THE JOHNS HOPKINS UNIVERSITY

human language technology
center of excellence

Nerit: Named Entity Recognition for Informal Text

**David Etter, Francis Ferraro, Ryan Cotterell,
Olivia Buzek, and Benjamin Van Durme**

TECHNICAL REPORT 11

JULY 5, 2013

Acknowledgements This work is supported, in part, by the Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

HLTCOE
810 Wyman Park Drive
Baltimore, Maryland 21211
http://hltcoe.jhu.edu

Nerit: Named Entity Recognition for Informal Text

David Etter
Department of Computer Science
George Mason University
detter@gmu.edu

Francis Ferraro and **Ryan Cotterell** and **Olivia Buzek** and **Benjamin Van Durme**
Human Language Technology Center of Excellence
Johns Hopkins University
ferraro@cs.jhu.edu, ryan.cotterell@jhu.edu, obuzek@cs.jhu.edu, vandurme@cs.jhu.edu

Abstract

We describe a multilingual named entity recognition system using language independent feature templates, designed for processing short, informal media arising from Twitter and other microblogging services. We crowdsource the annotation of tens of thousands of English and Spanish tweets and present classification results on this resource.

1 Introduction

Named entity recognition (NER) – the task of identifying and labeling salient mentions with a set of predefined tags such as PERSON, LOCATION, or ORGANIZATION — is a core task in information extraction and has motivated significant research into statistical models and the development of discriminative features (Grishman and Sundheim, 1996; Nadeau and Sekine, 2007). NER research has culminated in strong systems for formal newswire text, but recent work (Liu et al., 2011) has quantitatively shown these systems brittle to informal text, reporting a 45% decrease in F-Score when applying a CoNLL trained NER model to English tweets. Figure 1 illustrates some of the difficulties, such as novel vocabulary and inadherence to "standard" grammar rules.

Informal text presents additional problems. First, the number and diversity of people tweeting has created a vast multilingual dynamic source of open-domain text, necessitating the use of language independent features and resources. Second, generating annotated data for NER has traditionally been expensive; the lack of annotated data hinders progress in this domain.

To address the challenges of multilingual informal text, we propose a simple multilingual NER system that leverages previously validated techniques.

2 Related Research

Research on NER for Twitter is a recent direction for the information extraction community. To the best of our knowledge, there is no published work on Spanish Twitter NER and only a small sample of English based research. Ritter et al. (2011) developed an English POS tagger and Named Entity Chunker for Twitter using both in-domain and out-of-domain data. Their NER system uses output from a POS tagger, along with Labeled LDA (Ramage et al., 2009) over a FreeBase dictionary, to apply distant supervision. The system outperforms baseline formal text trained models on 2400 English annotated messages using 10 entity types. Locke and Martin (2009) applied microtext clustering to contextually linked tweets to improve NER performance over the four types: `Person, Location, Organization, and Digital ID`. Liu et al. (2012) proposed a joint NER and Named Entity Normalization (NEN) approach for English tweets.

Liu et al. (2011) proposed a semi-supervised approach to NER on English Tweets. Their system incorporates a two stage approach to extract both local and global features. The initial KNN model clusters similar tweets and assigns cluster level labels which are provided as features to the second stage CRF sequence tagging process. Experiments were conducted on over 12,000 annotated tweets using the entity types `Person, Product, Location, and Organization`. Li et al. (2012) proposes an unsupervised approach to Twitter NER. Their system leverages out of domain data from

@lopezdoriga1	para	la	twitteratura	son	idels	los	poeminimos	de	efraín	huerta
O	O	O	O	O	O	O	O	O	B-PER	I-PER

@Lopezdoriga1 for the twitterature, the poetweets by Efraín Huerta are ideal.

Figure 1: Example annotated tweet, with translation, illustrating many of the difficulties of working with multilingual informal text, such as novel words ("poeminimos") and misspellings ("idels").

Wikipedia and the Web to segment candidate entities and then performs a random walk to rank the entities. Finin et al. (2010) investigated the use of Crowdsourcing with MTurk and Crowdflower to annotate Named Entities in Twitter.

In contrast to the existing Twitter NER research, our system emphasizes Spanish data with over 30 thousand newly annotated messages and does not rely on language dependent features, such as dictionaries or POS tagging.

