Twitter at the Grammys: A Social Media Corpus for Entity Linking and Disambiguation

Mark Dredze, Nicholas Andrews, Jay DeYoung
Human Language Technology Center of Excellence
Johns Hopkins University
810 Wyman Park Drive
Baltimore, MD 20211 USA
{mdredze,noa}@cs.jhu.edu

Abstract

Work on cross document coreference resolution (CDCR) has primarily focused on news articles, with little to no work for social media. Yet social media may be particularly challenging since short messages provide little context, and informal names are pervasive. We introduce a new Twitter corpus that contains entity annotations for entity clusters that supports CDCR. Our corpus draws from Twitter data surrounding the 2013 Grammy music awards ceremony, providing a large set of annotated tweets focusing on a single event. To establish a baseline we evaluate two CDCR systems and consider the performance impact of each system component. Furthermore, we augment one system to include temporal information, which can be helpful when documents (such as tweets) arrive in a specific order. Finally, we include annotations linking the entities to a knowledge base to support entity linking. Our corpus is available: https://bitbucket.org/mdredze/tgx

1 Entity Disambiguation

Who is who and what is what? Answering such questions is usually the first step towards deeper semantic analysis of documents, e.g., extracting relations and roles between entities and events. Entity disambiguation identifies real world entities from textual references. Entity linking – or more generally Wikification (Ratinov et al., 2011) – disambiguates reference in the context of a knowledge base, such as Wikipedia (Cucerzan, 2007; McNamee and Dang, 2009; Dredze et al., 2010; Zhang et al., 2010; Han and Sun, 2011). Entity linking systems use the name mention and a context model to identify possible candidates and disambiguate similar entries. The context model includes a variety of information from the context, such as the surrounding text or facts extracted from the document. Though early work on the task goes back to Cucerzan (2007), the name entity linking was first introduced as part of TAC KBP 2009 (McNamee and Dang, 2009).

Without a knowledge base, cross-document coreference resolution (CDCR) clusters mentions to form entities (Bagga and Baldwin, 1998b). Since 2011, CDCR has been included as a task in TAC-KBP (Ji et al., 2011) and has attracted renewed interest (Baron and Freedman, 2008b; Rao et al., 2010; Lee et al., 2012; Green et al., 2012; Andrews et al., 2014). Though traditionally a task restricted to small collections of formal documents (Bagga and Baldwin, 1998b; Baron and Freedman, 2008a), recent work has scaled up CDCR to large heterogenous corpora, e.g. the Web (Wick et al., 2012; Singh et al., 2011; Singh et al., 2012).

While both tasks have traditionally considered formal texts, recent work has begun to consider informal genres, which pose a number of interesting challenges, such as increased spelling variation and (especially for Twitter) reduced context for disambiguation. Yet entity disambiguation, which links mentions across documents, is especially important for social media, where understanding an event often requires reading multiple short messages, as opposed to news articles, which have extensive background information. For example, there have now

been several papers to consider named entity recognition in social media, a key first step in an entity disambiguation pipeline (Finin et al., 2010; Liu et al., 2011; Ritter et al., 2011; Fromreide et al., 2014; Li et al., 2012; Liu et al., 2012; Cherry and Guo, 2015; Peng and Dredze, 2015). Additionally, some have explored entity linking in Twitter (Liu et al., 2013; Meij et al., 2012; Guo et al., 2013), and have created datasets to support evaluation. However, to date no study has evaluated CDCR on social media data,[1] and there is no annotated corpus to support such an effort.

In this paper we present a new dataset that supports CDCR in Twitter: the TGX corpus (Twitter Grammy X-doc), a collection of Tweets collected around the 2013 Grammy music awards ceremony. The corpus includes tweets containing references to people, and references are annotated both for entity linking and CDCR. To explore this task for social media data and consider the challenges, opportunities and the performance of state of the art CDCR methods, we evaluate two state-of-the-art CDCR systems. Additionally, we modify one of these systems to incorporate temporal information associated with the corpus. Our results include improved performance for this task, and an analysis of challenges associated with CDCR in social media.

2 Corpus Construction

A number of datasets have been developed to evaluate CDCR, and since the introduction of the TAC-KBP track in 2009, some now include links to a KB (e.g. Wikipedia). See Singh et al. (2012) for a detailed list of datasets. For Twitter, there have been several recent entity linking datasets, all of which number in the hundreds of tweets (Meij et al., 2012; Liu et al., 2013; Guo et al., 2013). None are annotated to support CDCR.

Our goal is the creation of a Twitter corpus to support CDCR, which will be an order of magnitude larger than corresponding Twitter corpora for entity linking. We created a corpus around the 2013 Grammy Music Awards ceremony. The popular ceremony lasted several hours generating many tweets. It included many famous people that are in Wikipedia, making it suitable for entity linking and aiding CDCR annotation. Additionally, Media personalities often have popular nicknames, creating an opportunity for name variation analysis.

