Diamonds in the Rough: Event Extraction from Imperfect Microblog Data

Ander Intxaurrondo†, Eneko Agirre†, Oier Lopez de Lacalle†, Mihai Surdeanu‡
†IXA NLP Group, University of the Basque Country
‡University of Arizona
{ander.intxaurrondo, e.agirre, oier.lopezdelacalle}@ehu.eus
msurdeanu@email.arizona.edu

Abstract

We introduce a distantly supervised event extraction approach that extracts complex event templates from microblogs. We show that this near real-time data source is more challenging than news because it contains information that is both approximate (e.g., with values that are close but different from the gold truth) and ambiguous (due to the brevity of the texts), impacting both the evaluation and extraction methods. For the former, we propose a novel, "soft", F1 metric that incorporates similarity between extracted fillers and the gold truth, giving partial credit to different but similar values. With respect to extraction methodology, we propose two extensions to the distant supervision paradigm: to address approximate information, we allow positive training examples to be generated from information that is similar but not identical to gold values; to address ambiguity, we aggregate contexts across tweets discussing the same event. We evaluate our contributions on the complex domain of earthquakes, with events with up to 20 arguments. Our results indicate that, despite their simplicity, our contributions yield a statistically-significant improvement of 33% (relative) over a strong distantly-supervised system. The dataset containing the knowledge base, relevant tweets and manual annotations is publicly available.

1 Introduction

Twitter is an excellent source of near real-time data on recent events, motivating the need for information extraction (IE) systems that operate on tweets rather than traditional news articles. However, using this data comes with its own challenges: tweets tend to use colloquial speech, noisy syntax and discourse, and, more importantly, the information reported is often inaccurate (e.g., reporting a different but similar magnitude for an earthquake) and ambiguous (e.g., reporting multiple potential earthquake locations, with insufficient context to guess which is the correct one).[1] The top rows in Table 1 show examples of these problems for an actual event in our dataset on earthquakes. This comes in contrast with "traditional" IE work on newswire documents, where information is considerably more accurate than microblog material, and none of the above observations hold (Grishman and Sundheim, 1996; Doddington et al., 2004).

As an example of the benefits of event extraction from a near real-time social-media resource, the last row in Table 1 lists a motivating example, where our system extracts the correct depth of an earthquake from the text tweeted by the U.S. Geological Survey, which is novel information that is missing in our manually-curated knowledge base.

In this work we take a classic event extraction (EE) task, where events are defined by templates containing a predefined set of arguments, and implement it using data from Twitter. We avoid the prohibitive cost of manual annotation through distant supervision (DS): we automatically generate train-

[1] We focus on microblogs here because they commonly contain inaccurate and/or ambiguous information. However, we believe that our contributions extend beyond microblogs because these innacuracies, especially inaccurate information, may appear in news article as well.

Approximate information	Earthquake in Honduras. So strong it strong it was felt in Guatemala as well. **7.1** offshore atlantic.
Ambiguous information	DTN **Indonesia**: *Peru* Earthquake Destroys Homes, Injures 100...
	6.9 magnitude earthquake rocks *Peru*.
	U.S.G.S. reports 6.9 Earthquake in *Peru*. NO TSUNAMI threat to Hawaii.
Information not in the knowledge base	#Earthquake M 7.0 – Ryukyu Islands, Japan T20:31:27 UTC , 25.95 128.40 depth: **22 km** <USGS URL> Local tsunami alert issued

Table 1: Challenges and opportunities for event extraction from Twitter. The first row shows a tweet with approximate information (in bold); the correct magnitude is 7.3 (cf. Table 2). The second row shows a first tweet with ambiguous information, which leads our baseline model to extract the incorrect country (in bold; correct country is *Peru*). The following two tweets help disambiguate the context. The last row shows a tweet containing information (in bold) that is missing in the knowledge base.

ing data by aligning a knowledge base of known event instances with tweets (Mintz et al., 2009; Hoffmann et al., 2011), which is then used to train a supervised extraction model (sequence tagger in our case). In seminal work on event extraction, (Benson et al., 2011) applied DS to both detect tweets about local events and then extracted values about two arguments (artist and venue). In our work, we work on automatically selected tweets, and scale the task to complex events with a large number of arguments. We focus on the domain of earthquakes, where each event has up to 20 arguments. Table 2 summarizes this task.