There has been a significant amount of Named Entity research on formal text. The Message Understanding Conference 6 (MUC6) (Grishman and Sundheim, 1996) provided one of the first formal evaluations for the English Named Entity task using the North American News Text Corpora. Ratinov and Roth (2009) evaluated the use of non-local features and external knowledge over the formal texts from MUC7 and CoNLL2003. Their work analyzed a number of class encoding approaches and found that BILOU significant outperformed BIO for that set of formal data. Nadeau and Sekine (2007) provides a detailed review of the types of features used by traditional NER systems. They organize the list of features into the categories: word level, list, and document and corpus. Word level includes features such as part-of-speech, morphology, punctuation, and case. The list category refers to gazetteer type features such as general dictionaries or derived list of organization names and locations. Document and corpus includes features such as meta information, word frequency, and position of the word in a sentence or document. Peng et al. (2003) studied the problem of language independent text categorization using character level n-gram language modeling.

The CoNLL Named Entity shared tasks have included a number of language independent evaluations. CoNLL 2002 (Tjong Kim Sang and De Meulder, 2003) included formal text for Dutch and Spanish and the 2003 task (Tjong Kim Sang and De Meulder, 2003) consisted of English and German. Kozareva et al. (2005) applied semi-supervised techniques such as self-training and co-training to unlabeled data in their work on Spanish Named Entity recognition. The work in Richman and Schone (2008) uses Wikipedia and its structural meta-data to generate large language specific annotated data sets which are evaluated on Spanish, French, and Ukrainian truth sets.

3 Detailed Approach

Our approach uses language independent features to jointly model the segmentation and classification tasks. We focus development on a small set features that are derived from the character composition of a word and its context in a message. These features have proved to be robust across different languages (Grishman and Sundheim, 1996; Nadeau and Sekine, 2007) and do not rely on part-of-speech taggers or gazetteers. When such resources are available for a target language they might be incorporated into our system for further gains in accuracy, however here we limit our study to the rapid development of data in a new language, and then constructing a system that should be portable to low-resource languages.

Table 1 presents all feature templates used in our experiments. The baseline feature is the word itself. Character n-grams features (Klein et al., 2003; Nadeau and Sekine, 2007) are frequently used in statistical natural language processing for tasks such as spam detection, speech recognition, and sequence searching. They provide a language independent approach to extract a token's root, prefix, or suffix. Compared to full tokens, they allow for a (limited) modeling of morphology and robustness to spelling errors.

The goal of a context feature is to identify common word patterns to the left or right of a named entity that may help to identify similar entities of that type. An English context example is the phrase *located in*, which provides a pattern to identify the entity type `Location`. Our feature extraction algorithm captures patterns of length $\pm n$ on each side of the current word, where n is a parameter to the system.

Feature	Description
Token	Binary identifier for current word
N-gram	character n-grams for current word
Context	$\pm k$ words to the left
Length	bin message length, bin word length
Position	bin word position in message

Table 1: All feature templates used in our language independent model.

Model	Precision	Recall	F-Score
word only	.71	.20	.31
3-gram	.68	.50	.58
7-gram	.68	.50	.58
$\pm$1 context	.76	.26	.39
$\pm$3 context	.70	.28	.40
7-gram $\pm$1 context	.70	.62	.66

Table 2: Spanish Feature Analysis

3.1 Model

We use a structural Support Vector Machine at training and test time.[1] This algorithm combines the discriminative learning of an SVM with a Hidden Markov Model (HMM). Altun et al. (2003) successfully applied the SVM-HMM to sequence tagging problems such as Named Entity Recognition and Part of Speech Tagging for newswire.

4 Experiments

We evaluate our system on the English-only dataset from Ritter et al. (2011), and a newly annotated annotated dataset of 29,056 Spanish and 9,609 English tweets.

We used Amazon Mechanical Turk over two weeks to annotate these fresh 40,000 tweets with 10 entity types; Figure 2 provides a breakdown of the entity distributions.

4.1 Features

Experiments for both Spanish and English were performed with a 90-10 train test split of our annotated set. To analyze the contribution of each feature to the overall model, we performed a set of experiments where each feature is evaluated independently. Table 2 shows the Spanish results of each single feature model and the combined feature models.

The baseline model is a word feature only model, consisting of a binary feature vector identifying the current word. The results in Table 2 and 3 show that the word model produces good precision but low recall. This type of model essentially memorizes words from the training set. An example is the named entity *febrero*, from the test set message, *que rapido te estas yendo febrero*. This entity is found 12 times in our training set and identified as an entity in 11 of those instances.

