Efficient Named Entity Annotation through Pre-empting

Leon Derczynski
University of Sheffield
S1 4DP, UK
leon@dcs.shef.ac.uk

Kalina Bontcheva
University of Sheffield
S1 4DP, UK
kalina@dcs.shef.ac.uk

Abstract

Linguistic annotation is time-consuming and expensive. One common annotation task is to mark entities – such as names of people, places and organisations – in text. In a document, many segments of text often contain no entities at all. We show that these segments are worth skipping, and demonstrate a technique for reducing the amount of entity-less text examined by annotators, which we call "pre-empting". This technique is evaluated in a crowdsourcing scenario, where it provides downstream performance improvements for the same size corpus.

1 Introduction

Annotating documents is expensive. Given the dominant position of statistical machine learning for many NLP tasks, annotation is unavoidable. It typically requires an expert, but even non-expert annotation work (cf. crowdsourcing) has an associated cost. This makes it important to get the maximum value out of annotation.

However, in entity annotation tasks, annotators sometimes are faced with passage of text which bear no entities. These blank examples are especially common outside of the newswire genre, in e.g. social media text (Hu et al., 2013). While finding good examples to annotate next is a problem that has been tackled before, these systems often require a tight feedback loop and great control over which document is presented next. This is not possible in a crowdsourcing scenario, where large volumes of documents need to be presented for annotation simultaneously in order to leverage crowdsourcing's scalability advantages. The loosened feedback loop, and requirement to issue documents in large batches, differentiate the problem scenario from classical active learning.

We hypothesise that these blank examples are of limited value as training data for statistical entity annotation systems, and that it is preferable to annotate texts containing entities over texts without them. This proposition can be evaluated directly, in the context of named entity recognition (NER). If correct, it offers a new pre-annotation task: predicting whether an excerpt of text will contain an entity we are interested in annotating.

The goal is to reduce the cost of annotation, or alternatively, to increase the performance of a system that uses a fixed amount of data. As this pre-annotation task tries to acquire information about entity annotations before they are actually created – specifically, whether or not they exist – we call the task "pre-empting".

Unlike many modern approaches to optimising annotated data, which focus on how to best leverage annotations (perhaps by making inferences over those annotations, or by using unlabelled data), we examine the step before this – selecting what to annotate in order to boost later system performance.

In this paper, we:

- demonstrate that entity-bearing text results in better NER systems;
- introduce an entity pre-empting technique;
- examine how pre-empting entities optimises corpus creation, in a crowdsourcing scenario.

2 Validating The Approach

The premise of entity pre-empting is that entity-bearing text is better NER training data than entity-less text. To check this, we compare performance with entity-bearing vs. entity-less and also unsorted text. Our scenario has a base set of 2 000 sentences annotated for named entities. We add different kinds of sentences to this base set, and see how an NER system performs when trained on them. This mimics the situation where one has a

Dataset	P	R	F1
Base: 2k sentences	76.55	70.65	73.48
2k sents + 2k without entities	78.03	66.12	71.58
2k sents + 2k random	79.29	76.36	77.80
2k sents + 2k with entities	79.80	77.78	**78.77**

Table 1: Adding entity-less vs. entity-bearing data to a 2 000-sentence base training set

Dataset	P	R	F1	Δ F1
Base: All sentences	85.70	84.08	84.88	-
- 2k without entities	84.89	84.41	84.65	-0.23
- 2k with entities	85.43	83.17	84.29	-0.59

Table 2: Removing data from our training set

base corpus of quality annotated data and intends to expand this corpus.

2.1 Experimental Setup

For English newswire, we use the CoNLL 2003 dataset (Tjong Kim Sang and Meulder, 2003). The training part of this dataset has 14 040 sentences; of these, 11 131 contain at least one entity and so 2 909 have no entities. We evaluate against the more challenging `testb` part of this corpus, which contains 5 652 entity annotations. We use Finkel et al. (2005)'s statistical machine learning-based NER system.