The contributions of this work are the following:

1. To our knowledge, this is one of the first works that analyzes the problem of distantly supervised extraction of complex events with many arguments from microblogs.

2. Our analysis shows (Section 3) that the biggest barrier is that information on Twitter can be *inaccurate* (containing approximately correct event argument values) and *ambiguous* (with insufficient context for accurate extraction). The top two blocks in Table 1 show an example of each. These challenges impact both evaluation and system development.

3. The analysis also highlights the need to adapt evaluation metrics to approximately correct information, which may appear both in text and in the knowledge base itself. For example, for a particular earthquake, the USGS reports a depth of 22 km., while NOAA reports 25 km[2]. We propose a new evaluation metric that gives partial credit to extracted argument values based on their similarity to existing values in the knowledge base.

4. We introduce two simple strategies that address the above barriers for system development: *approximate matching*, which addresses inaccurate values by allowing the distant supervision process to map values from the knowledge base to text even when they do not match exactly; and *feature aggregation*, which responds to small, ambiguous contexts by aggregating information across multiple tweets for the same event. For example, the first strategy considers the 7.1 magnitude in the first tweet in Table 1 as a training example because it is close to the value in the knowledge base (7.3). The second strategy classifies all instances of *Peru* jointly using a single set of features, extracted from all available tweets for the corresponding earthquake. For example, this feature set contains three values for the feature `previous-word` (*:*, *rocks*, and *in*). Each approach yields 19% relative improvement, 33% in combination.

5. We release a public dataset containing a knowledge base of earthquake instances and corresponding tweets for each earthquake[3].

2 Experimental framework

In this section we detail the creation of the knowledge base of earthquake events, the collection process for potentially-relevant tweets, and, lastly, our distant supervision framework, which serves as a platform for our contributions (Sections 5 and 6).

2.1 Knowledge base and tweet dataset creation

The **knowledge base (KB)** was created from the list of globally significant earthquakes during the 21st century, as reported by Wikipedia.[4] We se-

[2] `http://bit.ly/aq9Vxa` and `http://1.usa.gov/1p1gELB`

[3] `http://ixa.eus/Ixa/Argitalpenak/Artikuluak/1425465524/publikoak/earthquake-kb-dataset.zip`

[4] `https://en.wikipedia.org/wiki/List_of_21st-century_earthquakes`. Accessed on July 9th,

Argument Name	Arg. Type	# KB Values	Example Values	# DS Values	# MA Values
Date	D	108	2009-5-28	291	706
Time	T	108	T08:24:00	378	589
Country	L	108	Honduras	6294	6327
Region	L	77		2598	2663
City	L	77		1426	1723
Latitude	N	108	16.733	2	28
Longitude	N	108	-86.22	4	28
Dead	N	71	7	143	984
Injured	N	39		22	192
Missing	N	8		-	18
Magnitude	N	108	7.3	933	3403
Depth (km)	N	99	10	27	313
Countries affected(*)	L	37	Guatemala, Belize	436	357
Regions affected(*)	L	4		-	36
Landslides	B	8		7	9
Tsunami	B	10		408	273
Aftershocks	N	20		5	22
Foreshocks	N	3		6	-
Duration	T	7		-	1
Peak accel.	N	8		-	-
TOTAL		1,116		13,562	17,672

Table 2: Event arguments and types in the earthquake domain (first and second column), summary statistics for the knowledge base, i.e., the gold truth (third column), and values for one example earthquake (4th column). (*) indicates multi-valued arguments (all other are single-valued). The two rightmost columns give statistics for the number of mentions in the tweets per argument, as obtained through manual annotation (MA) or distant supervision (DS) (cf. Section 2.4). The argument types are the following: *D* date, *T* time, *L* location, *N* numeric, and *B* boolean.

lected earthquakes from the beginning of 2009, with the last reported earthquake happening on July 7th, 2013, and constructed the KB from the above Wikipedia list page and the individual infoboxes. Where necessary, argument values were normalized.[5] See Table 2 for a summary and an example.

