Making Sense of Microposts (#MSM2013) Concept Extraction Challenge

Amparo Elizabeth Cano Basave[1⋆], Andrea Varga[2], Matthew Rowe[3], Milan Stankovic[4], and Aba-Sah Dadzie[2⋆⋆]

[1] KMi, The Open University, Milton Keynes, UK
[2] The OAK Group, Dept. of Computer Science, The University of Sheffield, UK
[3] School of Computing and Communications, Lancaster University, UK
[4] Sépage, Paris, France
a.cano_basave@aston.ac.uk,a.varga@dcs.shef.ac.uk
m.rowe@lancaster.ac.uk,milstan@gmail.com,a.dadzie@cs.bham.ac.uk

Abstract. Microposts are small fragments of social media content that have been published using a lightweight paradigm (e.g. Tweets, Facebook likes, foursquare check-ins). Microposts have been used for a variety of applications (e.g., sentiment analysis, opinion mining, trend analysis), by gleaning useful information, often using third-party concept extraction tools. There has been very large uptake of such tools in the last few years, along with the creation and adoption of new methods for concept extraction. However, the evaluation of such efforts has been largely consigned to document corpora (e.g. news articles), questioning the suitability of concept extraction tools and methods for Micropost data. This report describes the Making Sense of Microposts Workshop (#MSM2013) Concept Extraction Challenge, hosted in conjunction with the 2013 World Wide Web conference (WWW'13). The Challenge dataset comprised a manually annotated training corpus of Microposts and an unlabelled test corpus. Participants were set the task of engineering a concept extraction system for a defined set of concepts. Out of a total of 22 complete submissions 13 were accepted for presentation at the workshop; the submissions covered methods ranging from sequence mining algorithms for attribute extraction to part-of-speech tagging for Micropost cleaning and rule-based and discriminative models for token classification. In this report we describe the evaluation process and explain the performance of different approaches in different contexts.

1 Introduction

Since the first Making Sense of Microposts (#MSM) workshop at the Extended Semantic Web Conference in 2011 through to the most recent workshop in 2013

⋆ A.E. Cano Basave has since changed affiliation, to: Engineering and Applied Science, Aston University, Birmingham, UK (e-mail as above).

⋆⋆ A.-S. Dadzie has since changed affiliation, to: School of Computer Science, University of Birmingham, Edgbaston, Birmingham, UK (e-mail as above).

we have received over 60 submissions covering a wide range of topics related to interpreting Microposts and (re)using the knowledge content of Microposts. One central theme that has run through such work has been the need to understand and learn from Microposts (social network-based posts that are small in size and published using minimal effort from a variety of applications and on different devices), so that such information, given its public availability and ease of retrieval, can be reused in different applications and contexts (e.g. music recommendation, social bots, news feeds). Such usage often requires identifying entities or concepts in Microposts, and extracting them accordingly. However this can be hindered by:

(i) the noisy lexical nature of Microposts, where terminology differs between users when referring to the same thing and abbreviations are commonplace;
(ii) the limited length of Microposts, which restricts the contextual information and cues that are available in normal document corpora.

The exponential increase in the rate of publication and availability of Microposts (Tweets, FourSquare check-ins, Facebook status updates, etc.), and applications used to generate them, has led to an increase in the use of third-party entity extraction APIs and tools. These function by taking as input a given text, identifying entities within them, and extracting entity type-value tuples. Rizzo & Troncy [12] evaluated the performance of entity extraction APIs over news corpora, assessing the performance of extraction and entity disambiguation. This work has been invaluable in providing a reference point for judging the performance of extraction APIs over well-structured news data. However, an assessment of the performance of extraction APIs over Microposts has yet to be performed.

This prompted the Concept Extraction Challenge held as part of the *Making Sense of Microposts Workshop (#MSM2013)* at the *2013 World Wide Web Conference (WWW'13)*. The rationale behind this was that such a challenge, in an open and competitive environment, would encourage and advance novel, improved approaches to extracting concepts from Microposts. This report describes the #MSM2013 Concept Extraction Challenge, collaborative annotation of the corpus of Microposts and our evaluation of the performance of each submission. We also describe the approaches taken in the systems entered – using both established and developing alternative approaches to concept extraction, how well they performed, and how system performance differed across concepts. The resulting body of work has implications for researchers interested in the task of extracting information from social data, and for application designers and engineers who wish to harvest information from Microposts for their own applications.

