Inside-Outside) and is widely used in other works. LSTM-CRF uses automatic features learned by bidirectional long short-term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) under CRFs with IOBES scheme. We use StanfordNER as the representative of the traditional methods and use LSTM-CRF as the representative of the neural network based methods that achieve similar results in the end-to-end NER task on CoNLL03 dataset; specifically, around 94.5% of F_1 on the development set and around 91.2% on the test set (Lample et al., 2016; Ma and Hovy, 2016; Chiu and Nichols, 2016; Xu et al., 2017; Liu et al., 2018).

In implementation, for StanfordNER, we use its 3.8.0 version and report the results on CoNLL03 by the model it provides, and use its source code to report its performance on OntoNotes*.[15] For LSTM-CRF, we use its provided model to report the results on CoNLL03, and use its source code[16] (trying) to report its performance on OntoNotes*.

Evaluation Metrics. We use the evaluation toolkit[17] of CoNLL03 shared task (Sang and Meulder, 2003) to report the results under the three traditional metrics: $Precision$, $Recall$, and F_1.

5.2 Experiment Results

5.2.1 Overall performance

Table 6 reports the overall performance of POM and the baseline methods on CoNLL03 and OntoNotes* datasets.[18] POM achieves better performance than StanfordNER and LSTM-CRF on CoNLL03 and achieves better performance than StanfordNER on OntoNotes*. In the perspective of error reduction, POM reduces 2.9% to 13.3% of errors compared with StanfordNER. That is because besides some standard features, POM makes full use of the non-common words.

The surprising result is that POM, StanfordNER, and LSTM-CRF achieve comparable results on CoNLL03 in named entity extraction; although in the end-to-end NER task, LSTM-CRF significantly outperforms StanfordNER. LSTM-CRF achieves 90.94% of F_1 in the NER task on CoNLL03's test set (Lample et al., 2016) while StanfordNER achieves only 86.86% (Finkel et al., 2005). This result indicates that simple handcrafted feature methods perform comparably with

[15] We use its default parameter setting except disuse the features of 'useSequences' and 'usePrevSequences' for saving memory. This setting training on CoNLL03 gets similar results compared with the provided model.

[16] `https://github.com/glample/tagger`

[17] `http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt`

[18] By the deadline of submission, LSTM-CRF still does not finish running on the OntoNotes* dataset, probably because the corpus size of OntoNotes* is significantly larger than the one of CoNLL03.

Table 7: Impact of factors. 'BIO' indicates the system that replace POM labeling tags by BIO tags. '−' indicates the kind of features or entity types that are removed from POM.

Dataset	Method	Dev.			Test		
		Pr.	*Re.*	F_1	*Pr.*	*Re.*	F_1
CoNLL03	POM	**96.71**	**97.13**	**96.91**	**95.27**	**94.72**	**94.99**
	BIO	95.66	95.62	95.64	93.12	93.52	93.32
	− POM Pre-tag	95.56	93.47	94.50	93.44	90.81	92.11
	− Word Cluster	95.23	94.77	95.00	92.97	92.78	92.87
	− Lexical & POS	93.07	91.94	92.50	91.25	90.47	90.86
	− Entity Types	95.31	95.42	95.36	94.96	94.43	94.69
OntoNotes*	POM	**93.25**	**91.96**	**92.60**	93.47	**92.79**	**93.13**
	BIO	92.60	90.93	91.75	**93.64**	90.63	92.11
	− POM PreTag	92.62	89.95	91.27	93.08	89.54	91.28
	− Word Cluster	92.63	89.86	91.23	93.00	89.94	91.45
	− Lexical & POS	85.30	80.69	82.93	85.60	80.81	83.14
	− Entity Types	92.10	90.48	91.28	92.36	91.87	92.11

Table 8: Runtime that POM and baselines requre to go through a whole process (unit: minutes)

Method	CoNLL03	OntoNotes*
StanfordNER	5	103
LSTM-CRF	3,240	-
POM	4	74

neural network based methods on the extraction task and suggests that the features learned by neural networks cannot provide more useful information for the extraction task. We should treat the extraction and classification tasks separately.

5.2.2 Computational Efficiency

Table 8 reports the runtime that POM and baseline methods require to go through a whole process (including training and test) on the two datasets on a Mac OS laptop (1.4GHz Processor and 8GB Memory). We can see that POM is more efficient than StanfordNER because POM refines the significant features. POM and StanfordNER are much much more efficient than LSTM-CRF.

5.2.3 Factor Analysis

We conduct controlled experiments to analyze the impact of POM scheme as labeling tags as well as the features and entity types that are used in POM. The results are presented in Table 7.

Impact of POM Labeling Tags. To analyze the impact of POM scheme as labeling tags, we replace the POM labeling tags by BIO tags (as well as by IOBES tags and IO tags) and keep other factors unchanged. IO, BIO, and IOBES schemes achieve similar results and Table 7 reports the result of BIO scheme as a representative. We can see that POM performs better than BIO, because POM scheme overcomes the problem of inconsistent tag assignment (Zhong and Cambria, 2018).