We used the Topsy API[6] to search for **tweets** that are potentially relevant for each earthquake. We formed a query using the word "earthquake" plus the location, encoded as a disjunction of city, region, and country arguments. We retrieved tweets from the day before the date and time of the earthquake, up to seven days after. This procedure might also retrieve tweets about aftershocks, which we consider to be different events. We applied an aggressive method to discard aftershock tweets: we only kept tweets up to the first tweet that mentions a time expression more than a minute different from that of the main earthquake (after adjusting for time zone). For example, this heuristic removes all tweets starting with *"A 4.9 earthquake occurred in Ryukyu Islands, Japan on 2010-2-27 T10:33:21 at epicenter."* because the main earthquake occurred on February 26th at 8:31PM UTC. It is important to note that identifying event-relevant tweets is not the focus of this work (hence the simple heuristics used for tweet extraction). We focus instead on the *extraction* of information from such tweets. In a complete system, our approach would follow a component that detects event tweets automatically (Benson et al., 2011). The final dataset contains 108 earthquakes and 7,841 tweets, 72 tweets per earthquake on average, a maximum of 654 and a minimum of 2. 19 earthquakes had less than 10 tweets.

2.2 Manual annotation of tweets

In order to analyze the challenges faced by our EE system based on distant supervision, we also manually annotated all tweets.[7] The manual annotation included any mention of an event argument in the tweets. This included information already in the KB, but also information that is missing, caused by: variations of dates and times, similar but not identical latitude/longitude values, different reported numbers for dead/injured/missing etc. The first tweet in Table 1 is an example of this situation: even though the reported magnitude is different from the value in the KB (cf. example in Table 2), it was annotated during this process. In total, we annotated 17,672 mentions (at an average of two event arguments per tweet). Table 2 shows the breakdown per argument (the MA column), compared to the automatic annotations generated through distant supervision (the DS column). Note that some of the arguments have a very different coverage in the tweets compared with the KB. For example, latitude and longitude are rarely present in tweets, but affected countries are commonly mentioned. The quality of the manual annotation was assessed on a 5% sample of the dataset, which was annotated by an additional expert. The agreement was very high: 90% ITA and 85% Fleiss Kappa. Disagreements were generally

2013, at 2PM CET.

[5]Time and date expressions were converted to TimeML. Numerical values in English were converted to numbers, latitude and longitudes were converted to decimal format.

[6]`http://api.topsy.com/doc/`

[7]These manual annotations are used solely for post-hoc analysis, *not* to train our system.

due to missed argument mentions. Note that the cost of annotation was around 75 hours, confirming the cost-saving properties of distant supervision.

2.3 Dataset and experiment organization

We sorted the list of earthquakes in the KB chronologically, and chose the earliest 75% of the earthquakes as the training dataset, and the most recent (25%) for testing. The training set contained 81 earthquakes and their corresponding 6078 tweets, while the testing set contained 27 earthquakes and 1763 tweets. All development experiments were performed using 5-fold cross-validation over the training partition, where the folds were organized randomly by earthquake. Each fold contained tweets for around 15 earthquakes, but the number of tweets varied widely, with one fold having 585 tweets and another 2229.

The evaluation compares the argument values induced by our system with those in the gold KB, and computes precision, recall and F1 using the official scorer from the Knowledge Base Population (KBP) Slot Filling (SF) shared task (Surdeanu, 2013). We also incorporated the notion of equivalence classes proposed in the SF task. For instance, if the system predicted *Guerrero State* for the argument `region`, when the KB contains just *Guerrero*, we consider this result correct because the two strings are equivalent in this context. Our equivalence classes also include countries, regions, and cities with hashtags, unnormalized temporal expressions, etc. Where applicable, we checked statistical significance of performance differences using the bootstrap resampling technique proposed in (Berg-Kirkpatrick et al., 2012), in which we draw many simulated test sets by sampling with replacement from the set of earthquakes in the test partition.

2.4 Distant supervision for event extraction

For the initial extraction experiment, we followed a traditional distant supervision approach (Mintz et al., 2009), which has four steps: the KB of past events is aligned to the text; a supervised system is trained on the resulting annotated text; the system is run on test data; and the output slot values are inferred from the annotations produced by the system. We thus started by aligning the information in the KB to the training tweets using strict matching[8]. Table 2 compares the number of mentions automatically generated through DS against the number of manually annotated mentions. As expected, the strict matching criterion yields fewer mentions than the manual annotation.