2 The Challenge

We begin by describing the goal of the challenge and the task set, and the process we followed to generate the corpus of Microposts. We conclude this section with the list of submissions accepted.

2.1 The Task and Goal

The challenge required participants to build semi-automated systems to identify concepts within Microposts and extract matching entity types for each concept identified, where *concepts* are defined as abstract notions of *things*. In order to focus the challenge we restricted the classification to four entity types:

(i) Person **PER**, e.g. Obama;
(ii) Organisation **ORG**, e.g. NASA;
(iii) Location **LOC**, e.g. New York;
(iv) Miscellaneous **MISC**, consisting of the following: film/movie, entertainment award event, political event, programming language, sporting event and TV show.

Submissions were required to recognise these entity types within each Micropost, and extract the corresponding entity type-value tuples from the Micropost. Consider the following example, taken from our annotated corpus:

```
870,000 people in canada depend on #foodbanks
-25% increase in the last 2 years - please give generously
```

The fourth token in this Micropost refers to the location *Canada*; an entry to the challenge would be required to spot this token and extract it as an annotation, as:

```
LOC/canada;
```

The complete description of concept types and their scope, and additional examples can be found on the challenge website[5], and also in the appendices in the challenge proceedings.

To encourage competitiveness we solicited sponsorship for the winning submission. This was provided by the online auctioning web site eBay[6], who offered a $1500 prize for the winning entry. This generous sponsorship is testimony to the growing industry interest in issues related to automatic understanding of short, predominantly textual posts – Microposts; challenges faced by major Social Web and other web sites, and increasingly, marketing and consumer analysts and customer support across industry, government, state and not-for-profits organisations around the world.

2.2 Data Collection and Annotation

The dataset consists of the message fields of each of 4341 manually annotated Microposts, on a variety of topics, including comments on the news and politics, collected from the end of 2010 to the beginning of 2011, with a 60% / 40% split between training and test data. The annotation of each Micropost in the training dataset gave all participants a common base from which to learn extraction patterns. The test dataset contained no annotations; the challenge task was for

[5] `http://oak.dcs.shef.ac.uk/msm2013/challenge.html`
[6] `http://www.ebay.com`

participants to provide these. The complete dataset, including a list of changes and the gold standard, is available on the #MSM2013 challenge web pages[7], accessible under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

To assess the performance of the submissions we used an underlying *ground truth*, or *gold standard*. In the first instance, the dataset was annotated by two of the authors of this report. Subsequent to this we logged corrections to the annotations in the training data submitted by participants, following which we release an updated dataset. After this, based on a recommendation, we set up a GitHub repository to simplify collaborative annotation of the dataset. Four of the authors of this report then annotated a quarter of the dataset each, and then checked the annotations that the other three had performed to verify correctness. For those entries for which consensus was not reached, discussion between all four annotators was used to come to a final conclusion. This process resulted in better quality and higher consensus in the annotations. A very small number of errors was reported subsequent to this; a final submission version with these corrections was used by participants for their last set of experiments and to submit their final results.

Figure 1 presents the entity type distributions over the training set, test set and over the entire corpus.