2.2 Validation Results

Results are shown in Table 1. Adding 2 000 entity-bearing sentences gives the largest improvement in F1, and is better than adding 2 000 randomly chosen sentences – the case without pre-empting. Adding only entity-free text decreases overall performance, especially recall.

To double check, we try removing training data instead of adding it. In this case, removing content *without* entities should hurt performance less than removing content *with* entities. From all 14k sentences of English training data, we remove either 2 000 entity-beering sentences or 2 000 sentences with no entities. Results are given in Table 2.

Although the performance drop is small with this much training data, the drop from removing entity-bearing data is over twice the size of that from removing the same amount of entity-free data. So, examples containing entities are often the best ones to add to an initial corpus, and have a larger negative impact on performance when removed. Being able to pre-empt entities is valuable, and can improve corpus effectiveness.

3 Pre-empting Entity Presence

Having defined the pre-empting task, we take two approaches to investigate the practicality of pre-empting named entities in English newswire text. The first is discriminative learning. We use maximum entropy and SVM classifiers (Daumé III, 2004; Joachims, 1999); we experiment with cost-weighted SVM in order to achieve high recall (Morik et al., 1999). The second is to declare sentences containing proper nouns as entity-bearing. We use a random baseline that predicts NE presence based on the prior proportion of entity-bearing to entity-free sentences ($\approx$4.8:1, entity-bearing is the dominant class, for any entity type).

For the machine learning approach, we use the following feature representations: character 1,2,3-grams; compressed word shape 1,2,3 grams;[1] and token 1,2,3 grams.

For the proper noun-based approach, we use the Stanford tagger (Toutanova et al., 2003) to label sentences. This is trained on Wall Street Journal data which does not overlap with the Reuters data in our NER corpus.

As data we use a base set of sentences as training examples, which are a mixture of entity-bearing and entity-free. We experiment with various sizes of base set. Evaluation is performed over a separate 4 000-sentence set, labelled as either having or not having any entities.

3.1 English Newswire, Any Entity

Intrinsic evaluation of these pre-empting approaches is made in terms of classification accuracy, precision, recall and F1. Results are given in Table 3. They indicate that our approach to pre-empting over all entity types in English newswire performs well. For SVM, few entity-bearing sentences were excluded by not being pre-empted (false negatives), and we achieved high precision. Maximum entropy achieved similar results, with the highest overall F-scores. We obtain close to oracle performance with little training data – a set of one hundred sentences affords a high overall performance. Repeating the experiment on the separate CoNLL evaluation set (gathered months after the training data, and so over some different entity

[1] Word shape reflects the capitalisations and classes of letters within a word; for example, "you" becomes "xxx" and "Free!" becomes "Xxxx." Compression turns runs of the same character into one, like an inverse + regex operator; this gives word shape representations "x" and "Xx." respectively.

Training sents.	Accuracy	P	R	F1
Random baseline				
	68.77	78.90	82.28	80.55
Proper nouns				
WSJ	72.43	92.29	92.14	92.22
MaxEnt				
10*	75	85	83	84
100*	83.3	83.9	97.5	90.1
1000	90.38	93.04	94.85	93.94
5000	94.25	96.25	96.44	96.35
10000	95.08	96.56	97.20	**96.88**
Plain SVM				
10*	79	79	100	88
100*	78.6	78.6	100	88.0
1000	90.58	92.12	96.25	91.34
5000	93.28	96.06	95.36	94.65
10000	94.22	96.46	96.18	**95.33**
SVM + Cost, $j = 5$				
10*	79	79	100	88
100*	78.6	78.6	100	88.0
1000	86.33	86.53	97.84	86.43
5000	92.12	92.36	98.09	92.24
10000	94.15	94.25	98.57	**94.20**

Table 3: Evaluating entity pre-empting on English newswire. *We report figures at 2s.f. and 3s.f. for results with 10 and 100 examples respectively, as the training set is small enough to make higher precision inappropriate.