As an example of this process, given the Honduras earthquake in Table 2, this procedure will annotate two argument mentions in the first tweet from Table 1, `country` and `affected-country`, as follows:

> Earthquake in <country>Honduras</country>. So strong it was felt in <affected-country>Guatemala</affected-country> as well. 7.1 offshore atlantic.

Note that the magnitude in the tweet is different from the one reported in the KB and it will thus be left unmarked (we revisit this issue in Section 5).

Using this automatically-generated data, we trained a sequential tagger based on Conditional Random Fields (CRF)[9]. Based on the output of the CRF, we inferred the arguments values using noisy-or (Surdeanu et al., 2012), which selects the value with the largest probability for each single-valued argument by aggregating the individual mention probabilities produced by the CRF.[10] In the case of multi-valued arguments (`affected-country` and `affected-region`) we choose all values that had been annotated by the sequential tagger.

3 Initial results and analysis

The left block in Table 3 reports the results on development (5-fold cross-validation) of the initial event

[8] We identified two types of arguments: those that have binary (yes/no) values (`tsunami` and `landslides`) and those having other values. For the first type, we search the tweets corresponding to the target earthquake for a small number of strings (e.g., *tsunami* and *tsunamis*), and annotate all matches (e.g., <*tsunami*> *tsunami* </*tsunami*>). For non-binary valued arguments, we searched the tweets for exact occurrences of the corresponding values, and annotated all matching strings. When the same value appears in more than one argument for the same earthquake (e.g., *7* as both magnitude and number of dead people), we choose the most common label (e.g., magnitude cf. Table 2).

[9] We used the linear CRF in Stanford's CoreNLP package, with the default features (word form, PoS, lemma, NERC) for the `macro` configuration: `http://nlp.stanford.edu/software/corenlp.shtml`.

[10] For multi-token mentions (e.g. *New Zealand*) we use the average of the token probabilities.

	Strict Evaluation		
System	**Prec.**	**Rec.**	**F1**
DS-CRF	53.1	22.0	31.1
MA-CRF	44.1	26.1	32.8
	Lenient Evaluation		
DS-CRF	67.4	27.9	39.4
MA-CRF	62.1	36.8	46.2

Table 3: Development: Results for the distant supervision system (DS-CRF). We also include results for the same CRF trained on manual annotations (MA-CRF). The regular evaluation is shown in the left columns and lenient evaluation (cf. Section 4) in the right.

extraction system based on a distantly-supervised CRF (DS-CRF), which notably attains higher precision than recall. These results are fair, e.g., they are comparable to those of (Benson et al., 2011), even though their events had much fewer argument types than ours (two vs. twenty). More importantly, we use this system's output to analyze where the approach could be improved. For the sake of comparison, we trained the same CRF with the manually annotated tweets, cf. Section 2 (MA-CRF). The MA-CRF results in Table 3 indicate that the main loss when doing distant supervision is in recall, but the overall F1 is close. This is remarkable, as the much more expensive MA-CRF (75 hours of human annotation) is taken to be an upperbound for DS-CRF.

Manual inspection showed that that DS-CRF returns fewer argument values than MA-CRF (328 vs. 469), from "easier" (more common) arguments which have a higher chance of appearing both in the text and the KB. Importantly, MA-CRF has lower precision than its distant supervision counterpart because it is trained on manual annotations, which included many mentions not in the KB. The consequence of this strategy is that MA-CRF tends to produce spurious mentions (i.e., mentions not in the KB) at evaluation time, which lowers precision.

In addition, we analyzed the annotations created through distant supervision[11], which produced 13,562 argument mentions in the training tweets (cf. Table 2, which also includes a breakdown by argument). This data contains incorrectly annotated strings (false positives) and also misses relevant argument values (false negatives). A comparison of these DS annotations against the manual annotations on all training tweets (17,672 mentions) yielded that 97.4% were correct, but that 27.4% of the gold manual annotations were missed. This is an important result: it demonstrates that, unlike in the problem of relation extraction (RE) where the major issue is the large percentage (higher than 30%) of false positives in automatically-created annotations (Riedel et al., 2010), here the fundamental roadblock is missing annotations (i.e., false negatives). We explain this difference by the fact that for this event extraction domain, it is trivial to identify domain-relevant tweets, which reduces the number of false positives for event arguments. We believe this generalizes to many other EE domains, e.g., airplane crashes (Reschke et al., 2014) or terrorist attacks, where the event context can be summarized accurately with a small number of keywords (e.g., flight number and date for the airplane crashes domain).