Sliding context window experiments were performed using $\pm$1 and $\pm$3 words to the left and right of the current token. The top sliding context window model shows a 12 point increase in F-Score over the baseline.

Model	Precision	Recall	F-Score
word only	.73	.26	.38
3-gram	.67	.45	.54
7-gram	.64	.46	.54
+-1 context	.74	.37	.50
+-3 context	.76	.34	.47
7-gram +-1 context	.71	.49	.58

Table 3: English Feature Analysis

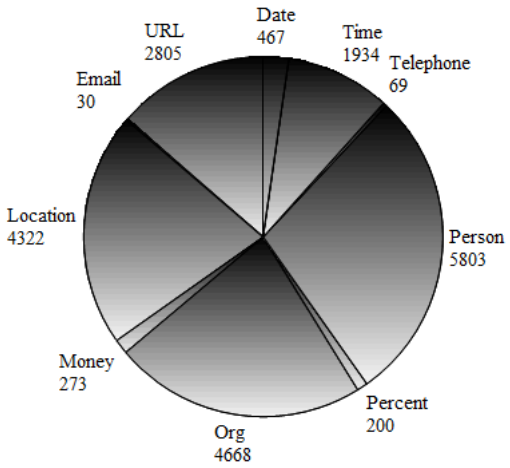
Character n-gram experiments were performed with lengths of 3 and 7 characters. In this series of experiments the n-gram includes all lengths $\leq n$. The n-gram models move beyond simple word memorization by identifying the root, prefix, and suffix of a word. The character n-gram models shows a 27 point F-Score increase over the baseline Spanish word model and a 16 point F-Score increase over the baseline English word model.

The hybrid models combine word, n-gram, and context features to create a comprehensive language independent feature set. These models show a 20 to 30 point increase over the baseline word model, while maintaining high precision. Tables 4 and 5 show a break down of the Precision, Recall, and F-Score for the 10 Spanish and English types.

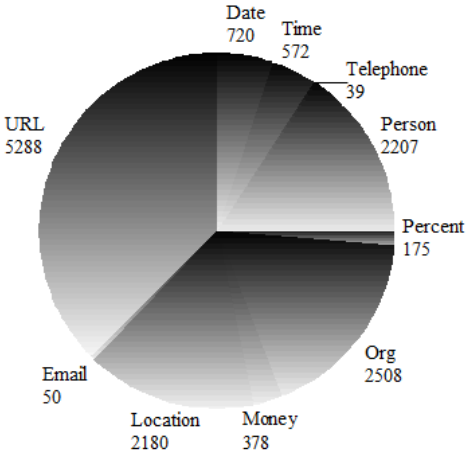
Type	Precision	Recall	F-Score
Date	.73	.70	.71
Email	.50	.13	.20
Location	.76	.64	.70
Money	.48	.63	.54
Organization	.63	.41	.50
Percent	.89	.99	.94
Person	.75	.60	.66
Telephone	.43	.75	.55
Time	.65	.75	.70
URL	.66	.80	.72

Table 4: Precision, recall and F1 for 10 types on Spanish data.

[1] www.cs.cornell.edu/people/tj/svm_light



(a) Spanish type distribution over 20,751 entities.



(b) English type distribution over 14,117 entities.

Figure 2: Distribution of Entity Types for the new Spanish (left) and English (right) datasets.

Type	Precision	Recall	F-Score
Date	.75	.75	.75
Email	.79	.79	.79
Location	.71	.62	.66
Money	.81	.71	.76
Organization	.65	.39	.49
Percent	.82	.72	.77
Person	.74	.37	.49
Telephone	.50	.13	.20
Time	.71	.37	.49
URL	.98	.95	.96

Table 5: Precision, recall and F1 for 10 types on new English data.

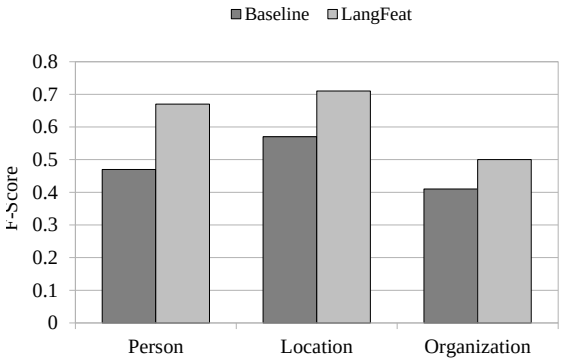


Figure 3: CoNLL Type Comparison

4.2 Model Comparisons

Figure 3 provides a comparison of our feature rich approach to existing formal text models. We compare a standard out of the box HMM trained on the CoNLL-2002 Spanish set with our informal model. The results clearly show that the out of domain Spanish models, trained on formal data, are not able to handle the informal nature of Twitter messages.