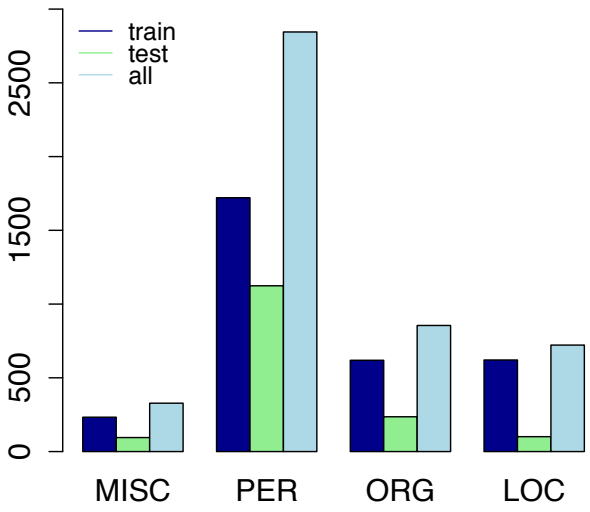


Fig. 1. Distributions of entity types in the dataset

[7] `http://oak.dcs.shef.ac.uk/msm2013/ie_challenge`

2.3 Challenge Submissions

Twenty-two complete submissions were received for the challenge; each of which consisted of a short paper explaining the system's approach, and up to three different test set annotations generated by running the system with different settings. After peer review, thirteen submissions were accepted; for each, the submission run with the best overall performance was taken as the result of the system, and used in the rankings. The accepted submissions are listed in Table 1, with the run taken as the result set for each.

Table 1. Submissions accepted, in order of submission, with authors and number of runs for each

Submission No.	Authors	No. of runs
submission_03	van Den Bosch, M. et al.	3
submission_14	Habib, M. et al.	1
submission_15	Van Erp, M. et al.	3
submission_20	Cortis, K.	1
submission_21	Dlugolinský, Š. et al.	3
submission_25	Godin, F. et al.	1
submission_28	Genc, Y. et al.	1
submission_29	Muñoz-García, O. et al.	1
submission_32	Hossein, A.	1
submission_30	Mendes, P. et al.	3
submission_33	Das, A. et al.	3
submission_34	de Oliveira, D. et al.	1
submission_35	Sachidanandan, S. et al.	1

2.4 System Descriptions

Participants approached the concept extraction task with *rule-based*, *machine learning* and *hybrid* methods. A summary of each approach can be found in Figure 2, with detail in the author descriptions that follow this report. We compared these approaches according to various dimensions: *state of the art (SoA) named entity recognition (NER) features* employed (columns 4-11) ([13,6]), *classifiers* used for both extraction and classification of entities (columns 12-13), additional *linguistic knowledge sources* used (column 14), special *pre-processing steps performed* (column 15), other *non-SoA NER features* used (column 16), and finally, the list of *off-the-shelf systems* incorporated (column 17).

From the results and participants' experiments we make a number of observations. With regard to the *strategy* employed, the best performing systems (from the top, 14, 21, 15, 25), based on overall F_1 score (see Section 3), were hybrid.

Strategy	System	Train	State-of-the-art features								Classifier used		Linguistic Knowledge	Prep.	Other Features	External Systems
			Token	Case	Morp.	POS	Function	Local syntax	List lookup	Context	Extraction	Classification				
Rule-based	20	TW		IsCap		ANNIE Pos			DBpedia Gazetteer, ANNIE Gazetteer		ANNIE		DBpedia	RT, @, #, Slang, MissSpell		ANNIE [1]
	29	TW	Ngram			PosFreeling 2012, isNP	Token Length		Wiki Gazetteer, IsStopWord		Rules	Rules	DBpedia	RT, @, URL, #, Punct, MissSpell, LowerCase		Freeling [8]
	28	TW	Ngram			NLTKPos			Wiki Gazetteer		Rules	Rules	Wiki	Punct, Capitalise		NLTK [4]
	32	TW									BabelNet WSD	Rules	DBpedia, BabelNet			BabelNet API [7]
Data-driven	3	TW, CoNLL03, ACE04, ACE05				PosTreebank			Geonames.org Gazetteer; JRC names corpus		2 IGTree memory-based taggers			LowerCase		
	33	TW	Stem	IsCap		TwPos2011		Follows FW	Country names Gazetter, City names Gazzetter, IsOOV		CRF		Samsad & NICTA dictionary	URL, #, @, Punct		
	34	TW		IsCap	Prefix, Suffix				Wiki Gazetter, Freebase Gazetter	size of TW	CRF					
Hybrid	14	TW		IsCap, AllCap		TwPos2011			Yago, Microsoft N-grams, WordNet, TW		CRF+ SVM RBF	AIDA	Yago, Microsoft N-grams, WordNet	#, Slang, MissSpell	AIDA Scores	AIDA [15]
	21	TW		IsCap, AllCap, Lower Case		isNP, isVP	Token Length		Google Gazetter		C4.5 decision tree				ConfScores	ANNIE [1], OpenNLP, Illinois NET [9], Illinois Wikifier [10], LingPipe, OpenCalais, StanfordNER [2], WikiMiner
	15	TW		IsCap, AllCap	Prefix, Suffix	TwPos2011		First Word, Last Word			SVM SMO					StanfordNER [2], NERD [12], TWNer [11]
	25	TW									Random Forest					Alchemy, DBpedia Spotlight, OpenCalais, Zemanta
	30	TW	Ngram	IsCap, AllCap, Lower Case					DBpedia Gazetteer, BALIE Gazetteer	2	DBpedia Spotlight	CRF	DBpedia	RT, #, @, URL		DBpedia Spotlight
	35	TW	Ngram						Yago, Wiki, WordNet		Pagerank	CRF	Yago, Wiki,WordNet			