We also did a post-hoc analysis of the quality of the arguments induced by DS-CRF. One of the most significant outcomes of the analysis is that a large portion of numeric values (31.3%) were partially correct, in that the returned values were very similar to those in the KB (see for instance the 7.1 vs. 7.3 example in Section 1). This strongly suggests that the evaluation metric should be more lenient, and give credit to argument values that are similar to the gold ones.

4 Lenient evaluation

The previous analysis suggests that traditional evaluation measures unnecessarily penalize arguments containing values that do not match the gold truth exactly. Rather than giving no credit when predicted values are different from gold ones, we devised a simple extension to the KBP evaluation measures that take into account the similarity between the values of system and gold arguments, where the similarity depends on the type of each slot (cf. Table 2). For numeric values, we use the following formula, where x is the predicted value, and g the gold value:

$$sim(x, g) = max\left(1 - \frac{|x - g|}{g}, 0\right) \qquad (1)$$

[11] Note that these are the argument mention annotations used to train DS-CRF, not the arguments inferred by the DS-CRF system.

For example, given a gold value of 7.3, a system value of 7.2 would have a similarity of 0.98, and a system value of 14.6 or larger would have a similarity 0. If both values are equal, similarity is 1.

For the other slot types, the similarity function is discrete, with values set to 1 (proposed slot is correct) or 0 (incorrect) as follows. We consider a proposed *temporal* argument as correct if it is within a span of 5 minutes of the corresponding gold temporal value. *Durations* are judged as correct if they are within 10 seconds of the gold values. We considered proposed *dates* as correct if they differ by at most one day from the gold date.[12]

For *location* arguments, we use GeoNames[13] to obtain the coordinates of the locations produced by the system that do not match the information in the KB. Based on the average size of countries, regions, and cities, we consider these additional locations as correct if they are at the following distance (or closer) from the gold locations: 500 kms for countries, 50 kms for regions, and 10 kms for cities.

The original KBP scorer increases the value of True Positives (TP) by 1 every time a predicted argument matches its gold value. In the proposed lenient scorer, TP is increased by the similarity between the predicted and gold values. The precision and recall will be thus calculated as follows (*SYS* for number of predicted argument values, *GOLD* for number of gold argument values):

$$prec = \frac{\sum sim(x, g)}{SYS} \quad (2)$$

$$rec = \frac{\sum sim(x, g)}{GOLD} \quad (3)$$

The right block in Table 3 lists the results under this lenient evaluation for the experiment initially reported in the left block in the same table. As expected, these results are higher than the ones using the strict measure, but maintain the relative order of the systems in each of the evaluation measures. The difference in precision between DS-CRF and MA-CRF decreases, indicating that the new measure assigns partial credit to the larger amount of argument values extracted by MA-CRF. The difference in recall values remains large. We address this in the next section.

[12]These thresholds might change in other domains, but adjusting these values is trivial.

[13]http://www.geonames.org/

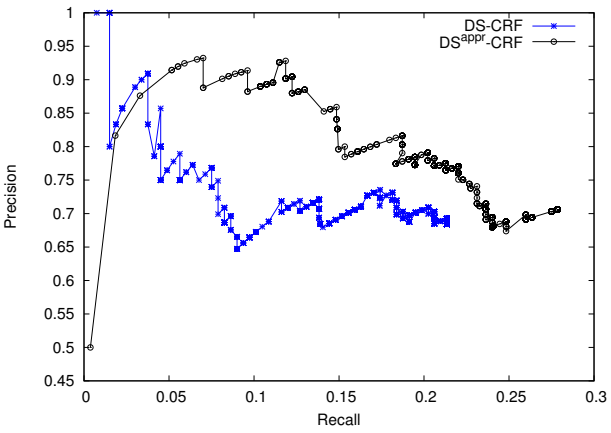

Figure 1: Test: Precision/Recall curves for regular DS and approximate DS on test (lenient evaluation).