Table 6 provides the results of our model trained and tested using the 10 entity types from the English Twitter data set of Ritter et al. (2011). Table 7 shows that our system outperforms Ritter et al on this data set.

4.3 Learning Curve

A question raised in every named entity task is how much annotated data do we need. The learning curve experiments attempt to answer that question for our Spanish data set. In this series of runs we build models over an increasingly large chunk

Type	Precision	Recall	F-Score
company	.54	.41	.46
facility	.60	.33	.43
geo-loc	.66	.53	.59
movie	.50	.25	.33
musicartist	.75	.37	.50
other	.46	.29	.36
person	.75	.54	.63
product	.50	.36	.42
sportsteam	.88	.70	.78
tvshow	.50	.50	.50

Table 6: Precision, recall and F1 for the 10 English types of Ritter et al. (2011).

Model	F-Score
Ritter et al. (2011)	.51
Multi-Model	**.54**

Table 7: English results on Ritter et al. (2011)'s data.

of the original training set and evaluate each model with the original test set. Table 4 shows our learning curve results which begin with 10 percent of the original training data and incrementally build to the entire training set. The results show the initial sharp F-Score increase from 10 to 40 percent of the training data and then a slow but steady increase as we add the remaining annotations.

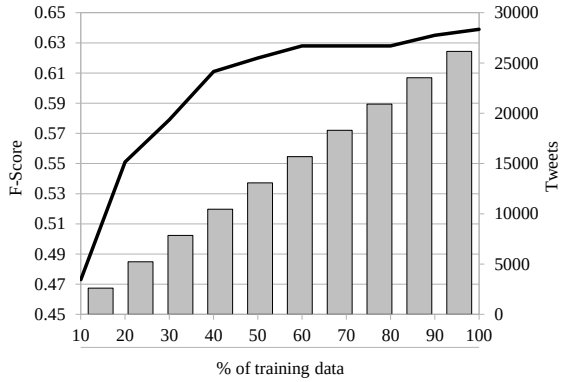


Figure 4: Learning curve for our Spanish system (black line). Number of tweets (bars) are provided for reference.

5 Conclusion

This work proposes novel research for Multilingual Named Entity Recognition on informal Twitter data. We annotate a large multilingual corpus of over 30 thousand Spanish and 10 thousand English tweets using the Amazon Mechanical Turk crowdsourcing marketplace. This large annotated corpus of informal text includes 10 entity types and is made available to the information extraction community for research. Our system extracts a set of language independent features which do not rely on part-of-speech taggers or gazetteers. The evaluation of our system on both Spanish and English Twitter messages shows significant improvements over formal text trained models.

Acknowledgments This work was performed as part of the 2012 Summer Camp for Applied Language Exploration (SCALE) at the Human Language Technology Center of Excellence (HLTCOE). Thanks to the staff at the HLTCOE, and SCALE leads: Edward Loper, Jason Duncan and Benjamin Van Durme.

References

Y. Altun, I. Tsochantaridis, T. Hofmann, et al. 2003. Hidden markov support vector machines. In *International Conference on Machine Learning*, volume 20, page 3.

T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics.

R. Grishman and B. Sundheim. 1996. Message understanding conference-6: A brief history. In *Proceedings of COLING*, volume 96, pages 466–471.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 180–183. Association for Computational Linguistics.

Z. Kozareva, B. Bonev, and A. Montoyo. 2005. Self-training and co-training applied to spanish named entity recognition. *MICAI 2005: Advances in Artificial Intelligence*, pages 770–779.

Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. 2012. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 721–730, New York, NY, USA. ACM.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 359–367, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. 2012. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 526–535, Stroudsburg, PA, USA. Association for Computational Linguistics.

B. Locke and J. Martin. 2009. Named entity recognition: Adapting to microblogging. *Senior Thesis, University of Colorado*.

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

F. Peng, D. Schuurmans, and S. Wang. 2003. Language and task independent text categorization with simple language models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 110–117. Association for Computational Linguistics.

D. Ramage, D. Hall, R. Nallapati, and C.D. Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics.

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

A.E. Richman and P. Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1–9.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

E.F. Tjong Kim Sang and F. De Meulder. 2003. 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*, pages 142–147. Association for Computational Linguistics.