Analysis and Enhancement of Wikification for Microblogs with Context Expansion

Taylor Cassidy, Heng Ji, Arkaitz Zubiaga, Hongzhao Huang
Computer Science Department and Linguistics Department
Queens College and Graduate Center
City University of New York
New York, NY, USA
taylorcassidy64@gmail.com

Lev Ratinov
Google Inc.
New York, NY, USA

ABSTRACT
Disambiguation to Wikipedia (D2W) is the task of linking mentions of concepts in text to their corresponding Wikipedia entries. Most previous work has focused on linking terms in formal texts (e.g. newswire) to Wikipedia. Linking terms in short informal texts (e.g. tweets) is difficult for systems and humans alike as they lack a rich disambiguation context. We first evaluate an existing Twitter dataset as well as the D2W task in general. We then test the effects of two tweet context expansion methods, based on tweet authorship and topic-based clustering, on a state-of-the-art D2W system and evaluate the results.

Author Keywords
Disambiguation to Wikipedia (D2W), Twitter, disambiguation context

INTRODUCTION
Determining the correct meaning of each word in a natural language text is a prerequisite for proper understanding. Disambiguation to Wikipedia (D2W) [23], the process of linking each *concept mention* in a text to a *concept* referent (i.e. a Wikipedia page), is a framework that supports the word sense disambiguation (WSD) task[1]. For example, consider the sentence, "BP said Halliburton destroyed Gulf Spill evidence". A D2W system should break the text into concept mentions and return a unique identifier (an article title, in the case of Wikipedia) for each concept. The intended meaning of each concept mention can be inferred in terms of its surface form and its context.

D2W may benefit both human end-users and natural language processing (NLP) systems. When a document is *Wikified* a reader can more easily grasp its contents as information about related topics is readily accessible[2]. From a system-to-system perspective, a disambiguated corpus has the meanings of many of its terms grounded in a structurally rich ontology, and indeed there is evidence that D2W output [29, 31] can improve NLP systems. Given a concept mention in a source text, and Wikipedia, D2W operates over a representation of the following:

Mention	Wikipedia title
BP	*BP*
said	*Press Release*
Halliburton	*Halliburton*
destroyed	*Spoliation of Evidence*
Gulf Spill	*Deepwater Horizon oil spill*
evidence	*Evidence*

Table 1. Desired D2W output

1. the content of the text, and how its elements are related to the concept mention.
2. the content of Wikipedia, and how its concepts are related to one another.
3. how individual elements of the text are related to elements of Wikipedia.
4. a method for generating candidate concepts for the concept mention.

Each of these items may be represented using the output of Natural Language Processing (NLP) techniques applied to the source text and Wikipedia, and/or an analysis of built-in structure (e.g. TF-IDF, Information Extraction techniques, relationships between documents, structural features of Wikipedia such as links, info boxes, and categories). Most successful D2W applications enumerate potential concept referents for a given concept mention based on the anchor text of already existing links within Wikipedia, as well as information from *redirects* and *disambiguation pages*. Context is extracted from throughout the document where a target concept mention occurs, which is then compared against Wikipedia content to narrow the hypothesis space of potential concepts. The task is therefore more challenging when concept mentions occur in short texts containing informal language.

[1]We use "concept" in both the usual sense and to refer to a Wikipedia page about a concept.

[2]http://en.wikipedia.org/wiki/Wikipedia:Glossary#Wikify.

Over 300 million Twitter users generate over 400 million tweets (posts) daily[3] [4]. The microblogging genre presents unique challenges for NLP tasks. Twitter posts (tweets) are limited to 140 characters and informal language is often used. Contextual evidence is important for accurate D2W, but for tweets it is scattered among various knowledge sources.

In this work we explore ways in which the disambiguation context of concept mentions in tweets can be enhanced. The novel contributions of the paper are as follows. First, we provide a qualitative analysis of a hand-annotated data set [22] and infer some properties of the contextual evidence most likely sought by annotators. Two sources of additional context useful for disambiguation are identified: tweets from the same author, and topically related tweets. In addition, we evaluate the contribution of these additional context types to the performance of GLOW, a state-of-the-art D2W system [30].

RELATED WORK

The task of linking expressions to Wikipedia concepts has received increased attention over the past several years, as the linking of all concept mentions in a single text [23, 24, 25, 18, 13, 30], the linking of a cluster of co-referent named entity mentions spread throughout different documents (Entity Linking) [20, 16, 15, 32, 10, 11], or the linking of a whole tweet to a single concept [8]. Most D2W work has been performed on newswire collections, and most work on tweets has been limited to a particular type of concept mention. For example, the Online Reputation Management Task [1] focused on filtering tweets containing company name to extract only those tweets that were actually related to the company.

For an n-gram deemed a concept mention, most D2W systems define candidate target concepts as a subset of those that were ever linked to using the n-gram in question as anchor text, from within Wikipedia itself (though [33] expanded this set using search engine click results). The relative frequency with which a given n-gram links to each target concept is referred to as its *commonness* distribution[5]. Disambiguation is then couched as re-ranking, computed based on similarity between the concept mention along with its surrounding context, and a candidate concept, The systems of [7, 30, 24, 4, 12] take into account the coherence of all concepts linked to in a given document, based on concept similarity. [22] created the hand-labeled dataset that we use in our work. Their best performing system based on random forests outperforms commonness accord, though it does not ensure any global coherence over the concepts assigned to a given tweet.