Using the Twitter streaming API[2], we collected tweets during the event on Feb 10, 2013 between 8pm and 11:30pm Eastern time (01:00am and 04:30 GMT). We used Carmen geolocation[3] (Dredze et al., 2013) to identify tweets that originated in the United States or Canada and removed tweets that were not identified as English according to the Twitter metadata. We then selected tweets containing "grammy" (case insensitive, and including "#grammy"), reducing 564,892 tweets to 50,429 tweets. Tweets were processed for POS and NER using Twitter NLP Tools [4] (Ritter et al., 2011). Tweets that did not include a person mention were removed. Using an automated NER system may miss some tweets, especially those with high variation in person names, but it provided a fast and effective way to identify tweets to include in our data set. For simplicity, we randomly selected a single person reference per tweet.[5] The final set contained 15,736 tweets.

We randomly selected 5,000 tweets for annotation, a reasonably sized subset for which we could ensure consistent annotation. Each tweet was examined by two annotators who grouped the mentions into clusters (CDCR) and identified the corresponding Wikipedia page for the entity if it existed (entity linking). As part of the annotation, annotators fixed incorrectly identified mention strings. Similar to Guo et al. (2013), ambiguous mentions were removed, but unlike their annotations, we kept all persons including those not in Wikipedia. Mentions that were comprised of usernames were excluded.

The final corpus contains 4,577 annotated tweets, 10,736 unlabeled tweets, and 273 entities, of which 248 appear in Wikipedia. The corpus is divided into five folds by entity (about 55 entities per fold),

[1] Andrews et al. (2014) include CDCR results on an early version of our dataset but did not provide any dataset details or analysis. Additionally, their results averaged over many folds, whereas we will include results on the official dev/test splits.

[2] https://dev.twitter.com/streaming/reference/get/statuses/sample

[3] https://github.com/mdredze/carmen

[4] https://github.com/aritter/twitter_nlp

[5] In general, **within** document coreference is run before CDCR, and the cross-document task is to cluster within-document coreference chains. In our case, there were very few mentions to the same person within the same tweet, so we did not attempt to make within-document coreference decisions.

Mentions per entity: mean	16.77
Mentions per entity: median	1
Number of entities	273
Number of mentions (total tweets)	15,313
Number of unique mention strings	1,737
Number of singleton entities	166
Number of labeled tweets	4,577
Number of unlabeled tweets	10,736
Words/tweet (excluding name): mean	10.34
Words/tweet (excluding name): median	9

Table 1: Statistics describing the TGX corpus.

where splits were obtained by first sorting the entities by number of mentions, then doing systematic sampling of the entities on the sorted list. The first split is reserved for train/dev purposes and the remaining splits are reserved for testing. This allows for a held out evaluation instead of relying on cross-validation, which ensures that future work can conduct system development without the use of the evaluation set. Some summary statistics appear in Table 1 and examples of entities in Table 2. The full corpus, including annotations (entity linking and CDCR), POS and NER tags are available at `https://bitbucket.org/mdredze/tgx`.[6]

3 Models

We consider two recent models that represent state-of-the-art performance on CDCR. While TGX has entity linking annotations, we focus on CDCR since Twitter entity linking has been previously explored.

Green et al. (2012) (GREEN) developed a pipeline system for cross document entity disambiguation. First, entities with dissimilar mention strings are identified via "cannot-link" constraints. Then, subject to these constraints, entities are disambiguated based on context via a hierarchical clustering step. Neither of the two steps requires explicit supervision, but instead relies on the careful tuning of hyperparameters. In our experiments, we use a grid search to find the hyperparameters that yield the highest score on the development split, and then use those same hyperparameters for testing with no further tuning. We compare the performance of the full pipeline (FULL), as well as a variation which does no disambiguation (NO-CONTEXT).

[6]Permitted by the Twitter terms of service: `https://dev.twitter.com/overview/terms/agreement-and-policy`

Andrews et al. (2014) (PHYLO) developed a generative model for clustering entities across documents based on name and context similarity.[7] Their work extended a phylogenetic name model (Andrews et al., 2012) that learns groups of name variations through string transducers by composing a phylogeny of name variation based on unlabeled data. As above, we present versions of the model with both context and name matching (FULL) as well as without context (NO-CONTEXT). Parameters are tuned on dev data as with GREEN.

A unique property of TGX is its temporal ordering, where documents are timestamped and time impacts entity priors. Figure 4 shows the number of mentions for the top 10 entities over time. The curves are highly peaked, suggesting that there is a small window in time in which the entity is popular, though there are occurrences over the whole event.