Training data	P	R	F1
500 base + 500 random	74.33	68.56	71.33
500 base + 500 pre-empted	74.80	69.43	**72.01**

Table 4: Entity recognition performance with random vs. pre-empted sentences

names) gives similar results; for example the pre-empting SVM trained on 100 examples from the training set performs with 79.81% precision and full recall, and with 1000 examples, 87.92% precision and near-full recall (99.53%). Even though entity-bearing sentences are the dominant class, we can still increase entity presence in a notable proportion of the training corpus.

3.2 Extrinsic Evaluation

It is important to measure the real impact of pre-empting on the resulting NER training data. To this end, we use 500 hand-labelled sentences as base data to train a pre-empting SVM, and add a further 500 sentences to this. We compare NER performance of a system trained on the base 500 + 500 random sentences, to that of one using 500 + 500 pre-empted entity-bearing sentences. As before, evaluation is against the `testb` set. Table 4 show results. Performance is better with pre-empted annotations, though so many sentences bear entities that the change in training data – and resultant effect – is small.

Language	Accuracy	P	R	F1
Random baseline				
Dutch	49.0	46.7	46.4	46.5
Spanish	63.2	76.1	75.4	75.8
Hungarian	57.1	69.3	68.5	68.9
SVM				
Dutch	92.9	89.9	98.2	93.9
Spanish	76.2	76.2	100	86.5
Hungarian	70.7	70.4	99.9	82.6

Table 5: Pre-empting performance for Dutch, Spanish and Hungarian

Training data	P	R	F1
Dutch, 3 926 entities			
100 base + 500 random	63.57	48.80	**55.22**
100 base + 500 pre-empted	62.46	44.93	52.27
Spanish, 3 551 entities			
100 base + 500 random	68.38	61.71	64.90
100 base + 500 pre-empted	73.00	66.91	**69.82**
Hungarian, 2 432 entities			
100 base + 500 random	76.55	67.52	**71.75**
100 base + 500 pre-empted	72.84	61.43	66.65

Table 6: Entity recognition performance with random vs. pre-empted sentences for Dutch, Spanish and Hungarian

3.3 Other Languages

Pre-empting is not restricted to just English. Similar NER datasets are available for Dutch, Spanish and Hungarian (Tjong Kim Sang, 2002; Szarvas et al., 2006). Results regarding the effectiveness of an SVM pre-empter for these languages are presented in Table 5. In each case, we train with 1 000 sentences and evaluate against a 4 000-sentence evaluation partition.

Strong above-baseline performance was achieved for each language. For Dutch and Spanish, this pre-empting approach performs in the same class as for English, with a low error rate. The error rate is markedly higher in Hungarian, a morphologically-rich language. This could be attributed to the use of token n-gram features; one would expect these to be sparser in a language with rich morphology, and therefore being harder to build decision boundaries over.

For extrinsic evaluation, we use a pre-empter trained with 100 sentences and then compare the performance benefits of adding either 500 randomly-selected sentences or 500 pre-empted sentences to this training data. The same NER system is used to learn to recognise entities. Results are given in Table 6. Pre-empting did not help in Hungarian and Dutch, though was useful for Spanish. This indicates that the pre-empting hypothesis

may not hold for every language, or every genre. But as far as we can see, it certainly holds for English, and also for Spanish.

4 Crowdsourced Corpus Annotation

As pre-empting entities is useful during corpus creation, in this section we examine how to apply it with an increasingly popular new annotation method: crowdsourcing. Crowdsourcing annotation works by presenting a many *microtasks* to non-expert workers. They typically make their judgements over short texts, after reading a short set of instructions (Sabou et al., 2014). Such judgments are often simpler than those in linguistic annotation by experts; for example, workers might be asked to annotate only a single class of entity at a time. Through crowdsourcing, quality annotations can be gathered quickly and at scale (Aker et al., 2012).

There also tends to be a larger variance in reliability over crowd workers than in expert annotators (Hovy et al., 2013). For this reason, crowdsourced annotation microtasks are often all performed by at least two different workers. E.g., every sentence would be examined for each entity type by at least two different non-expert workers.