Some TAC-KBP Entity Linking [15] systems utilized all entities in the context of a given query, disambiguating all entities simultaneously using a graph-based re-ranking algorithm [6, 28, 5, 9, 10, 11] or a collaborative/ensemble ranking algorithm [27, 3, 17] to ensure global consistency. [21] demonstrated that co-occurring named entities are particularly helpful for Cross-lingual Entity Linking (CLEL). None of the TAC-KBP systems performed full-document D2W to include concept mentions of different types, including non-entities.

For a given concept mention, all-concept D2W work we are aware of makes use of context that is part of or derived from its containing document, whereas we explore ways to obtain supporting context in the form of additional (tweet) documents.

MOTIVATION

Analysis of human annotation task

Although there is a consensus that WSD is best suited for evaluation *in vivo* (i.e. as a component of another system), a reliable gold standard data set for *in vitro* evaluation is desirable, even if the output is not intended for a human end-user [26]. While annotation reliability depends in part on robust guidelines designed to maximize inter-annotator agreement (IAA), IAA tends to degrade as the sense repository becomes more fine-grained [26], as is the case in D2W. On one hand, if a D2W task is limited to named entities, and the set of mentions to be linked is given in advance, agreement can be rather high – e.g. 91.53%, 87.5%, and 92.98% was observed for Person, Geo-political, and Organization type entities in the TAC2010 data [16] – in spite of a sense repository which is *a priori* quite vast. In contrast, the task of linking whichever concept mentions appear important in a corpus of very small documents should prove difficult, as it is more demanding in spite of a dearth of contextual evidence. A D2W task may be characterized along two dimensions: whether concept mentions to be disambiguated are given in advance, and whether the target domain of concepts consists of all of Wikipedia or from a limited subset (e.g. only named entities). We refer to the task of linking whichever concept mentions appear important to a (largely) unrestricted domain of concepts (i.e. all Wikipedia pages) as *open-ended concept linking*.

Annotating every word without regard to its syntactic or semantic category, or its prominence in the discourse, is probably unnecessary for any application [26]. The criteria for determining which concept mentions to annotate must be specified in terms of (1) the properties of the target domain of concepts, (2) whether a concept exists in the target domain, and (3) the extent to which a mention is deemed ambiguous. A concept mention can be said to lack a (Wikipedia) concept referent in two distinct ways: it may be deemed *unlinkable* because the string in question, in the context in question, does not refer to a *valid concept* (i.e. one that could, in principle, appear in Wikipedia). On the other hand, the mention may refer to a valid concept, but there is not yet a corresponding Wikipedia page (see [19] for further discussion). Similarly, a concept mention can qualify as ambiguous in two ways: it may obviously refer to some valid concept, but even if each candidate has a corresponding Wikipedia page, the intended concept may be impossible to determine; on the other hand, the (Wikipedia-independent) concept being referred to may

[3] http://blog.twitter.com/2011/08/your-world-more-connected.html as of August 2011.

[4] http://www.mediabistro.com/alltwitter/twitter400million-tweets_b23744 as of August 2012.

[5] For an n-gram m, concept $t \in T$, $COMMONNESS(m,t) = \frac{c(m \to t)}{\sum_{t' \in T} c(m \to t')}$, where $c(m \to t)$ denotes the number of times m serves as a hyperlink to the concept t.

be clear, but there may be more than one (Wikipedia) concept that constitutes a correct answer in accordance with the annotation guidelines (e.g. concepts for which article mergers have been suggested might be considered equivalent, for annotation purposes; also c.f. "Gators" and "Pine nut" in section *Information potentially used by annotators* regarding taxonomic granularity). Concept mentions that unambiguously refer to a Wikipedia concept may still present difficulties. Specification of which concepts constitute valid targets must be done in terms of the property space of all concepts, which is arguably quite complex. In the case of D2W a concept's content derives not only from explicit (e.g. infobox, category, and link structure) but implicit (article text) facts, and may be difficult to separate from personal knowledge and experience with the (Wikipedia-independent) concept in question. Such a separation potentially limits annotation richness but may reduce inconsistency across annotators. Furthermore, determining which mentions to annotate depends not only on the properties of potential target concepts but on the prominence of the mention in question in the context in which it occurs. Perhaps a concept mentioned in passing, which does not pertain to the main point, should not be annotated. Finally, a concept might be relevant to an entire tweet though not denoted by any word or phrase therein. For example, *2011 Tohoku earthquake and tsunami* is clearly related to the tweet, "my thoughts and prayers go out to the Japanese people". We are aware of no annotation schemes that account for all of these variables, and leave a more precise formulation to future work.

Information potentially used by annotators

Annotators use information from different sources when annotating a concept mention. When short and informal texts such as tweets are analyzed in isolation, identifying the context necessary to disambiguate the concept mentions therein is non-trivial. Informative context for a given concept mention might be derived from the mention alone, within the tweet, or within the authors other tweets. Information about the author in general, his or her interests, recent events in the author's life, , and world knowledge may be informative as well. We inferred that annotators made use of several different sources of information, often simultaneously, and that world knowledge is supplemented by information acquired from Wikipedia during annotation. We aim to determine what sort of additional tweet context might have provided for an improved disambiguation context[6]. In what follows we give examples in which annotators either (1) appeared to use, or (2) failed to take advantage of, a given type of contextual support, along with analysis. Table 2 illustrates cases in which it appears that annotators *have* taken advantage of the type of supporting context in question.