We modify PHYLO to include consider temporal information. The model is a generative account of the process by which authors choose particular name spellings, either by copying some existing spelling (possibly introducing variation) or coming up with new names from scratch. This process is modeled in two parts: (1) a name model which assigns probabilities to different spelling variations, and a (2) parent model which assigns probabilities to different parent-child relationships. The parent-child relations give ancestral lineages which form a phylogenetic tree; the connected components of this tree give a partition of names into entity clusters.

Andrews et al. proposed a log-linear model for the parent model to incorporate topic features in order to disambiguate between entities with similar names. By incorporating different features in this log-linear model we give the model more flexibility in explaining the choice of parent. To incorporate temporal information, we introduce features that look at the proximity of pairs of named-entities in time. There are several options for incorporating temporal features; we use a simple overlapping sliding window approach. We use a width of 10 minutes with 5 minute overlaps; every tweet is in two windows except for the first and last 5 minutes. The indicator of a shared bucket fires if a parent and child appear in the same bucket. Unsupervised training can learn

[7]Code available: `https://bitbucket.org/noandrews/phyloinf`

Entity Name	# Mentions	Example Mentions
Taylor Swift	742	taylor,t-swizzle,swift,tswift,taylor freaken swiift,tay,t swift,taylor alison swift
Adelle	370	adel,adelle,adele
Miranda Lambert	266	miranda lambert,lambert,amanda miranda,miranda lamberts,miranda
Carrie Underwood	264	carrie,underwood,carrie underwear,kerry underwoods
Elton John	227	elton j,sir elton,elton,elton john
Johnny Depp	204	johnny deep,johnny,johnny d,johnny jack sparrow,johhny depp,john depp
Ed Sheeran	189	ed sharon,sherran,ed shee-ran,ed sheerannn,ed sheeren,ed sheeeeeeran,ed sheerin
Miguel	182	miguel
Wiz Khalifa	141	khalifa,wizard,wiz kalifa,wiz kahalifa,wiz
Marcus Mumford	140	marcus,marcus mumford,mark mumford,munford

Table 2: The 10 largest entities. 90% of the labeled tweets refer to the 38 most common entities.

Model		Dev. B^3	Test B^3
	EXACT	67.8	69.9
GREEN	NO-CONTEXT	78.0	77.2
	FULL	88.5	79.7
PHYLO	NO-CONTEXT	96.9	72.3
	FULL	97.4	72.1
	FULL+TIME	97.7	72.3

Table 3: CDCR performance (larger B^3 is better).

positive weights for these features by observing that mentions with similar names and contexts, which makes them likely to paired in the phylogeny, are also likely to appear in the same time buckets. We refer to this model as FULL+TIME.

Finally, we compare to an exact mention match baseline (EXACT), which clusters all mentions with identical string mentions.

4 Experiments

Following Green et al. (2012) and other CDCR papers, we report results in terms of B^3 (Bagga and Baldwin, 1998a) in Table 3. First, we note that the performance of EXACT is relatively high. This is attributable to popular artists that use a single alias as a stage name, such as Adele or Beyonce. The result is that these artists are not susceptible to name variation, except for common misspellings. Only 3.6% of the mentions are singletons, so they are unlikely to significantly help this method.

Next, both CDCR models in all configurations improve over the EXACT baseline. While all versions of PHYLO improve over GREEN on development data, the PHYLO models overfit and do worse on test. These results differ from Andrews et al. (2014), which may be due to our hyper-parameter selection method. Additionally, for both models, adding context improves over clustering based on names alone, but test data suffers for PHYLO. Judging by the resulting clusters, context primarily aided in identifying two references to the same entity that had a low name similarity score.

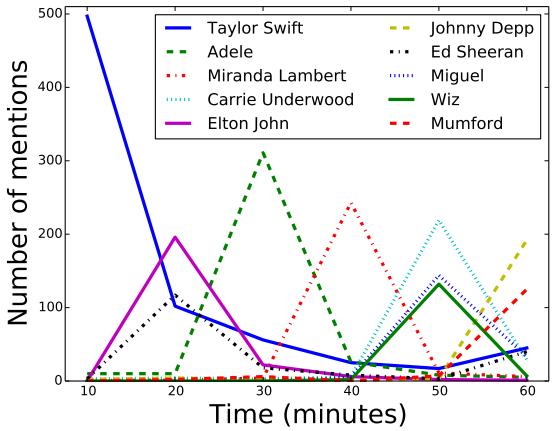

Figure 1: The number of (labeled) mentions for the 10 most common entities shown in 10 minute bins. The entities clearly spike at given points in the dataset. For example, Taylor Swift is most popular in the first few minutes of the data because she performed the opening number.