We investigate entity pre-empting of crowdsourced corpora for a challenging genre: social media. Newswire corpora are not too hard to come by, especially for English, and the genre is somewhat biased in style, mostly being written or created by working-age middle-class men (Eisenstein, 2013), and in topic, being related to major events around unique entities that one might refer to by a special name. In contrast, social media text has broad stylistic variance (Hu et al., 2013) while also being difficult for existing NER tools to achieve good accuracy on (Derczynski et al., 2013; Derczynski et al., 2015) and having no large NE annotated corpora.

In our setup, we subdivide the annotation task according to entity type. Workers perform best with light cognitive loads, so asking them to annotate one kind of thing at a time increases their agreement and accuracy (Krug, 2009; Khanna et al., 2010). Person, location and organisation entities are annotated, giving three annotation sub-tasks, following Bontcheva et al. (2015). Jobs were created automatically using the GATE crowdsourcing plugin (Bontcheva et al., 2014). An example sub-task is shown in Figure 1. This means that we must pre-empt according to entity type, instead of just pre-empting whether or not an excerpt contains any entities at all, which has the additional effect of changing entity-bearing/entity-free class distributions.

Entity type	Messages with	Messages without
Any	45.95%	54.05%
Location	9.52%	90.48%
Organisation	11.16%	88.84%
Person	32.49%	67.51%

Table 7: Entity distribution over twitter messages

Dataset	P	R	F1
Base: 500 messages	70.39	31.66	43.67
500 msgs + 1k without entities	85.00	25.15	38.81
500 msgs + 1k random	76.14	44.38	56.07
500 msgs + 1k with entities	71.21	54.14	**61.51**

Table 8: Adding entity-less vs. entity-bearing data to a 500-message base training set

We use two sources that share entity classification schemas: the UMBC twitter NE annotations (Finin et al., 2010), and the MSM2013 twitter annotations (Rowe et al., 2013). We also add the Ritter et al. (2011) dataset, mapping its geo-location and facility classes to location, and company, sports team and band to organisation. Mixing datasets reduces the impact of any single corpus' sampling bias on final results. In total, this gives 3 854 twitter messages (tweets). Table 7 shows the entity distribution over this corpus. From this we separated a 500 tweet training set, used as base NER training data and pre-empting training data, and another set of 500 tweets for evalution. Note that each message can contain more than one type of entity, and that names of people are the most common class of entity.

4.1 Re-validating the Hypothesis

As we now have a new dataset with potentially much greater diversity than newswire, our first step is to re-check our initial hypothesis – that entity-bearing text contributes more to the performance of a statistical NER system than entity-free or random text. Results are shown in Table 8.

The effect of entity-bearing training data is clear here. Only data without annotations to the base is harmful (-4.8 F1), adding randomly chosen messages is helpful (+14.4 F1), and adding only messages containing entities is the most helpful (+17.8 F1). The corpus is small; in this case, the evaluation data has only 338 entities. Even so, the difference between random and entity-only F1 is signif-

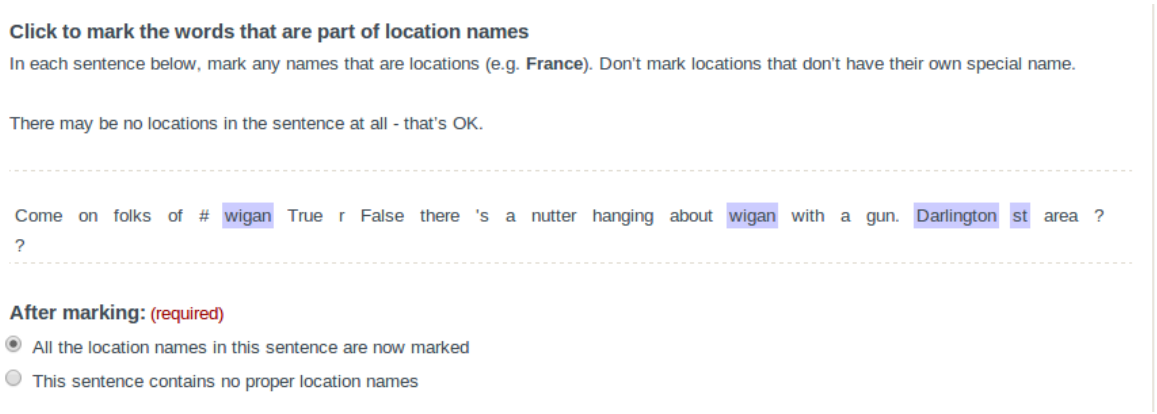

Figure 1: An example crowdsourced entity labelling microtask.