First, "St. Patrick's Day" is unambiguous regardless of context. That "Hawks" refers to a sports team is implied by "Slump" and the pattern "Go ... !", but "Hawks" also may refer to the teams *Fukuoka Softbank Hawks* or *Chicago Blackhawks*, in addition to the correct referent *Atlanta Hawks*. However, only the Atlanta Hawks have players named Jeff (Teague) and Damien (Wilkins), and knowing this requires either being a member of a subculture that possesses enough knowledge to make this distinction, or having searched for this information, which can be done with a Wikipedia search and very few clicks. That "Gators" refers to a sports team is implied by "Go ... !". Whether the mention can be reliably linked to *Florida Gators men's basketball* may depend on mentions in other tweets written by the same author. In the first supporting tweet, "Sweet 16" refers to *NCAA Men's Division I Basketball Championship* as opposed to *Sweet Sixteen (birthday)*, as evidenced by the sports context; the situation is analogous for "March Madness" in the second supporting tweet. A candidate target like *Sweet Sixteen (KHSAA State Basketball Championship)*, a less prominent basketball tournament, is ruled out by the presence of "March Madness" and "Gators" (as both are associated with only the NCAA tournament). In addition, time of publication and author attributes provide ample evidence, independent of these supporting tweets: the tweet date was March 18th, during the NCAA Division I Men's Basketball Tournament, and the author played basketball at the University of Florida. Commonness alone would not suffice as "Gators" links most commonly to *Florida Gators*, the Wikipedia page about the University of Florida's athletics in general, which is not specific enough[7]. Some additional source of information is required to link to *Florida Gators men's basketball*. Table 3 illustrates annotation errors; presumably, annotators did not take advantage of the type of context in question.

"Detroit Tigers" is unambiguously associated with *Detroit Tigers*. The given annotation for "nuts" is *Nut (fruit)*, which is reasonable, but *Pine nut* is more appropriate as it is the nut ingredient used in pesto according to Wikipedia[8]. Ben Rhodes was the deputy National Security Advisor (NSA) to Barack Obama in March of 2011. This is not clear from the tweet text, but supporting tweets each provide evidence in favor of the target *Ben Rhodes (speechwriter)*. The American Political context indicates the target concept for "Clinton" is either *Bill Clinton* or *Hillary Rodham Clinton*. To inter that Hillary Clinton went on such a trip at the time of publication requires either American political knowledge or access to the URL in the tweet.

We observe that world knowledge, including what can quickly be obtained by looking through Wikipedia, helps annotation. Many such on-the-fly inferences would be difficult to make automatically, thus additional textual context is needed in order to generate a more comprehensive disambiguation context. We consider two methods for providing such content: (1) disambiguating mentions in the context of

[6]Note that in general by *disambiguation context* we mean all information that is applicable to the disambiguation task. Later in our description of GLOW **??** we take a narrower definition of this term.

[7]We base this judgement on the Gricean maxim of quantity: "Be as informative as required" (c.f. http://plato.stanford.edu/entries/implicature/). We leave an analysis in this vane to future work.

[8]Pesto may be made with other nuts, but according to the article *Pesto* this does not correspond with the classic recipe. The existence of multiple correct options for candidate targets at varying taxonomic levels makes evaluation more difficult because some arbitrary choices about what constitutes "close enough" or "specific enough" must be made.

Type	Tweet text	Mention
Mention Alone	Are you a college kid who likes drinking, dressing up, and making irish immigrants roll in their graves? Then St. Patrick's Day is for you!	St. Patrick's Day
Within Tweet	Slump is over! Way to ball out Jeff and Damian. Much needed win. Go Hawks!!	Hawks
Within Author's Tweets	Go Gators!!! A1: Sweet 16! What a good feeling. Keep it going... Go Gators!!!' A2: What's good everyone, catching up on these Tourney games and already see some upsets... March Madness! Go Gators!	Gators

Table 2. Context type used by annotators

Type	Tweet text	Mention
Mention Alone	So excited to announce I'll be singing "God Bless America" during the 7th Inning Stretch at the Detroit Tigers..	Detroit Tigers
Within Tweet	Making pesto! I had to soak my nuts for 3 hours	nuts
Within Author's Tweets	It was a pool report typo. Here is exact Rhodes quote: "this is not gonna be a couple of weeks. It will be a period of days." A1: At a WH briefing here in Santiago, NSA spox Rhodes came with a litany of pushback on idea WH didn't consult with Congress. A2: Rhodes singled out a Senate resolution that passed on March 1st which denounced Khaddafy's atrocities. WH says UN rez incorporates it	Rhodes
URL Content	Awesome post from wolfblitzercnn: Behind the scenes on Clinton's Mideast trip - URL - #cnn	Clinton

Table 3. Context type *not* used by annotators

all tweets in the dataset by the same author, and (2) disambiguating mentions in the context of all tweets in the same cluster (section *Tweet document creation*)[9].

SYSTEM

Global coherence

Some D2W systems aim to maximize the *global coherence* of their output, i.e., the concepts linked to in a given *source document*. Essentially, some measure of relatedness among these concepts informs the selection process for a given concept mention. A relatedness metric based on the Wikipedia link structure can leverage the co-occurrence of concept mentions in a document to the extent that the relationships expressed therein are captured in the links between their referent concepts. Concept mentions in microblog messages often lack explicit supporting context, therefore systems and annotators alike must look elsewhere for disambiguation context. We hypothesize that with the right additional context, given the resulting enriched disambiguation context, a D2W system that relies on optimizing its output for global coherence should perform better. In our experiments we do this in two ways: to a given tweet, we (1) append additional tweets by the same author, and (2) append tweets based on a clustering algorithm. We constrain the term *disambiguation context* in what follows to a set of concepts, each deemed a candidate referent of any concept mention in the source document. This definition is analogous to that used in previous sections; world knowledge, including that gained by reading tweets and examining Wikipedia, is represented approximately, via the extension of the disambiguation context that results from augmenting tweets with related tweets to create multi-tweet documents.