Fig. 2. Automated approaches for #MSM2013 Concept Extraction. Columns correspond to the strategies employed by the participants (Strategy), the id of the systems (System), the data used to train the concept extractors (Train), state of the art features [6], Token, Case, Morphology (Morph.), Part-of-speech (POS), Function, Local context, List lookup, Context window size (Context)), classifiers used for both entity extraction (Extraction) and classification, additional linguistic knowledge used for concept extraction (Linguistic Knowledge), preprocessing steps performed on the data (Prep.), additional features used for the extractors (Other Features), a list of off-the-self systems employed (External Systems).

The success of these models appears to rely on the application of off-the-shelf systems (e.g. AIDA [15], ANNIE [1], OpenNLP[8], Illinois NET [9], Illinois Wikifier [10], LingPipe[9], OpenCalais[10], StanfordNER [2], WikiMiner[11], NERD [12], TWNer [11], Alchemy[12], DBpedia Spotlight[5][13], Zemanta[14]) for either *entity extraction* (identifying the boundaries of an entity) or *classification* (assigning a semantic type to an entity). For the best performing system (14), the complete concept classification component was executed by the (existing) concept disambiguation tool AIDA. Other systems (21, 15, 25), on the other hand, made use of the output of multiple off-the-shelf systems, resulting in additional features (such as the confidence scores of each individual NER extractors – **ConfScores**) for the final concept extractors, balancing in this way the contribution of existing extractors.

Among the rule-based approaches, the winning strategy was also similar. Submission 20 achieved the fourth best result overall, by taking an existing rule-based system (ANNIE), and simply increasing the coverage of captured entities by building new gazetteers[15]. We also find that for entity extraction the participants used both rule-based and statistical approaches. Considering current state of the art approaches, statistical models are able to handle this task well.

Looking at *features*, the gazetteer membership and part-of-speech (POS) features played an important role; the best systems include these. For the gazetteers, a large number of different resources were used, including Yago, WordNet, DBpedia, Freebase, Microsoft N-grams and Google. Existing POS taggers were trained on newswire text (e.g. ANNIEPos [1], NLTKPos [4], POS trained on Treebank corpus (PosTreebank), Freeling [8]). Additionally, there appears to be a trend on incorporating recent POS taggers trained on Micropost data (e.g. **TwPos2011** [3]).

Considering *pre-processing* of Microposts, we find the following:

- removal of Twitter-specific markers, e.g. hashtags (#), mentions (**@**), retweets (**RT**),
- removal of external URL links within Microposts (**URL**),
- removal of punctuation marks (**Punct**), e.g. points, brackets,
- removal of well-known slang words using dictionaries[16] (**Slang**), e.g. "lol", "tmr", – unlikely to refer to named entities,

[8] `http://opennlp.apache.org`
[9] `http://alias-i.com/lingpipe`
[10] `http://www.opencalais.com`
[11] `http://wikipedia-miner.cms.waikato.ac.nz`
[12] `http://www.alchemyapi.com`
[13] `http://dbpedia.org/spotlight`
[14] `http://www.zemanta.com`
[15] Another off-the-shelf entity extractor employed was BabelNet API [7], in submission 32.
[16] `http://www.noslang.com/dictionary/full`
`http://www.chatslang.com/terms/twitter`
`http://www.chatslang.com/terms/facebook`

– removal of words representing exaggerative emotions (**MissSpell**), e.g. "nooooo", "goooooood", "hahahaha",
– transformation of each word to lowercase (**LowerCase**),
– capitalisation of the first letter of each word (**Capitalise**).

