Multi-source named entity typing for social media

Reuth Vexler
Information Systems Dep.
University of Haifa
Haifa, Israel, 31905
reuthvex@gmail.com

Einat Minkov
Information Systems Dep.
University of Haifa
Haifa, Israel, 31905
einatm@is.haifa.ac.il

Abstract

Typed lexicons that encode knowledge about the semantic types of an entity name, e.g., that 'Paris' denotes a *geolocation*, *product*, or *person*, have proven useful for many text processing tasks. While lexicons may be derived from large-scale knowledge bases (KBs), KBs are inherently imperfect, in particular they lack coverage with respect to long tail entity names. We infer the types of a given entity name using multi-source learning, considering information obtained by alignment to the Freebase knowledge base, Web-scale distributional patterns, and global semi-structured contexts retrieved by means of Web search. Evaluation in the challenging domain of social media shows that multi-source learning improves performance compared with rule-based KB lookups, boosting typing results for some semantic categories.

1 Introduction

Typed lexicons associate lexical entity names with a set of semantic types, e.g., the name 'Paris' may be mapped into the semantic categories of *geolocation*, *product*, or *person*. Such world knowledge has proven valuable for various text processing tasks, including the classification of search queries (Pasca, 2004), question answering (Bordes et al., 2014; Yao and Durme, 2014), named entity recognition (Ling and Weld, 2012; Yamada et al., 2015), entity linking (Fang and Chang, 2014; Hoffart et al., 2014), and relation extraction (Ling and Weld, 2012; Chang et al., 2014).

Over the last years, large scale knowledge bases (KBs) have become available, like Freebase (Bollacker et al., 2008), DBPedia (Auer et al., 2007) and YAGO (Suchanek et al., 2007), which organize entities and their lexical aliases into semantic categories (or, *types*). KBs are inherently incomplete however in their coverage of world knowledge (West et al., 2014). This is partly because updates to the database with emerging entities and facts exhibit some delay (Wang et al., 2012; Hoffart et al., 2014). In addition, the 'long tail' of unpopular, local or domain-specific concepts is under-represented in general knowledge bases (Lin et al., 2012; Ling and Weld, 2012).

In this work, we seek to associate semantic types of interest with a given entity name. We refer to this problem as *named entity typing*. Our main focus is on names mentioned on social media. In this domain, local contextual information is often very limited, so that background world knowledge is especially useful for semantic processing purposes. For example, the commercial system described by Gattani *et al.* (2013) performs named entity recognition in tweets by matching the tweet text against known entity names in a reference knowledge base (KB). But, emerging and long tail entity names, which are frequently mentioned on social media, e.g., the names of local businesses, as well as informal names, abbreviations, and acronyms, are often missing from KBs (Ritter et al., 2011).

Previously, researchers have proposed to infer the semantic types of entity names based on local distributional evidence harvested from Web documents (Lin et al., 2012; Yaghoobzadeh and Schütze, 2015), however such distributional evidence is sparse and costly to obtain for long tail names. We claim that in order to infer the semantic types of infrequent entity names, it is neces-

Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, pages 11–20,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

sary to model global, semi-structured, contextual evidence of the respective name mentions. Specifically, we consider in this work paragraph-level contexts, Webpage titles and URLs, obtained by means of Web search. As suggested by Rüd *et al.* (2011), Web search has several advantages as a source of background knowledge in cross-domain settings. We believe this work to be the first to examine Web search as an information source for the task of entity name typing, and the first to evaluate this task in the social media domain.

We apply a discriminative learning framework using an ensemble of information sources, naming this approach as Multi-Source Named Entity Typing (MS-NET). Relevant evidence about a given entity name is extracted from the various sources and represented as features in a joint feature space. Specifically, we model as features in this work related Freebase category tags, distributional information in the form of ReVerb lexico-syntactic patterns (Fader et al., 2011), and semi-structured Web search results.

In our experiments, we consider a set of nine target semantic entity types found to be highly popular on social media (Ritter et al., 2011). While training examples are obtained by means of distant supervision, our evaluation of MS-NET is conducted using a gold-labeled dataset of entity names extracted from tweets, in which each name string has been annotated with its full set of semantic types. We make this evaluation dataset available to the research community.