Enforcing constraints can be potentially harmful. The system of [24] performs poorly on the tweet dataset because it relies on unambiguous concept mentions for disambiguation, the guaranteed existence of which is implausible for the microblog genre [22]. TAGME [7] begins with commonness but enforces global coherence through a "voting" scheme in which the score associated with an n-gram m and a target concept t is derived from the vote of each other n-gram m' in the tweet. The vote of m' is the average of the relatedness scores [25] between each of its candidate concepts t' with t, weighted according to $COMMONNESS(m', t')$, and though links may be pruned, this system performs poorly on the tweet dataset as well [22]. GLOW [30], on the other hand, optimizes for global coherence using two supervised classifiers, and is conducive to a balanced disambiguation context, neither prohibitively small, nor large and noisy. Their notion of disambiguation context consists of the top candidates returned by a *local model* (described below) that for a given concept mention takes into account surrounding textual context while remaining agnostic to candidate concepts for surrounding mentions. A *global model* finalizes linking choices so as to optimize global coherence of the output. We chose to use GLOW because of its state-of-the-art performance on benchmark D2W datasets and its focus on a balanced disambiguation context.

Pipeline

The pipeline consists of three phases: first a *tweet document* is generated, then the document is fed to the D2W system, and finally results are extracted from the D2W system output.

[9]Other dimensions in terms of which tweets could be clustered to filter out noise include hashtags, timestamps and the mention/retweet structure for the tweet in question. Unfortunately Twitter API restrictions render these extensions slightly less accessible for older tweets.

Tweet document creation

The first phase consists of grouping individual tweets into documents. We create tweet documents for each experimental case, as described in Table 4.

Case	Tweet document content
By file	Each document consists of a single tweet
By author	Each document consists of all tweets by a given author
By cluster	Each document consists of all tweets in the same cluster

Table 4. Description of experimental cases

All tweets are pre-processed such that URLs are removed, and the @ and # characters are removed from user mentions and hashtags respectively. Tweets in documents are ordered chronologically by publication date, and those labeled ambiguous or non-referential are omitted.

A number of well-known probabilistic topic modeling approaches such as Probabilistic Latent Semantic Analysis (PLSA) [14] and Latent Dirichlet Allocation (LDA) [2], have been explored to discover topics from a set of documents. However, due to the shortness and lack of context, these topic modeling approaches may not work well with tweets. To overcome this difficulty, we explicitly smooth the topic distributions of tweets by building linkages between tweets, weighted by cosine similarity in terms of TF-IDF. A random walk-based approach is used to propagate the topic distribution probabilities across the linkages:

$$\begin{aligned} \hat{P}(z_k|x_i) &= \sum_{x_j \in X} w_{ji} P(z_k|x_j), \\ P(z_k|x_i) &= (1-\lambda) P(z_k|x_i) + \lambda \frac{\hat{P}(z_k|x_i)}{\sum_i \hat{P}(z_k|x_i)} \end{aligned} \quad (1)$$

where $P(z_k|x_i)$ is the probability of topic z_k for tweet x_i, w_{ij} is the similarity between x_i and x_j, and λ is a parameter that controls the balance between the previous topic distribution $P(z_k|x_i)$ and propagated topic distribution. We utilize PLSA to initialize the topic distributions. We cluster tweets using this PLSA+Random Walk-based Propagation (PRP) method by assigning a tweet x_i to the topic z_k that maximizes $P(z_k|x_i)$.

GLOW: a D2W system

In the second phase we use GLOW [30], a D2W system that disambiguates terms by attempting to optimize the global coherence of its output. Given a document d consisting of mentions $M = \{m_1, \ldots, m_N\}$, the system output consists of an N-tuple of target concepts, $\Gamma =< t_1, \ldots, t_N >$, a subset of all available concepts $T = \{t_1, \ldots, t_{|T|}\}$. Formally, one element of T is a null concept $t_\emptyset$, such that linking m to $t_\emptyset$ is akin to not linking m at all. *Local* feature functions ϕ assign $< m, t >$ pairs a high score to the extent that the context surrounding m is similar to t, and are meant to measure the likelihood that m links to t irrespective of the concepts referred to by m's surrounding mentions. *Global* feature functions ψ assign a high score to Γ to the extent that its contents are coherent. Coherence is calculated on a pairwise basis. Each global feature is either the Pointwise mutual information (PMI) or normalized Google distance (NGD) of a pair of concepts in the set, calculated in terms of the sets of concepts that either (1) link to each concept in the pair, (2) are linked to from each concept in the pair, or (3) are in the intersection of the sets defined in (1) and (2), for each concept in the pair[10]. Thus, GLOW attempts to solve the following optimization problem for a given document d:

$$\Gamma^* = \arg\max_{\Gamma} [\sum_{i=1}^{N} \phi(m_i, t_i) + \psi(\Gamma)] \quad (2)$$

Where Γ^* is the optimal output. This problem is NP hard, so inter-concept relatedness is calculated pairwise to reduce complexity, reformulating the problem as:

$$\Gamma^* \approx \arg\max_{\Gamma} \sum_{i-1}^{N} [\phi(m_i, t_i)] + \sum_{t_j \in \Gamma'} [\psi(t_i, t_j)] \quad (3)$$

The optimization is performed in two stages. First, in the *ranker* stage, Γ^* is found but without allowing any mention to be linked to $t_\emptyset$. Next, in the *linker* stage, whether each mention's top candidate should be replaced by $t_\emptyset$ is determined. In the system output, mentions linked to $t_\emptyset$ have a negative linker score while others have a positive linker score.