With respect to the *data* used for training the entity extractors, the majority of submissions utilised the challenge training dataset, containing annotated Microcpost data (**TW**) alone. A single submission, (3, the sixth best system overall), made use of a large silver dataset (CoNLL 2003 [14], ACE 2004 and ACE 2005[17]) with the training dataset annotations, and achieved the best performance among the statistical methods.

3 Evaluation of Challenge Submissions

3.1 Evaluation Measures

The evaluation involved assessing the correctness of a system (S), in terms of the performance of the system's entity type classifiers when extracting entities from the test set (TS). For each instance in TS, a system must provide a set of tuples of the form: (entity type, entity value). The evaluation compared these output tuples against those in the gold standard (GS). The metrics used to evaluate these tuples were the standard precision (P), recall (R) and f-measure (F_1), calculated for each entity type. The final result for each system was the average performance across the four defined entity types.

To assess the correctness of the tuples of an entity type t provided by a system S, we performed a *strict match* between the tuples submitted and those in the GS. We consider a *strict match* as one in which there is an exact match, with conversion to lowercase, between a system value and the GS value for a given entity type t. Let $(x, y) \in S_t$ denote the set of tuples extracted for entity type t by system S, $(x, y) \in GS_t$ denote the set of tuples for entity type t in the gold standard. We define the set of True Positives (TP), False Positives (FP) and False Negatives (FN) for a given system as:

$$TP_t = \{(x, y) \mid (x, y) \in (S_t \cap GS_t)\} \tag{1}$$

$$FP_t = \{(x, y) \mid (x, y) \in S_t \wedge (x, y) \notin GS_t\} \tag{2}$$

$$FN_t = \{(x, y) \mid (x, y) \in GS_t \wedge (x, y) \notin S_t\} \tag{3}$$

Therefore TP_t defines the set of true positives considering the entity type and value of tuples; FP_t is the set of false positives considering the unexpected results for an entity type t; FN_t is the set of false negatives denoting the entities that were missed by the extraction system, yet appear within the gold standard. As we require matching of the tuples (x, y) we are looking for strict extraction matches, this means that a system must both detect the correct entity type (x)

[17] the ACE Program: `http://projects.ldc.upenn.edu/ace`

and extract the correct matching entity value (y) from a Micropost. From this set of definitions we define precision (P_t) and recall (R_t) for a given entity type t as follows:

$$P_t = \frac{|TP_t|}{|TP_t \cup FP_t|} \tag{4}$$

$$R_t = \frac{|TP_t|}{|TP_t \cup FN_t|} \tag{5}$$

As we compute the precision and recall on a per-entity-type basis, we define the average precision and recall of a given system S, and the harmonic mean, F_1 between these measures:

$$\bar{P} = \frac{P_{PER} + P_{ORG} + P_{LOC} + P_{MISC}}{4} \tag{6}$$

$$\bar{R} = \frac{R_{PER} + R_{ORG} + R_{LOC} + R_{MISC}}{4} \tag{7}$$

$$F_1 = 2 \times \frac{\bar{P} \times \bar{R}}{\bar{P} + \bar{R}} \tag{8}$$

3.2 Evaluation Results and Discussion

We report the differences in performance between participants' systems, with a focus on the differences in performance by entity type. The following subsections report results of the evaluated systems in terms of precision, recall and F-measure, following the metrics defined in subsection 3.1.

Precision. We begin by discussing the performance of the submissions in terms of precision. Precision measures the accuracy, or '*purity*', of the detected entities in terms of the proportion of false positives within the returned set: high precision equates to a low false positive rate. Table 3.2 shows that hybrid systems are the top 4 ranked systems (in descending order, 14, 21, 30, 15), suggesting that a combination of rules and data-driven approaches yields increased precision. Studying the features of the top-performing systems, we note that maintaining capitalisation is correlated with high precision. There is, however, clear variance in other techniques used (classifiers, extraction methods, etc.) between the systems.