System	Prec.	Rec.	F1
DS-CRF	68.4	21.3	32.5
DS^{appr}-CRF	70.6	27.8	39.9 †

Table 4: Test: Regular (DS-CRF) and approximate DS (DS^{appr}-CRF) results, with lenient evaluation. † indicates statistically significant improvement over DS-CRF ($p < 0.05$).

5 Approximate distant supervision

The previous section demonstrated that many tweets contain argument values which are similar but not identical to the data in the knowledge base. These values would not be annotated during alignment by traditional distant supervision, which expects an exact match between knowledge base values and tweet texts. This means that DS-CRF will be trained with less data than what is available (e.g., without the 7.1 magnitude example in the tweet in Section 2.4). Here we demonstrate that a simple extension to distant supervision that annotates values close to the values in the knowledge base, results in improved performance.

The proposed alignment algorithm scans the training tweets, and labels named and numeric entities as positive argument examples (with the corresponding label from the KB), if they are deemed similar to the gold values according to the similarity formulas introduced in the previous section. This

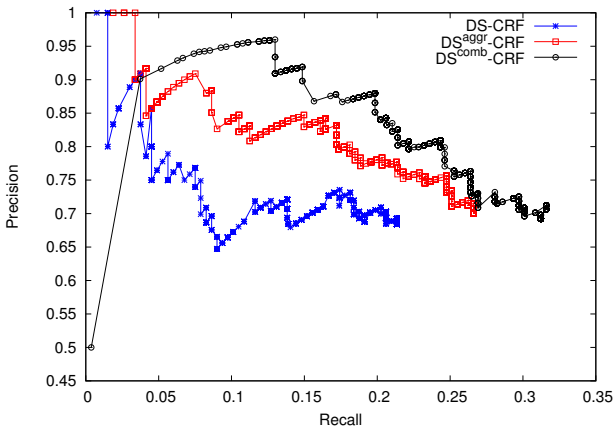

Figure 2: Test: P/R curves for DS-CRF, feature aggregation and combination with approximate DS (lenient evaluation).

System	Prec.	Rec.	F1
DS-CRF	68.4	21.3	32.5
DS^{aggr}-CRF	70.1	26.6	38.6 †
DS^{comb}-CRF	69.2	31.2	43.1 †
MA-CRF	69.1	37.9	48.9

Table 5: Test: Results for regular DS (DS-CRF), DS with feature aggregation (DS^{aggr}-CRF), and the DS model that combines feature aggregation and approximate matching (DS^{comb}-CRF), with lenient evaluation. † indicates statistically significant improvement over DS-CRF ($p < 0.05$). We include the results of the CRF trained on manual annotations (MA-CRF) as a performance ceiling for this task.

is a trivial process for discrete similarities, but requires some care for continuous similarity functions, which are triggered for numeric arguments. In this situation, numeric entities are considered as positive examples only if their similarity function returns a value over a certain threshold with a known argument in the KB. If a numeric mention has more than one matching argument in the KB, the algorithm chooses the argument label with the highest similarity value; if all have the same similarity, the algorithm chooses the most frequent label in training.

We tuned the threshold hyper parameter for numeric values over the training dataset using 5-fold cross validation, which yielded 0.95 as the optimal value. Table 4 shows the results for the test partition using this threshold, and Figure 1 shows the corresponding P/R curves. Both results are generated using the proposed lenient evaluation. The results in the table show that, despite its simplicity, the proposed alignment algorithm yields considerable, statistically-significant improvements. The P/R curves show that the improvement holds for all recall points[14].