Extracting output

For a given case, each tweet document d is fed to the D2W system separately, the output of which consists of mentions that were linked (including those ultimately linked to $t_\emptyset$ and their associated target concepts). Each mention is associated with a *linker score* - the confidence associated with the choice to link that term - while each of its candidate target concepts is associated with a *ranker score* - the confidence associated with that particular concept. Thus for each linked mention m_{di} we have its *result tuple*, $R(m_{di})$ which consists of a linker score and a list of k targets, ordered according to their ranker score.

$$R(m_{di}) =< ls(m_{di}), (< t^1_{m_{di}}, rs(t^1_{m_{di}}) >, \ldots, < t^k_{m_{di}}, rs(t^k_{m_{di}}) >) > \quad (4)$$

We abbreviate the first and second elements of $R(m_{di})$ as $R(m_{di})_{ls}$ and $R(m_{di})_{rs}$. The output for each set of surface-identical mentions in d is then aggregated into one result tuple as follows. For a surface string s_d associated with one or more mentions in d, the set of associated result tuples is denoted R_{s_d}. Then $R(s_d)$, the result tuple for s_d, is defined by:

$$R(s_d) =< max_{R(m_{di}) \in R_{s_d}} R(m_{di})_{ls}, \bigcup_{R(m_{di}) \in R_{s_d}} R(m_{di})_{rs} > \quad (5)$$

In other words for any surface string, we consider all target concepts and associated ranker scores, and associate the string with the highest linker score of any matching mention.

Output aggregation is informed by two parameters: *longest-n-gram*, a binary parameter indicating whether or not the

[10] See [30] for a detailed explanation including the adaptations of PMI and NGD used.

"longest n-gram heuristic" is used (as opposed to "all terms"), and a *linker score threshold* λ. If the longest n-gram heuristic is used, then if both "Houston Rockets" and "Rockets" are disambiguated, for example, "Rockets" will be ignored. Finally, $R(s_d)$ will only be included in the final output if $R(s_d)_{ls} > \lambda$.

DATA AND SCORING METRIC

In this section we describe the dataset, provide a critical evaluation, and explain how system output is evaluated.

Construction, content, and annotation

We use the dataset described in [22], which we refer to as *gold1*. A random sample of verified twitter accounts were selected, and up to their 20 most recent tweets were extracted. The original dataset had 562 tweets, but due to tweets having been deleted, the dataset consists of 502 tweets from 28 authors. Annotators used an interface enabling them to read and annotate tweets, searching Wikipedia as needed, and were instructed to, where possible, indicate which concepts were "contained in, meant by, or relevant" to a particular tweet. Alternatively they were permitted to label tweets as ambiguous or as having referents outside of Wikipedia; 127 tweets were labeled as such and discarded[11]. The gold standard consists of the union of annotations from two annotators which amounts to 812 annotations (not including discarded tweets). URLs were removed entirely while mentions and hashtags were edited to remove leading @ and # characters respectively[12].

System false positives

Some system errors are the result of human annotation omissions [22] . There were 229 false positives when applying the GLOW to single tweets, using the longest n-gram heuristic, with the linker score threshold at -0.04. We looked at each one and rated it incorrect (110), partially correct (49), or correct (70). False positives deemed correct (FPDC) were labeled as follows: "@" (2), "#" (13), "lol" (5), "replace" (6), "new" (35), "equivalent" (9).

The *gold2* dataset is the result of adding all FPDC to *gold1*. For each FPDC type we provide representative system results followed by analysis. Table 5 gives some examples of each FPDC type, along with from the system output or *gold1*.

FPDC labeled *new* consist of a mention that annotators previously did not link and a target concept deemed correct. Table 5 gives three examples; "support", in this case, is an example of an analogous annotation in *gold1*. In the first a song was omitted in one tweet whereas in another a song was linked, and similarly so for the dates in the second example and its counterpart. In the third, a governmental acronym and an associated term are omitted, whereas in its counterpart they are annotated.

FPDC labeled *replace* consist of mentions that were originally annotated, but we believe the annotation provided by GLOW was significantly better. Table 5 contains three examples; "support", in this case, illustrates the change made by the system. In the first example some evidence was available in the tweet itself (though more conclusive evidence is available in the author's other tweets, as alluded to in section *Information potentially used by annotators*). In the second example note that *Grammy Nominees* is an album containing Grammy-nominated songs for a given year, but the URL in the tweet links to a page where only the album "Infinite Arms" can be purchased, revealing that the original annotation is incorrect (note that annotators did not have access to URLs in tweets). In the third the original annotation is too general. Note that the vast majority of false positives deemed *partially correct* are of this type.

FPDC labeled *eq* are instances where GLOW's target was deemed equivalent to the target in the original annotation. Table 5 lists three such examples followed by justification. FPDC labeled @ were user mentions that were not annotated, even though the user is identifiable and is prominent enough to have a Wikipedia page. FPDC labeled # were hash marked mentions that were not annotated. FPDC labeled *lol* were mentions expressing that the user laughed, e.g. "lol", "ROFL", "LMAO", etc. Annotating such mentions depends on whether we want to annotate actions the user indicates he or she performs in conjunction with the tweet.