To the best of our knowledge, this is the first work to combine diverse information sources, including Web search results, for the purpose of entity name typing. Our results clearly show that the information sources modeled are complimentary– learning using all sources yields the best overall typing performance with respect to both recall and precision. MS-NET improves overall performance compared with a plausible baseline of rule-based alignment to a KB. In particular, a boost in performance is obtained for some of the semantic types evaluated, such as *facility* and *product*, which apply to a long tail of infrequent entity names.

2 Related Work

2.1 Named entity typing

Several recent works have considered a similar problem setting to ours. These works mainly rely on Web-scale lexico-syntactic distributional profiles of the target name mentions to induce its types.

Lin *et al.* (2012) represent KB-indexed as well as out-of-KB entity names in terms of *subject–relation–object* triplets with which they cooccur in Web documents, extracted using the ReVerb algorithm (Fader et al., 2011). Freebase semantic types are then propagated from indexed to unknown names over common ReVerb triplets. We represent ReVerb distributional patterns as features in our framework. Similarly to previous researchers (Yaghoobzadeh and Schütze, 2015), we find ReVerb triplets to perform poorly when used as a standalone information source. We discuss the reasons for this in detail below.

Yaghoobzadeh and Schütze (2015) presented an embedding based approach for the typing of named entities. They compute embedding vectors for words, entities and semantic types using an annotated version of the ClueWeb Web corpus, in which entity mentions have been tagged with their Freebase IDs. They report poor typing performance for infrequent entity names, possibly because the embedding approach is sensitive to context sparsity (Cui et al., 2015).

In this work, we model distributional lexico-syntactic information within a multi-source learning framework. This distributional information is integrated with knowledge extracted from structured datasets and with global semi-structured contextual signals, obtained via web search. We believe this work to be the first to evaluate named entity typing on twitter messages, which contain relatively many unknown and irregular entity name forms. Our contribution is orthogonal to the works mentioned above, as various types of distributional features can be easily integrated using multi-source learning.

2.2 Semantic type prediction using Web Search

Web search results have been used as an information source for several related semantic tasks. It is a common strategy to enrich Web queries with search results in order to classify them into semantic categories (Shen et al., 2006; Ullegaddi and Varma, 2011). Recently, Web searches have been proven useful for the disambiguation and linking of entities mentioned in search queries to Wikipedia (Carmel et al., 2014; Cornolti et al., 2014). Web queries and micro-blogs are similar to Web queries in that they are short, ambiguous and noisy.

Rüd *et al.* (2011) claim to be first to consider the idea of using search engines as an information source for an NLP problem. They outline a variety of features extracted from search engine results, designed to achieve cross-domain named entity recognition robustness. While their work is evaluated on the classification of Web queries, they conjecture that their approach is similarly applicable to related domains like Twitter and SMS.

This work continues these previous efforts, applying Web search features to the task of named entity typing, and evaluating this task in the social media domain. We find that modeling of global semi-structured contexts obtained via Web search is necessary in order to maintain high cross-domain performance. Unlike previous works, we put emphasis on investigating the synergy between Web search results and other sources of world knowledge, including structured KBs. A detailed evaluation reveals that the best configuration of information sources depends on the characteristics of the target semantic class.

3 Multi-source named entity typing

Let us first formally define the named entity typing task. Given a named entity string e and a set of target semantic types of interest L, our goal is to output a membership function $m : e \times L \mapsto \{0, 1\}$, such that $m(e, \ell)$=1 if e denotes some entity of type $\ell \in L$.[1] The membership relation is not functional, and it is possible that e be associated with any subset of the target types $S \subseteq L$, including the empty set.

We address this multi-class multi-label classification task by learning binary classifiers per label, $C_1, ..., C_{|L|}$, having each classifier C_j predict a binary membership, $m(e, \ell_j) \in (0, 1)$. A main advantage of the binary prediction setting is that it is intuitive, simple, and scalable (Read et al., 2009).