Fine-grained insight into the disparities between precision performance was obtained by inspecting the performance of the submissions across the different concept types (person, organisation, location, miscellaneous). Figure 3a presents the distribution of precision values across these four concept types and the macro average of these values. We find that systems do well (above the median of average precision values) for person and location concepts, and perform worse than the median for organisations and miscellaneous. For the entity type '*miscellaneous*', this is not surprising as it features a fairly *nuanced* definition, including

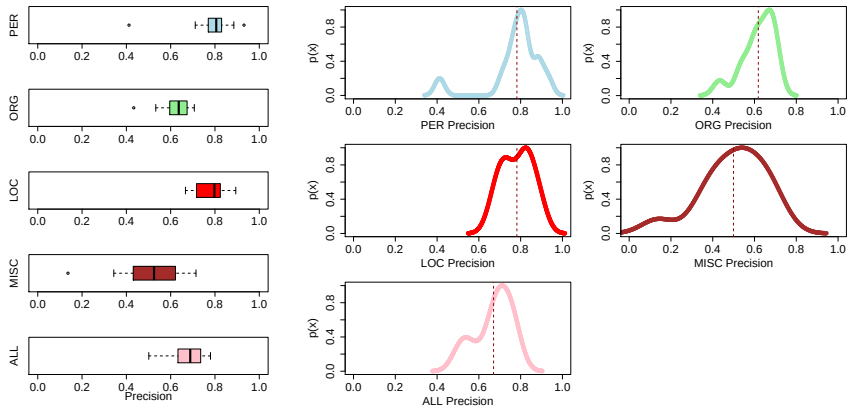
films and movies, entertainment award events, political events, programming languages, sporting events and TV shows. We also note that several submissions used gazetteers in their systems, many of which were for locations; this could have contributed to the higher precision values for location concepts.

Table 2. Precision scores for each submission over the different concept types

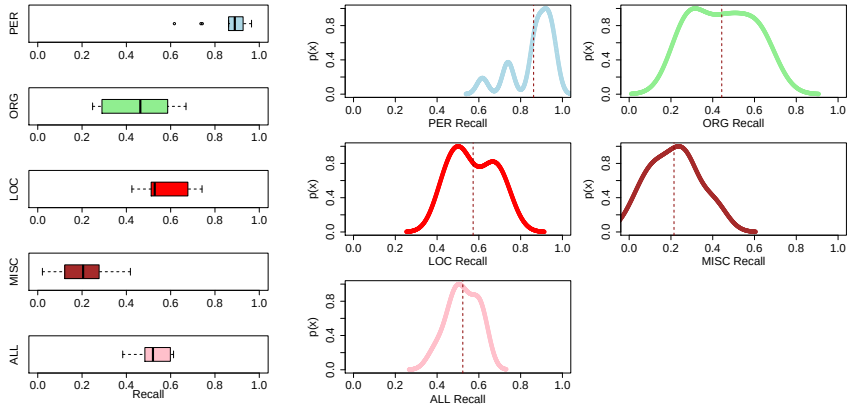
Rank	Entry	PER	ORG	LOC	MISC	ALL
1	14 - 1	0.923	0.673	0.877	0.622	0.774
2	21 - 3	0.876	0.603	0.864	0.714	0.764
3	30 - 1	0.824	0.648	0.800	0.667	0.735
4	15 - 3	0.879	0.686	0.844	0.525	0.734
5	33 - 3	0.809	0.707	0.746	0.636	0.724
6	25 - 1	0.771	0.606	0.824	0.548	0.688
7	03 - 3	0.813	0.696	0.794	0.435	0.685
8	29 - 1	0.785	0.596	0.800	0.553	0.683
9	28 - 1	0.765	0.674	0.711	0.500	0.662
10	20 - 1	0.801	0.636	0.726	0.343	0.627
11	32 - 1	0.707	0.433	0.683	0.431	0.564
12	35 - 1	0.740	0.533	0.712	0.136	0.530
13	34 - 1	0.411	0.545	0.667	0.381	0.501

Recall. Although precision affords insight into the accuracy of the entities identified across different concept types, it does not allow for inspecting the detection rate over all possible entities. To facilitate this we also report the recall scores of each submission, providing an assessment of the entity coverage of each approach. Table 3 presents the overall recall values for each system and for each and across all concept types. Once again, as with precision, we note that hybrid systems (21, 15, 14) appear at the top of the rankings, with a rule-based approach (20) and a data driven approach (3) coming fourth and fifth respectively.