Note that these omissions and errors drawn from a subset of those mentions whose annotation was corrected by GLOW; however, other errors and omissions exist (e.g. when both humans and GLOW made mistakes). The purpose of this analysis is not to discredit the dataset. Classification of annotations or omissions as erroneous is highly subjective in that it depends on both the user's interpretation of the annotation guidelines, which in this case were rather open-ended, along with their own world knowledge. We believe the formation of guidelines and annotation methods that are more robust to such discrepancies is an important avenue of research.

EXPERIMENTS

In this section we present and discuss experimental results. For each case we generate tweet documents (see section *Tweet document creation*), each of which is fed to the D2W system, and final output is extracted from system output (see section *Extracting output*). We calculate precision, recall, and MRR.

Evaluation metric

Output is evaluated against *gold1* and *gold2* (see section *Data and scoring metric*). Final output for a tweet document distinguishes identical mentions allowing each tweet to be associated with a list of targets. Precision (P), recall (R), and F-measure (F1) are calculated on a by-tweet basis as follows:

$$P = \frac{\sum_i^{N_S} |T(x_i) \cap G(x_i)|}{N_S} \quad (6)$$

[11]We acknowledge that ignoring non-referential tweets makes the task easier. Work that focuses on a system's ability to ignore irrelevant content is needed. Tweets were deemed ambiguous if annotators identified more than one correct answer, a case our system did not accommodate.

[12]@ and # characters were visible to human annotators, who were asked to ignore hash tagged terms unless their meaning is obvious; they were stripped during pre-processing. For further details and access to the dataset: http://ilps.science.uva.nl/resources/wsdm2012-adding-semantics-to-microblog-posts/ [22].

Type	False Positives Deemed Correct (FPDC)	Support from system output or *gold1*
New	So excited to announce I'll be singing "God Bless America" during the 7th Inning Stretch at the Detroit Tigers... URL	#NP "Crazy"- The Boys - *heeeeey*
New	Enter to win FREE tickets to my Houston show March 29th! URL	Ben has announced a benefit show in Charleston on December 10th for the family of Andy Kotowitcz. Details here: URL
New	DOE approves $102 million loan aid for Maine wind farm - URL	So nothing has changed from last night. Timetable for handover of no-fly zone enforcement is still "days" according to WH
Replace	Sweet 16! What a good feeling. Keep it going... Go Gators!!!	Sweet sixteen (birthday) → NCAA Men's Division I Basketball Championship
Replace	The deluxe version of the Grammy Nominated, Infinite Arms, is available for a special holiday price. Get it here URL	Grammy Nominees → Grammy Award
Replace	RT @user: @user I always spend my summer here! An old-growth forest within the Sipalay island in the Philippines!	Forest → Old-growth forest
Eq	Photos are great for engaging with your audiences. Upload images to Flickr.com and create slideshows with URL	Slideshow redirects to Slide show
Eq	Jalen said "How did Santa make my presents and it says Made in China?! Santa ain't Chinese!" lmao	People's Republic of China was merged with China
Eq	The Devil is a liar! Thank God for giving you to chance to see this beautiful morning. I'm thankful and very blessed.	Satan was deemed conceptually equivalent to Devil
Eq	Which childhood story would you miss the most? Peter Pan and Mary Plain for me. URL #IdMiss	Peter and Mary is the fairy tale whose main character is Peter Pan
#	#NATO to enforce arms embargo against #Libya - URL #Gaddafi	The situation in Libya is of great concern. NATO can act as an enabler and coordinator if and when member states will take action
@	RT @user: Tweets to 6.5 million followers in the name of #girlseducation: Thanks @Shakira, @user and @user! URL	Obama set to deliver a response on #Libya soon

Table 5. A mention is underlined to indicate it was annotated.

$$R = \frac{\sum_i^{N_G} |T(x_i) \cap G(x_i)|}{N_G} \quad (7)$$

$$F_1 = \frac{2PR}{P+R} \quad (8)$$

Where N_S is the number of $\langle m, t \rangle$ pairs in the system output, each x_i is a tweet, $T(x)$ contains the top target concept from each mention in tweet x, $G(x)$ contains each concept associated with x by an annotator, and N_G is the total number of gold standard annotations. Mean Reciprocal Rank (MRR) is calculated over all gold annotation tuples $\langle x, t \rangle \in G$ as follows:

$$MRR = \frac{1}{|G|} \sum_{i=1}^{|G|} \frac{1}{rank_i} \quad (9)$$

Where $rank_i$ is r if $< t_i^r, rs(t_i^r) >$ is in $R(s_d)_{rs}$, where t_i is the target of the ith gold annotation $< x_i, t_i >$, and d is the document that contains x_i. Otherwise, $1/rank_i = 0$.

Results

In order to investigate the most effective way to extend tweet context to improve D2W, we augmented single tweets using either the *by author* or *by cluster* methods (see Table 4)

For the case of single tweets, each tweet was input one at a time into GLOW. For cases where tweets were aggregated, a document containing the tweets, delimited by a line break and in chronological order by publication date, was input into GLOW.

Table 6 presents the results of applying these different methods to augment tweets. *By author* outperforms *by cluster*. Table 7 shows details for the top performing systems of each type. The systems that achieve the top Mean Reciprocal Rank (MRR), as well as MRR for the systems with the top F measure, are shown in Table 8.