3.1 Knowledge sources

Each name e is represented as a multinomial vector of features, extracted from multiple sources as described below. We illustrate the proposed feature scheme using the running example e=‘eagles’, which maps to the types *sportsteam* and *music artist* according to our annotation.

```
Data: Entity name $e$, knowledge base $k$
Result: $S(e)$, a set of semantic types associated with $e$ according to $k$
initialization;
1. Extract $A(e)$: the set of entities in $k$, for which $e$ is a possible alias;
2. Initialize $S(e)$;
3. for $a \in A(e)$ do
    Retrieve S(a): the semantic types of entity $a$ according to $k$;
    $S(e) = S(e) \cup S(a)$;
end
```

Algorithm 1: Extraction of semantic types associated with e from a KB

KB tags. Knowledge bases model distinct entities and the relations between them. Typically, entities are associated with a hierarchical structure of semantic types, as well as with lexical aliases (Dalvi et al., 2015). In this work, we extract and represent as features semantic categories associated with the name entity string e in Freebase (FB).

We apply the procedure outlined in Alg. 1 to extract $S^F(e)$, the set of Freebase semantic types associated with the name string e. As described, step (1) of the algorithm involves alignment of the target entity name with possibly matching entities in the KB. In order to increase recall, one can apply proximate string matching in performing this alignment. In this work, we use Freebase API for this purpose, enabling inexact search against known entity aliases.[2] Once relevant entities are identified, their known types are unified to create the set of associated FB types.

47 semantic tags in total were extracted in this fashion for the string ‘eagles’. The extracted tags are of high-quality in part, e.g., *sports/professional sports team*, but include also noisy tags, e.g., *base/websites/website*. Noisy tags can be explained by annotation inconsistencies that exist in Freebase, and also result from the proximate alignment of ‘eagles’ to FB–while increasing recall, proximate matching involves some decrease in precision. In general, databases may include noise also as the result of applying imperfect automatic knowledge population methods.

The set of the KB extracted types has to be mapped to the target type schema L. It is common practice to make use of manually constructed mapping rules to align KB categories against a target semantic schema (Ling and Weld, 2012). We rather apply learning to perform this alignment, having the tags $S^F(e)$ represented as binary fea-

[1] Named entity detection is out of scope of this work. Entity names of interest may be identified automatically (Lin et al., 2012; Hoffart et al., 2014) or manually.

[2] We use the filter: ”(all name:entity alias:entity)”, considering the top three results in the returned ranked hit list. These choices were tuned in preliminary experiments using training examples.

Freebase cateogry tags
fb.base/popstra/topic, fb.base/popstra/sww_base
fb.base/losangelesbands/topic, base/ontologies/ontology_instance
sports/professional_sports_team, base/usnris/nris_listing
base/ovguide/country_musical_groups, base/schemastaging/topic
...
ReVerb patterns
reverbRel.be also a master of
reverbRel.win in
...
Web search features
url.philadelphiaeagles
title.philadelphia.-1, title.official, title.site, title.philadelphia
snippet.philadelphia.-1, snippet.official.1, snippet.team.2
snippet.site.3, snippet.official, snippet.team, snippet.site
snippet.philadelphia, snippet.roster, snippet.new, snippet.history
snippet.youth, snippet.program, snippet.ticket and *snippet.inform*

Table 1: An illustration of the feature types derived from the various sources for $e =$‘eagles’.

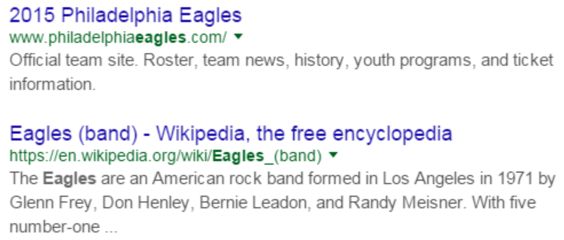
tures, as illustrated in Table 1. Ideally, learning should discard or downweight the features that correspond to irrelevant KB categories.

As discussed above, a main weakness of general knowledge bases as information source is that even large-scale KBs such as Freebase are inherently incomplete. Long tail and emerging entities, as well as informal name forms that are creatively and dynamically used in social media, are likely to be missing or from the KB. In such cases, relevant evidence must be obtained from other sources.