6 Feature aggregation

The second block in Table 1 illustrates a common scenario on Twitter, where a short, ambiguous tweet derails the extraction. We address this problem of insufficient local context with a method inspired by work in relation extraction, where relation instances between identical entities are classified jointly using the conjunction of features from all instances (Mintz et al., 2009). We adapt this idea to our sequence tagging EE model as follows:

1: We focus on location, date and temporal entities (both earthquake time and duration) which are argument candidates that are often ambiguous, i.e., they may be classified as more than one argument type. For example, a location entity may be labeled as `country`, `region`, `country-affected`, etc. We exclude numeric entities due to potential feature collisions between different argument types: we observed that, in training, several earthquakes had different numeric arguments with the same value. For example, the magnitude and depth of the 2012 Zohan earthquake were 5.6. Applying feature aggregation to examples of these arguments would lead to collisions between features from different classes.[15]

2: For each token that appears in one of these named entities, we identify all its instances across the relevant tweets, and share features across all these token instances. For example, for the tweets in the second block in Table 1, our approach identifies *Peru* as an argument mention candidate. All three instances of *Peru* are then classified using the same shared features, e.g., using three values for the fea-

[14]The curves for the strict evaluation are similar, and were omitted for brevity.

[15]Initial experiments confirmed this hypothesis: feature aggregation did not improve results for numeric arguments in development. In future work, we will explore multi-instance multi-label algorithms to handle this situation (Surdeanu et al., 2012).

ture `previous-word` (*:*, *rocks*, and *in*). This process is repeated for each earthquake individually, because tokens may be labeled differently in different earthquakes. This approach produced 37% more features than the DS-CRF baseline.[16]

The positive effect of feature aggregation is confirmed by the formal evaluation on the test dataset. Table 5 shows a statistically significant improvement in overall F1, for the lenient evaluation. The P/R curves (Fig. 2) indicate that DS$^{\text{aggr}}$-CRF's improvement comes from both better recall and better precision that the DS-CRF baseline.

Table 5 and Fig. 2 also show that the combination of approximate matching and aggregation outperforms the individual models, demonstrating that feature aggregation is complementary to approximate matching. The combined model attains a relative improvement of 33% over the DS-CRF baseline, reaching approximately 88% of the ceiling performance for this task (MA-CRF row, the CRF trained on manual annotations).

7 Related work

There has been considerable recent interest in IE from Twitter. However, in general, these works use supervised learning frameworks (Popescu et al., 2011; Ritter et al., 2012), and/or they use either a coarse representation of events, which reduces to topic modeling or classification of entire tweets (Popescu et al., 2011; Becker et al., 2011; Ritter et al., 2012), or a simplified representation of events with few arguments (Sakaki et al., 2010; Popescu et al., 2011; Benson et al., 2011; Ritter et al., 2012). In contrast, our work uses a complex event representation with 20 arguments, and does not require any manual annotation of tweets. Our work is closest, but complementary to the work of (Benson et al., 2011), which also uses distant supervision for event extraction: We provide solutions for two problems they do not address (inaccurate and ambiguous information) and we focus on more complex events (20 arguments vs. two).

This paper is also complementary to systems which detect event-relevant tweets (Sakaki et al., 2010; Petrović et al., 2010). In future work, we plan to replace our simple method of extracting relevant tweets by one of these approaches, producing a system that monitors microblogs in realtime to automatically construct event-specific knowledge bases.

Our work uses the framework of distant supervision, which has also received considerable attention recently. Nevertheless, most of these works focus on the extraction of binary relations from well-formed documents (Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012). We use the much noisier Twitter as the underlying text, and extract complex events instead of binary relations. We note, however, that the idea of feature aggregation is inspired by these works (Mintz et al., 2009; Riedel et al., 2010), but, to our knowledge, we are the first to apply it to event extraction and sequence tagging. In the DS space, our work is closest to (Reschke et al., 2014), which use it to extract complex events (airplane crashes) from newswire text. Because they focus on newswire, they do not need to address the potential for inaccurate or ambiguous information, which is the main focus of our work.

System	Prec.	Rec.	F1
DS-CRF	66.21	20.66	31.49
DS$^{\text{aggr}}$-CRF	68.27	25.92	37.58 †
DS$^{\text{comb}}$-CRF	61.53	27.61	38.25 †
MA-CRF	68.76	27.61	39.40

Table 6: Test: Replica of the experiments in Table 5 using a threshold of 0.95 for the lenient evaluation measure. All other settings are identical to the experiments in Table 5. † indicates statistically significant improvement over DS-CRF ($p < 0.05$).