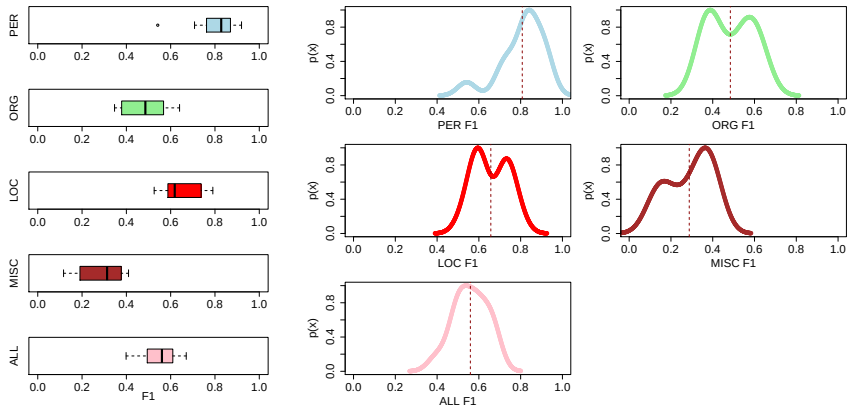
Looking at the distribution of recall scores across the submissions in Figure 3c we see a similar picture as before when inspecting the precision plots. For instance, for the person and location concepts we note that the submissions exceed the median of all concepts (when the macro-average of the recall scores is taken), while for organisation and miscellaneous lower values than the median are observed. This again comes back to the nuanced definition of the miscellaneous category, although the recall scores are higher on average than the precision score. The availability of person name and place name gazetteers also benefits identification of the corresponding concept types. This suggests that additional effort is needed to improve the *organisation* concept extraction and to provide information to seed the detection process, for instance through



(a) Concept type Precision (b) Probability densities of concept type Precision



(c) Concept type Recall (d) Probability densities of concept type Recall



(e) Concept type F_1 (f) Probability densities of concept type F_1.

Fig. 3. Distributions of performance scores for all submissions; dashed line is the mean.

the provision of organisation name gazetteers. Interestingly, when we look at the best performing system in terms of recall over the organisation concept we find that submission 14 uses a variety of third party lookup lists (Yago, Microsoft n-grams and Wordnet), suggesting that this approach leads to increased coverage and accuracy when extracting organisation names.

Table 3. Recall scores for each submission over the different concept types

Rank	Entry	PER	ORG	LOC	MISC	ALL
1	21 - 3	0.938	0.614	0.613	0.287	0.613
2	15 - 3	0.952	0.485	0.739	0.269	0.611
3	14 - 1	0.908	0.611	0.620	0.277	0.604
4	20 - 1	0.859	0.587	0.517	0.418	0.595
5	03 - 3	0.926	0.463	0.682	0.122	0.548
6	25 - 1	0.887	0.405	0.685	0.205	0.546
7	28 - 1	0.864	0.290	0.692	0.155	0.500
8	29 - 1	0.736	0.489	0.444	0.263	0.483
9	32 - 1	0.741	0.289	0.506	0.391	0.482
10	35 - 1	0.920	0.346	0.506	0.102	0.468
11	33 - 3	0.877	0.248	0.518	0.077	0.430
12	34 - 1	0.787	0.283	0.439	0.098	0.402
13	30 - 1	0.615	0.268	0.444	0.204	0.383

F-Measure (F_1). By combining the precision and recall scores together for the individual systems using the f-measure (F_1) score we are provided with an overall assessment of concept extraction performance. Table 4 presents the f-measure (F_1) score for each submission and performance across the four concept types. We note that, as previously, hybrid systems do best overall (top-3 places), indicating that a combination of rules and data-driven approaches yields the best results. Submission 14 records the highest overall F_1 score, and also the highest scores for the person and organisation concept types; submission 15 records the highest F_1 score for the location concept type; while submission 21 yields the highest F_1 score for the miscellaneous concept type. Submission 15 uses Google Gazetteers together with part-of-speech tagging of noun and verb phrases, suggesting that this combination yields promising results for our nuanced miscellaneous concept type.