The by file system performs the worst in each category. By author improves recall while by cluster improves precision. The Wilcoxon matched pairs signed rank text shows that improvement in f-measure from by file to by author method was significant ($p < .01$); improvement from by file to by cluster was significant as well ($p < .013$)[13]. The Adjusted Rand Index (ARI) is a measure of cluster similarity, corrected for chance. The ARI between the top author based and cluster based methods is low (.0128), indicating that there is very little overlap.

Detailed results for the highest performing systems are shown in Table 7. The differences in output moving from by file to by author systems consisted of 23 gains and 12 losses. Gains resulted for the following reasons: (i) because the top candidate was correct in both cases but in the by author case the linker score exceeded 0.0, but in the by file case it did not exceed -0.4; (ii) the top candidate was incorrect in the by file case but correct in the by author case; (iii) a surface-identical mention in another tweet either had a better linker score and/or it was linked to the correct target[14]. Some gains were deemed neutral (4) or bad (1), meaning that we deemed the change made incorrect, contrary to *gold1*. Examples of changes are illustrated in Table 9 and explained below. Losses were categorized in an analogous way.

The first change is due to additional supporting context in the author's other tweets, which include entities from modern politics (e.g. politician names and organizations). This additional context alleviates the noisy mention "Allies" which is strongly associated with *World War II* and hence *Empire of Japan*. In the second case the author had later mentioned "Whistler", a popular winter sports destination, near mentions of "slopes", "snowboarding", and "jet lag". In the third case, the author frequently mentions "St. Louis" in other tweets.

CONCLUSIONS AND FUTURE WORK

D2W systems that attempt to maximize the global coherence of output have been successful in formal genres, but the required supporting concept mentions are hidden in the Twitter domain. Our approach to this apparent data sparsity is orthogonal to that taken by [22], who designed features in terms of individual n-grams and candidate concepts, rarely dependent on the entire tweet (5 out of 33), never attempting to achieve global coherence. We showed that for a given tweet, adding tweets based on both authorship and topical similarity provided GLOW sufficient information to enhance the disambiguation context for concept mentions therein, yielding statistically significant gains over the *by file* base.

We have provided a qualitative analysis of an existing hand-labeled dataset, which raised questions about both definition and evaluation of the D2W task, elucidating various sources of difficulty. In future work we plan to generate comprehensive annotation and evaluation guidelines for D2W. Second, it is clear that sometimes there is more than one appropriate target concept for a given concept mention. In some cases two concepts are equally plausible targets (*Devil* vs. *Satan* for the n-gram "the devil"), while in other cases returning a concept slightly higher up in the *is-a* taxonomic structure would plausibly still be useful for downstream applications (e.g. returning *Florida Gators* instead of the more accurate *Florida Gators men's basketball*, given only "go Gators!!"). We plan to explore principled criteria for Wikipedia concept equivalence that go beyond the provided redirects, as well as evaluation methods that do not penalize such "not so bad" deviation from human annotation. Finally, we plan to evaluate the effects of expanding tweet context based on Twitter-centric features such as the mention/retweet structure and hashtags, as well as websites linked to from within tweets.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053, the U.S. NSF Grants IIS-0953149 and IIS-1144111

[13]We randomly split tweets into 17 groups, yielding 17 lists of annotations. We calculated F-measure for each group using both methods and the resulting F-measure pairs served as input to the test.

[14]Gains are determined with respect to the gold standard. The best by file and by author systems had linker score thresholds of -0.4 and 0.0, respectively.

Table of f-measures for different experimental parameters						
	By file		**By cluster**		**By author**	
	gold1	***gold2***	***gold1***	***gold2***	***gold1***	***gold2***
all terms	44.19%	51.00%	46.35%	51.82%	47.00%	52.56%
longest ngram	45.58%	52.79%	47.50%	53.13%	48.07%	53.92%
Link Threshold for F-Measures Shown Above						
	gold1	***gold2***	***gold1***	***gold2***	***gold1***	***gold2***
all terms	-0.2	-0.2	0	0.1	0.1	0
longest ngram	-0.4	-0.4	0	-0.2	0	-0.2
Cluster Size for F-Measures Shown Above						
	gold1	***gold2***	***gold1***	***gold2***	***gold1***	***gold2***
all terms	n/a	n/a	28	28	n/a	n/a
longest ngram	n/a	n/a	28	50	n/a	n/a

Table 6. Overview of different methods

Statistics for top performing systems of each type							
System	**Correct**	**Missed**	**False Positives**	**Total Output**	**Precision**	**Recall**	**F1**
by file	307	505	228	535	0.5738	0.3781	0.4558
by author	318	494	193	511	0.6223	0.3916	0.4807
by cluster	309	503	180	489	0.6319	0.3805	0.4750

Table 7. Detailed results by system type using the optimal parameters for each

		MRR1		**MRR2**	
		Best Params	**Best F Params**	**Best Params**	**Best F Params**
by File	All terms	44.20%	41.62%	43.77%	41.29%
	Longest ngram	40.75%	39.70%	40.50%	39.53%
by Author	All terms	45.82%	42.27%	45.44%	42.03%
	Longest ngram	42.23%	40.21%	42.06%	40.05%
by Cluster	All terms	44.89%	41.86%	44.42%	41.56%
	Longest ngram	41.52%	39.35%	41.32%	39.25%