Analysis An analysis of the mistakes made by the CDCR systems point to several sources of error. While some entities had little name variation (e.g., Adelle and Miguel) aside from spelling errors, others had significant diversity. Table 2 shows the 10 most referenced entities, including number of mentions and variations. People like Taylor Swift have numerous name variations, which include references to nicknames and full names. This name

variation accounted for many of the errors in our output. For instance, the system produced a large high-precision cluster consisting of "Taylor Swift" mentions, and another cluster consisting of the following three mentions: T-Swizzle, TSwift, T-Swift. Similarly, LLCoJ, Llcoolj, LLCOOLJ and LLCoolJ, were incorrectly placed in their own cluster separate from another high-precision cluster consisting of primarily "LL Cool J" mentions. These errors highlight challenges of dealing with informal communications.

Similarly, we found several errors due to superficial name similarity. For instance, the system placed Jessica Biel and Melissa in the same cluster. The system also produced a low-precision cluster LL and Allison Williams, where LL refers to "LL Cool J."

While abbreviations are common sources of errors in newswire for organizations and countries, we saw this for people: Neil Patrick Harris vs. NPH. We also saw more typical variations due to forms of address, e.g., Taylor vs. Taylor Swift, and Mayer vs. John Mayer. We did not see many errors where two entities were confused with each other due to context. Instead, low recall clusters were of the type described above.

Finally, there are several properties of the data unique to social media that could help improve results. First, since our simple time features were helpful, but more sophisticated temporal models could further improve the results. Second, Twitter specific properties, such as hashtags and links, could be integrated into a modified generative model. Third, conversations could provide a larger context for resolution, or aid in identifying name variations for a mention. We plan to consider these directions.

References

Nicholas Andrews, Jason Eisner, and Mark Dredze. 2012. Name phylogeny: A generative model of string variation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Nicholas Andrews, Jason Eisner, and Mark Dredze. 2014. Robust entity clustering via phylogenetic inference. In *Association for Computational Linguistics (ACL)*.

A. Bagga and B. Baldwin. 1998a. Algorithms for scoring coreference chains. In *LREC*.

A. Bagga and B. Baldwin. 1998b. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL*.

A. Baron and M. Freedman. 2008a. Who is Who and What is What: Experiments in cross-document coreference. In *EMNLP*.

Alex Baron and Marjorie Freedman. 2008b. Who is Who and What is What: Experiments in cross-document coreference. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Colin Cherry and Hongyu Guo. 2015. The unreasonable effectiveness of word representations for twitter named entity recognition. In *North America Chapter of Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 708–716.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Conference on Computational Linguistics (Coling)*.

Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *NAACL Workshop on Creating Speech and Language Data With Mechanical Turk*.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating NER for Twitter# drift. In *LREC*.

Spence Green, Nicholas Andrews, Matthew R Gormley, Mark Dredze, and Christopher D Manning. 2012. Entity clustering across languages. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 60–69. Association for Computational Linguistics.

Stephen Guo, Ming-Wei Chang, and Emre Kıcıman. 2013. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of NAACL-HLT*, pages 1020–1030.

Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 945–954, Stroudsburg, PA, USA. Association for Computational Linguistics.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the tac2011 knowledge base population track. In *Text Analysis Conference (TAC)*.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500. Association for Computational Linguistics.

Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. 2012. Twiner: Named entity recognition in targeted twitter stream. In *SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 721–730, New York, NY, USA. ACM.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Association for Computational Linguistics (ACL)*, pages 359–367. Association for Computational Linguistics.

Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. 2012. Joint inference of named entity recognition and normalization for tweets. In *Association for Computational Linguistics (ACL)*, ACL '12, pages 526–535, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaohua Liu, Yitong Li, Haocheng Wu, Ming Zhou, Furu Wei, and Yi Lu. 2013. Entity linking for tweets. In *Association for Computational Linguistics (ACL)*.

Paul McNamee and Hoa Dang. 2009. Overview of the TAC 2009 knowledge base population track. In *Text Analysis Conference (TAC)*.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 563–572. ACM.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Delip Rao, Paul McNamee, and Mark Dredze. 2010. Streaming cross document entity coreference resolution. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1050–1058. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 793–803. Association for Computational Linguistics.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia. Technical report, Technical report, University of Massachusetts.

Michael Wick, Sameer Singh, and Andrew McCallum. 2012. A discriminative hierarchical model for fast coreference at large scale. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 379–388. Association for Computational Linguistics.

Wei Zhang, Jian Su, Chew Lim Tan, and Wen Ting Wang. 2010. Entity linking leveraging: automatically generated annotation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1290–1298, Stroudsburg, PA, USA. Association for Computational Linguistics.