Figure 3e shows the distribution of F_1 scores across the concept types for each submission. We find, as before, that the systems do well for person and location and poorly for organisation and miscellaneous. The reasons behind the reduced performance for these latter two concept types are, as mentioned, attributable to the availability of organisation information in third party lookup lists.

Table 4. F_1 scores achieved by each submission for each and across all concept types

Rank	Entry	PER	ORG	LOC	MISC	ALL
1	14 - 1	0.920	0.640	0.738	0.383	0.670
2	21 - 3	0.910	0.609	0.721	0.410	0.662
3	15 - 3	0.918	0.568	0.790	0.356	0.658
4	20 - 1	0.833	0.611	0.618	0.377	0.610
5	25 - 1	0.828	0.486	0.744	0.298	0.589
6	03 - 3	0.870	0.556	0.738	0.191	0.589
7	29 - 1	0.762	0.537	0.587	0.356	0.561
8	28 - 1	0.815	0.405	0.705	0.236	0.540
9	32 - 1	0.727	0.347	0.587	0.410	0.518
10	30 - 1	0.708	0.379	0.578	0.313	0.494
11	33 - 3	0.846	0.367	0.616	0.137	0.491
12	35 - 1	0.823	0.419	0.597	0.117	0.489
13	34 - 1	0.542	0.372	0.525	0.155	0.399

4 Conclusions

The aim of the MSM Concept Extraction Challenge was to foster an open initiative for extracting concepts from Microposts. Our motivation for hosting the challenge was born of the increased availability of third party extraction tools, and their widespread uptake, but the lack of an agreed formal evaluation of their accuracy when applied over Microposts, together with limited understanding of how performance differs between concept types. The challenge's task involved the identification of entity types and value tuples from a collection of Microposts. To our knowledge the entity annotation set of Microposts generated as a result of the challenge, and thanks to the collaboration of all the participants, is the largest annotation set of its type openly available online. We hope that this will provide the basis for future efforts in this field and lead to a standardised evaluation effort for concept extraction from Microposts.

The results from the challenge indicate that systems performed well which: (i) used a hybrid approach, consisting of data-driven and rule-based techniques; and (ii) exploited available lookup lists, such as place name and person name gazetteers, and linked data resources. Our future efforts in the area of concept extraction from Microposts will feature additional hosted challenges, with more complex tasks, aiming to identify the differences in performance between disparate systems and their approaches, and inform users of extraction tools on the suitability of different applications for different tasks and contexts.

5 Acknowledgments

We thank the participants who helped us improve the gold standard used for the challenge. We also thank eBay for supporting the challenge by sponsoring the prize for winning submission.

A.E. Cano is funded by the EPSRC project *violenceDet* (grant no. EP/J020427/1). A.-S. Dadzie was funded by the MRC project *Time to Change* (grant no. 129941).

References

1. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Annual Meeting of the ACL*, 2002.
2. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 2005.
3. K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
4. E. Loper and S. Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 62–69. Somerset, NJ: Association for Computational Linguistics, 2002.
5. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, 2011.
6. D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), 2007.
7. R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193, 2012.
8. L. Padró and E. Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2473–2479, Istanbul, Turkey, May 2012. ACL Anthology Identifier: L12-1224.
9. L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, 2009.
10. L.-A. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *ACL*, 2011.
11. A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *EMNLP*, 2011.
12. G. Rizzo and R. Troncy. NERD: evaluating named entity recognition tools in the web of data. In *ISWC 2011, Workshop on Web Scale Knowledge Extraction (WEKEX'11), October 23-27, 2011, Bonn, Germany*, Bonn, GERMANY, 10 2011.

13. S. Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1:261–377, 2008.
14. E. F. Tjong Kim Sang and F. 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*, CONLL '03, pages 142–147. Association for Computational Linguistics, 2003.
15. M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum. Aida: An online tool for accurate disambiguation of named entities in text and tables. *PVLDB*, 2011.