Table 8. Best MRR & MRR for parameters yielding best F1

Tweet	**By file**	**By author**	**Type**
Japan is one of NATOs global partners. On behalf of our Allies I want to extend our heartfelt condolences to those who have lost loved ones	Empire of Japan	Japan	Good change
Ejoying myself in Whistler!	Whistler, British Columbia	Whistler, British Columbia	Greater LS for identical mention
RT @kmoxnews: Section of I-55 Closed Until Monday: I-55 will be closed in both directions between Carondelet and the 4500 block of...	Carondelet, St. Louis	Carondelet, St. Louis	Context
Obama says he doesn't expect harmful levels of radiation to hit the U.S. ... public health experts say no precautionary measures needed	Ionizing radiation	Radiation	Neutral Change
Making pesto! I had to soak my nuts for 3 hours!	Pine nut	Nut (fruit)	Bad change

Table 9. Gains from by file to by author system

and the U.S. DARPA BOLT program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

1. Amigó, E., Artiles, J., Gonzalo, J., Spina, D., Liu, B., and Corujo, A. Weps-3 evaluation campaign: Overview of the online reputation management task. In *CLEF 2010 (Notebook Papers/LABs/Workshops)* (2010).
2. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3* (Mar. 2003).
3. Chen, Z., and Ji, H. Collaborative ranking: A case study on entity linking. In *Proc. EMNLP2011* (2011).
4. Cucerzan, S. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL 2007* (2007).
5. Cucerzan, S. Tac entity linking by performing full-document entity extraction and disambiguation. In *Proc. TAC 2011 Workshop* (2011).
6. Fernandez, N., Fisteus, J. A., Sanchez, L., and Martin, E. Webtlab: A cooccurence-based approach to kbp 2010 entity-linking task. In *Proc. TAC 2010 Workshop* (2010).
7. Ferragina, P., and Scaiella, U. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, ACM (2010), 1625–1628.
8. Genc, Y., Sakamoto, Y., and Nickerson, J. V. Discovering context: classifying tweets through a semantic transform based on wikipedia. In *Proceedings of the 6th international conference on Foundations of augmented cognition: directing the future of adaptive systems*, FAC'11 (2011), 484–492.
9. Guo, Y., Che, W., Liu, T., and Li, S. A graph-based method for entity linking. In *Proc. IJCNLP2011* (2011).
10. Han, X., and Sun, L. A generative entity-mention model for linking entities with knowledge base. In *Proc. ACL2011* (2011).
11. Han, X., Sun, L., and Zhao, J. Collective entity linking in web text: A graph-based method. In *Proc. SIGIR2011* (2011).
12. Han, X., and Zhao, J. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM 2009 (2009).
13. He, J., de Rijke, M., Sevenster, M., van Ommering, R., and Qian, Y. Generating links to background knowledge: A case study using narrative radiology reports. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, ACM (2011), 1867–1876.
14. Hofmann, T. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99 (1999).
15. Ji, H., Grishman, R., and Dang, H. Overview of the tac 2011 knowledge base population track. In *Text Analysis Conference (TAC) 2011* (2011).
16. Ji, H., Grishman, R., Dang, H., Griffitt, K., and Ellis, J. Overview of the tac 2010 knowledge base population track. In *Text Analysis Conference (TAC) 2010* (2010).
17. Kozareva, Z., Voevodski, K., and Teng, S. Class label enhancement via related instances. In *Proc. EMNLP2011* (2011).
18. Kulkarni, S., Singh, A., Ramakrishnan, G., and Chakrabarti, S. Collective annotation of wikipedia entities in web text. In *KDD* (2009), 457–466.
19. Lin, T., Mausam, and Etzioni, O. No noun phrase left behind: Detecting and typing unlinkable entities. In *EMNLP-CoNLL* (2012), 893–903.
20. McNamee, P., and Dang, H. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC) 2009* (2009).
21. McNamee, P., Mayfield, J., Lawrie, D., Oard, D. W., and Doermann, D. Cross-language entity linking. In *Proc. IJCNLP2011* (2011).
22. Meij, E., Weerkamp, W., and de Rijke, M. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, ACM (New York, NY, USA, 2012).
23. Mihalcea, R., and Csomai, A. Wikify!: linking documents to encyclopedic knowledge. In *CIKM*, vol. 7 (2007), 233–242.
24. Milne, D., and Witten, I. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, ACM (2008), 509–518.
25. Milne, D., and Witten, I. Learning to link with wikipedia. In *An effective, low-cost measure of semantic relatedness obtained from wikipedia links*, the Wikipedia and AI Workshop of AAAI (2008).
26. Navigli, R. Word Sense Disambiguation: a survey. *ACM Computing Surveys 41*, 2 (2009), 1–69.
27. Pennacchiotti, M., and Pantel, P. Entity extraction via ensemble semantics. In *Proc. EMNLP2009* (2009).
28. Radford, W., Hachey, B., Nothman, J., Honnibal, M., and Curran, J. R. Cmcrc at tac10: Document-level entity linking with graph-based re-ranking. In *Proc. TAC 2010 Workshop* (2010).
29. Ratinov, L., and Roth, D. Learning-based multi-sieve co-reference resolution with knowledge. In *Proc. EMNLP* (2012).

30. Ratinov, L., Roth, D., Downey, D., and Anderson, M. Local and global algorithms for disambiguation to wikipedia. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)* (2011).

31. Vitale, D., Ferragina, P., and Scaiella, U. Classification of short texts by deploying topical annotations. In *ECIR* (2012), 376–387.

32. Zhang, W., Su, J., and Tan, C. L. A wikipedia-lda model for entity linking with batch size changing. In *Proc. IJCNLP2011* (2011).

33. Zhou, Y., Nie, L., Rouhani-Kalleh, O., Vasile, F., and Gaffney, S. Resolving surface forms to wikipedia topics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10 (2010), 1335–1343.