Training sents.	Accuracy	P	R	F1
Random baseline				
	51.6	47.1	48.2	47.6
Proper nouns				
From WSJ	54.0	49.8	85.4	62.9
SVM + Cost, $j = 5$				
10	46	46	100	63
100	69.5	63.0	80.3	70.6
200	72.4	66.9	78.4	72.2
500	71.4	64.8	81.7	72.3
1000	47.7	68.0	83.6	75.1

Table 9: Evaluating any-entity tweet pre-empting.

Entity type	Acc.	P	R	F1
Random baseline				
Person	56.63	33.33	33.87	33.60
Location	83.17	10.91	11.32	11.11
Organisation	80.08	8.86	9.09	8.97
SVM + Cost, $j = 5$				
Person	74.87	65.69	70.10	67.77
Location	91.27	64.81	13.21	21.95
Organisation	89.55	60.42	9.42	16.30
Maximum entropy				
Person	80.15	60.67	73.39	66.43
Location	90.85	7.92	55.26	13.86
Organisation	89.38	7.79	55.81	13.68

Table 10: Per-entity pre-empting on tweets.

icant at p<0.00050, using compute-intensive χ^2 testing following Yeh (2000).

4.2 Pre-empting Entities in Social Media

We construct a similar pre-empting classifier to that for newswire (Section 3.1). We continue using the base 500 messages as a source of training data, and evaluate pre-empting using the remainder of the data. The random baseline follows the class distribution in the base set, where 47.2% of messages have at least one entity of any kind.

We also evaluate pre-empting performance per entity class. The same training and evaluation sets are used, but a classifier is learned to pre-empt each entity class (person, location and organisation), as in Derczynski and Bontcheva (2014). This may greatly impact annotation, due to the one-class-at-a-time nature of the crowdsourced task and low occurrence of individual entity types in the corpus (see Table 7). We took 300 of the base set's sentences and used these for our training data, with the same evaluation set as before.

4.3 Results

Results for any-entity pre-empting on tweets are given in Table 9. Although performance is lower than on newswire, pre-empting is still possible in this genre. Only results for cost-weighted SVM are given.

We were able to learn accurate per-entity classifiers despite having a fairly small amount of data. Results are shown in Table 10. A good reduction is achieved over the baseline in all cases, though specifically predicting locations and organisations is hard. However, we do achieve high precision, meaning that a good amount of less-useful entity-free data is rejected. The SVM figures are with a reasonably high weighting in favour of recall. Conversely, while achieving similar F-scores to SVM, the maximum entropy pre-empter scores much better in terms of recall than precision.

These results are encouraging in terms of cost reduction. In this case, once we have annotated the first few hundred examples, we can avoid a lot of un-needed annotation by only paying crowd workers to complete microtasks on texts we suspect (with great accuracy) bear entities. From the observed entity occurrence rates in Table 7, given our pre-empting precision, we can avoid 41% of person microtasks, 59% of location microtasks and

Removed features	Acc.	P	R
Baseline			
None	90.58	92.12	96.25
-gram shortening			
3-grams	90.50	92.29	95.93
2-grams	90.15	91.62	96.28
1-grams	89.09	90.13	96.69
Removed feature classes			
Char-grams	87.47	89.46	95.29
Shape-grams	87.20	87.73	97.33
Token-grams	90.33	92.56	95.36

Table 11: Pre-empting feature ablation results.

58% of organisation microtasks where no entities occur – excluding a large amount material in preference for content that will give better NER performance later.