Impact of Features. To analyze the impact of features, we remove each of them from TOMN at a time. After POM pre-tag features are removed, the performance is affected in all the measures, with absolute decreases of 1.3% to 2.8% in F_1. That means POM pre-tag features improve the performance and validates the predictive power of the non-common words. Like previous works, word cluster features and lexical & POS features improve the performance on the extraction task.

Impact of Entity Types. Entity types also improve the extraction performance, due to its function as separator to separate consecutive named entities. Some named entities appear consecutively; without separator, they are treated as a named entity. Figure 5(b) shows such an example.

6 Conclusion and Future Work

We analyze two benchmark datasets and summarize three common characteristics about named entities. The characteristics drive us to design a learning method with a constituent-based tagging scheme to extract named entities from free text. Experiments show the effectiveness of our method against representative state-of-the-art methods. In the future we will try to answer the central question of classification: what differentiate certain categories of named entities from each other?

References

Masayuki Asahara and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 8–15.

Albert-Laszlo Barabasi and Reka Albert. 1999. Emergence of scaling in random networks. *Science*, 286:509–512.

Daniel M. Bikel, Scott L. Miller, Richard M. Schwartz, and Ralph M. Weischedel. 1997. Nymble: A high-performance learning name-finder. In *Proceedings of the fifth Conference on Applied Natural Language Processing*, pages 194–201.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Nyu: Description of the mene named entity system as used in muc-7. In *Proceedings of the 7th Message Understanding Conference*.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Hsin-Hsi Chen and Jen-Chang Lee. 1996. Identification and classification of proper nouns in chinese texts. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 1, pages 222–229.

Nancy A. Chinchor. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference*, volume 29.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Machael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Michael Collins. 2002. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 489–496.

Ronan Collobert, Jason Weston, Leon Bottou, Machael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(2493-2537).

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference - 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics*.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. In *https://arxiv.org/abs/1508.01991v1*.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1158.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. Overview of the chemical compound and drug name recognition (chemdner) task. In *BioCreative Challenge Evaluation Workshop*, volume 2, pages 2–33.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning*, pages 281–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architecture for named entity recognition. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 260–270.

Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology.

Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015a. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530.

Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015b. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence*.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank F. Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Proceedings of the 32nd AAAI Conference on Artifical Intelligence*.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint named entity recognition and disambiguation. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing*, pages 879–888.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Benoit Mandelbrot. 1961. On the theory of word frequencies and on related markovian models of discourse. *Structure of Language and its Mathematical Aspects*, 12:190–219.

Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham, and Yorick Wilks. 2001. Named entity recognition from diverse text types. In *Proceedings of 2001 Recent Advances in Natural Language Processing Conference*, pages 257–274.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the 7th Conference on Computational Natural Language Learning*.

Einat Minkov, Richard C. Wang, and William W. Cohen. 2005. Extracting personal names from email: Applying named entity recognition to informal text. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 443–450.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Ndapandula Nakashole, Tomasz Tylenda, and Gerhard Weikum. 2013. Fine-grained semantic typing of emerging entities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1488–1497.

Mark EJ. Newman. 2005. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT 2013*, pages 380–390.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of the 8th Conference on Computational Language Learning*, pages 78–86.

Thierry Poibeau and Leila Kosseim. 2001. Proper name extraction from non-journalistic texts. *Language and Computers*, 37:144–157.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeny Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of Workshop on Semantic Evaluation 2016*, pages 19–30.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Bjorkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the 7th Conference on Computational Natural Language Learning*, pages 143–152.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155.

Yael Ravin and Nina Wacholder. 1997. Extracting names from natural-language text. Technical report, IBM Research Division.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147.

Cicero Nogueira Dos Santos and Victor Guimaraes. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the 5th Named Entities Workshop*, pages 25–33.

Satoshi Sekine. 1998. Nyu: Description of the japanese ne system used for met-2. In *Proceedings of the 7th Message Understanding Conference*.

Bengt Sigurd, Mats Eeg-Olofsson, and Joost van de Weijer. 2004. Word length, sentence length and frequency - zipf revisited. *Studia Linguistica*, 58(1):37–52.

Joaquim Ferreira Da Silva, Zornitsa Kozareva, and Jose Gabriel Pereira Lopes. 2004. Cluster analysis and classification of named entities. In *Proceedings of the 2004 Conference on Language Resources and Evaluation*.

Koichi Takeuchi and Nigel Collier. 2005. Bio-medical entity extraction using support vector machines. *Artificial Intelligence In Medicine*, 33(2):125–137.

Liang-Jyh Wang, Wei-Chuan Li, and Chao-Huang Chang. 1992. Recognizing unregistered names for mandarin word identification. In *Proceedings of the 14th Conference on Computational Linguistics*, volume 4, pages 1239–1243.

Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A local detection approach for named entity recognition and mention detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1237–1247.

Shihong Yu, Shuanhu Bai, and Paul S. Wu. 1998. Description of the kent ridge digital labs system used for muc-7. In *Proceedings of the 7th Message Understanding Conference*.

Xiaoshi Zhong and Erik Cambria. 2018. Time expression recognition using a constituent-based tagging scheme. In *Proceedings of The Web Conference 2018*.

George Zipf. 1936. *The Psychobiology of Language*. London: Routledge.

George Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Inc.