8 Discussion: An alternate evaluation measure

Designing relevant measures for lenient evaluations, such as the one discussed here, is an open research issue. For example, the method proposed in Section 4 gives partial credit to all reported (positive) numeric values in the interval $[0, 2g]$, where g is the correct value for the corresponding slot (see the equation in Section 4). But other, stricter, measures

[16] We also tried skip-chain CRFs (Getoor and Taskar, 2007), but found that our simpler approach converges considerably faster and produces slightly better results. We do not show those results for brevity.

are certainly possible.[17] For example, one stricter variant of our proposed measure would assign partial credit only for predicted values that have a similarity of 0.95 or higher with the gold truth (inline with our approximate DS training process). For example, for the same gold numeric value g, the measure assigns partial credit only for predicted values in the interval $[0.95g, 1.05g]$.

We repeated the experiments in Table 5 using this alternate evaluation measure. The result are summarized in Table 6. The results reported in Table 5 do not alter the findings of the paper. In fact, under this stricter evaluation measure, our results are stronger: DS^{comb}-CRF, which combines both our ideas, approaches with nearly 1 F1 point MA-CRF, which trains on manually annotated data.

9 Conclusions

To our knowledge, this is one of the first works that analyzes the problem of distantly supervised complex event extraction on microblogs. This near real-time data source is challenging, with inaccurate information and short, ambiguous texts, as shown by our empirical analysis of the dataset. We proposed two simple techniques to address these problems: (a) a novel distant supervision paradigm, which implements an alignment algorithm that allows text snippets that are similar but not identical to argument values in the knowledge base to be annotated (thus producing better training data); and (b) a feature aggregation strategy that provides richer information across tweets to cope with ambiguity. Our results on earthquake-related tweets show that each improvement yields 19% significant improvement when applied on top of a strong system based on sequence tagging (CRFs). We show that these contributions are complementary: a model that combines both performs better than each of the above individual models, with an improvement of 33% over the baseline. All in all, our approach attains approximately 88% of the ceiling performance for this task, which is obtained by a system trained on manually-annotated tweets, validating the hypothesis that distant supervision is useful for a complex event extraction task.

In addition, we devised a lenient evaluation measure which incorporates the similarity between the extracted argument values and the gold truth, rather than considering as correct only the extractions that exactly match the gold values. We show that this evaluation models the event extraction task better, and, furthermore, is more realistic, especially in view of imperfect knowledge bases.

Lastly, we release a dataset containing an event knowledge base constructed from Wikipedia information on earthquakes, which contains 108 earthquakes, 20 different argument types, and 1,116 argument values. The dataset also includes a collection of relevant tweets about these earthquakes, totaling 7,841 tweets. The dataset is publicly available.[18]

Acknowledgements

This work was partially funded by MINECO (CHIST-ERA READERS project – PCIN-2013-002-C02-01, EXTRECM project – TIN2013-46616-C2-1-R, SKaTeR project – TIN2012-38584-C06-02), and the European Commission (QTLEAP – FP7-ICT-2013.4.1-610516). Ander Intxaurrondo is supported by a PhD grant from the Basque Country Government. The IXA group is funded by the Basque Government (A type Research Group).

[17] We thank the anonymous reviewer for the suggestion.

[18] http://ixa.eus/Ixa/Argitalpenak/Artikuluak/1425465524/publikoak/earthquake-kb-dataset.zip

References

Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identication on twitter. In *Proceedings of the Conference of the Association for the Advancement of Articial Intelligence*.

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 995–1005, Stroudsburg, PA, USA. Association for Computational Linguistics.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The automatic content extraction (ace) program – tasks, data, and evaluation. In *Proceedings of LREC*.

L. Getoor and B. Taskar, 2007. *Introduction to statistical relational learning*. MIT Press.

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

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189, Los Angeles, California, June. Association for Computational Linguistics.

Ana-Maria Popescu, Marco Pennacchiotti, and Deepa Paranjpe. 2011. Extracting events and event descriptions from twitter. In *Proceedings of the 20th International Conference on World Wide Web*.

Kevin Reschke, Martin Jankowiak, Mihai Surdeanu, Christopher D. Manning, and Daniel Jurafsky. 2014. Event extraction using distant supervision. In *Proceedings of LREC*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD '10)*.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of KDD*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA. ACM.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*.

Mihai Surdeanu. 2013. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *Proceedings of the TAC-KBP 2013 Workshop*.