5 Analysis

5.1 Feature Ablation

The SVM system we have developed for pre-empting named entities is effective. To investigate further, we performed feature ablation along two dimensions. Firstly, we hid certain feature n-gram lengths (which are 1, 2 or 3 entries long). Secondly, we removed groups of features i.e. word n-grams, character n-grams or compressed word shape n-grams. We experimented using 1 000 training examples, on the newswire all-entities task, evaluating against the same 4 000-sentence evaluation set, with an SVM pre-empter. This makes figures comparable to those in Table 3.

Ablation results are given in Table 11. Shape grams, that is, subsequences of word characters, have the least overall impact on performance. Unigram features (across all character, shape and token groups) have the second-largest impact. This suggests that morphological information is useful in this task, and that the presence of certain words in a sentence acts as a pre-empting signal.

5.2 Informative Features

When pre-empting certain features are more helpful than others. The maximum entropy classifier implementation used allows output of the most informative features. These are reported – for newswire – in Table 12. In this case, the model was trained on 10 000 examples, and is the one for which results were given in Table 3, that achieved an F-score of 96.88.

Word shape features are the strongest indicators of named entity presence, and the strongest indicators of entity absence are all character grams.

Feature type	Feature value	Weight
shape	X_.	0.99558
char-gram	K	1.06190
shape	._	1.10804
shape	Xx_Xx_x	1.17046
shape	X	1.39189
shape	x_Xx_x	1.40092
shape	Xx_Xx	1.56733
shape	x_Xx	1.77390
shape	._.	-1.40075
char-gram	”	-1.03842
shape	x	-0.96047
char-gram	G_	-0.85378
char-gram	T_	-0.80422
char-gram	H_e_	-0.77069
n-gram	He	-0.77069
char-gram	I_	-0.75819

Table 12: Strongest features for pre-empting in English newswire.

Many shapes that indicate entity presence have one or more capitalised words in sequence, or linked to all-lower case words surrounding them. Apparently, sentences containing quote marks are less likely to contain named entities. Also, the characters sequence “He” suggests that a sentence does not contain an entity, perhaps because the target is being referred to pronomially.

5.3 Observations

Our experiments have begun with a base set of annotated sentences, mixing entity-bearing and entity-free. This not only serves a practical purpose of providing the pre-empter with training data and negative examples. It is also important to include some entity-free text in the NER training data so that systems based on it can observe that some sentences may have no entities. Without this observation, there is a risk that they will handle entity-free sentences poorly when labelling previously-unseen data.

It should be noted that segmenting into sentences risks the removal of long-range dependencies important in NER (Ratinov and Roth, 2009). However, overall performance in newswire – on longer documents – is not harmed by our approach. In the social media context we examined, entity co-reference is rare, due to its short texts.

6 Related Work

Avoiding needless annotation is a constant theme in NLP, and of interest to researchers, who often go to great lengths to avoid it. For example, recently, Garrette and Baldridge (2013) demon-

strated the impressive construction of a part-of-speech tagger based on just two hours' annotation.

Similar to our work, Shen et al. (2004) proposed active learning for named entity recognition annotation, reducing annotation load without hurting NER performance, based on three metrics for each text batch and an iterative process. We differ from Shen et al. by giving a one-shot approach which does not need iterative re-training and is simple to implement in an annotation workflow, although we do not reduce annotation load as much. Our simplification means that pre-empting is easy to integrate into an annotation process, especially important for e.g. crowdsourced annotation, which is cheap and effective but gives a lot less control over the annotation process.

Laws et al. (2011) experiment with combining active learning and crowdsourcing. They find that not only does active learning generate better quality than randomly selecting crowd workers, it can be used to filter out miscreant workers. The goal in this work was to improve annotation quality and reduce cost that way. Recent advances in crowdsourcing technology offer much better quality than at the time of this paper. Rather than focusing on finding good workers, we aim for the extrinsic goal improving system performance by choosing which annotations to perform in the first place.

7 Conclusion

Entity pre-empting makes corpus creation quicker and more cost-effective. Though demonstrated with named entity annotation, it can apply to other annotation tasks, especially when for corpora used in information extraction, for e.g. relation extraction and event recognition.

This paper presents the pre-empting task, shows that it is worthwhile, and demonstrates an example approach in two application scenarios. We demonstrate that choosing to annotate texts that are rich in target entity mentions is more efficient than annotating randomly selected text. The example approach is shown to successfully pre-empt entity presencce classic named entity recognition. Applying pre-empting to the social media genre, where annotated corpora are lacking and NER is difficult, also offers improvement – but is harder.

Further analysis of the effect of pre-empting in different languages is also warranted, after the mixed results in Table 6. Larger samples can be used for training social media pre-empting; though we only outline an approach using 1 000 examples, up to 15 000 have been annotated and made publicly available for some entity types. For future work, the pre-empting feature set could be first adapted to morphologically rich languages, and then also to languages that do not necessarily compose tokens from individual letters, such as Mi'kmaq or Chinese.

Acknowledgments

This work is part of the uComp project (http://www.ucomp.eu//), which receives the funding support of EPSRC EP/K017896/1, FWF 1097-N23, and ANR-12-CHRI-0003-03, in the framework of the CHIST-ERA ERA-NET.

References

A. Aker, M. El-Haj, M.-D. Albakour, and U. Kruschwitz. 2012. Assessing crowdsourcing quality through objective tasks. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 1456–1461.

K. Bontcheva, I. Roberts, L. Derczynski, and D. Rout. 2014. The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

K. Bontcheva, L. Derczynski, and I. Roberts. 2015. Crowdsourcing named entity recognition and entity linking corpora. In N. Ide and J. Pustejovsky, editors, *The Handbook of Linguistic Annotation (to appear)*. Springer.

H. Daumé III. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`, August.

L. Derczynski and K. Bontcheva. 2014. Passive-aggressive sequence labeling with discriminative post-editing for recognising person entities in tweets. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 69–73.

L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. 2013. Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM.

L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, and K. Bontcheva. 2015. Analysis of named entity recognition and linking for

tweets. *Information Processing and Management*, 51:32–49.

J. Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369. Association for Computational Linguistics.

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.

J. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

D. Garrette and J. Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of NAACL-HLT*, pages 138–147.

D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy. 2013. Learning Whom to trust with MACE. In *Proceedings of NAACL-HLT*, pages 1120–1130.

Y. Hu, K. Talamadupula, S. Kambhampati, et al. 2013. Dude, srsly?: The surprisingly formal nature of Twitter's language. *Proceedings of ICWSM*.

T. Joachims. 1999. Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 19(4).

S. Khanna, A. Ratan, J. Davis, and W. Thies. 2010. Evaluating and improving the usability of Mechanical Turk for low-income workers in India. In *Proceedings of the first ACM symposium on computing for development*. ACM.

S. Krug. 2009. *Don't make me think: A common sense approach to web usability*. Pearson Education.

F. Laws, C. Scheible, and H. Schütze. 2011. Active learning with amazon mechanical turk. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1546–1556. Association for Computational Linguistics.

K. Morik, P. Brockhausen, and T. Joachims. 1999. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Proceedings of the 16th International Conference on Machine Learning (ICML-99)*, pages 268–277, San Francisco.

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. Ritter, S. Clark, Mausam, and O. Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK.

M. Rowe, M. Stankovic, A. Dadzie, B. Nunes, and A. Cano. 2013. Making sense of microposts (#msm2013): Big things come in small packages. In *Proceedings of the WWW Conference - Workshops*.

M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the 9th international conference on language resources and evaluation (LREC14)*, pages 859–866.

D. Shen, J. Zhang, J. Su, G. Zhou, and C.-L. Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

G. Szarvas, R. Farkas, L. Felföldi, A. Kocsor, and J. Csirik. 2006. A highly accurate named entity corpus for hungarian. In *Proceedings of International Conference on Language Resources and Evaluation*.

E. F. Tjong Kim Sang and F. D. Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

E. F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pages 173–180.

A. Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the conference on Computational linguistics*, pages 947–953. Association for Computational Linguistics.