Lexico-syntactic distributional patterns. Named entities and their types may be extracted based on distinctive lexical patterns that surround the target entity mentions (Banko et al., 2007; Carlson et al., 2010).

Distributional semantics may be represented in terms of key phrases (Hoffart et al., 2014), distributional clusters (Roth, 2009), topics (Ritter et al., 2011) or word embeddings (Rong et al., 2016). Similarly to previous works (Lin et al., 2012; Yaghoobzadeh and Schütze, 2015), we consider here lexico-syntactic contexts in the form of <*subject*,*relation*,*object*> triplets extracted from Web documents using the ReVerb algorithm (Fader et al., 2011).

We utilize a collection of 15 million ReVerb assertions extracted with high confidence from the ClueWeb09 corpus[3]. Following Lin *et al.* (2012), for each entity name e, we consider the set of assertions in which e appears as subject. The string ‘eagles’, for example, appears as the subject argument in 140 different patterns in this corpus, including the (lemmatized) triplets <*eagles*, be also a master of, craftsmanship>, and <*eagles*, win in a war of, attrition>. As shown in Table 1, we represent the *relation* argument cooccuring with e as a full string. In preliminary experiments, we also considered a bag-of-word representation of the subject and object arguments, but found those representations to give inferior results.



2015 Philadelphia Eagles
www.philadelphiaeagles.com/ ▾
Official team site. Roster, team news, history, youth programs, and ticket information.

Eagles (band) - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Eagles_(band) ▾
The Eagles are an American rock band formed in Los Angeles in 1971 by Glenn Frey, Don Henley, Bernie Leadon, and Randy Meisner. With five number-one ...

Figure 1: Top two search results for the query ‘eagles’

A manual inspection of the extracted Reverb patterns reveals them to be quite noisy. One reason is ambiguity of the subject argument, which may denote a general noun phrase as well as an entity name. For example, triplets that include *eagles* may refer to mentions of the bird species that the sportsteam is inspired from. As name mentions in social media are often short and ambiguous, they tend to match general noun phrases. Another downside of using distributional statistics as information source is that limited data exists for long tail entity names. Given only a few Web mentions, it is often the case that a long tail name be matched with no Reverb pattern. Finally, as indicated elsewhere (Yaghoobzadeh and Schütze, 2015), these well-formed patterns may not maintain their effectiveness across genres.

Web search results. We wish to further model global Web contexts of the target name e, including the full paragraphs in which it appears, and semi-structured evidence in the form of the Webpage title and source. We obtain this information by means of Web search.

There are several advantages of modeling Web search results that make it appropriate for the processing of social media. Mainly, search engines provide access to nearly real-time raw Web data. They therefore provide good coverage of emerging entities, as well as long tail and noisy name forms that appear frequently on social media.

Having submitted the named entity string as a query to a search engine, we consider the top k retrieved results. Figure 1 shows the top two results for the query ‘eagles’. We derive several types of features from these search results.

[3] http://reverb.cs.washington.edu/

Given the paragraphs in which e is mentioned ('snippets'), we encode local context information using positional features, denoting the distance and direction of neighboring words within a window of 3 tokens to the right and left of the name mention. These local features are highly specific, and are therefore expected to be sparse. Useful context information can be further derived at paragraph-level (Pritsker et al., 2015)–we therefore additionally represent the whole snippet using count-weighted unigram word features.

We further model the respective Webpage title and source URL, both providing useful semi-structured evidence. Webpage titles are often thematic, and are sometimes generated using wrappers indicative of a source table or list. Likewise, the source URL may map to a domain-specific resource, e.g., *imdb.com* is a Web database that indexes entities in the entertainment domain. We represent titles using positional and bag-of-word features similarly to the snippet. The URL string is split by period and backslash symbols, having each token map to a feature. The full list of generated features per the top search result for 'eagles' are detailed in Table 1.

3.2 Distant learning

As described above, we learn $|L|$ binary classifiers $C_1, ..., C_{|L|}$, having each classifier C_j predict a binary membership $m(x, \ell_j) \in (0, 1)$.

Given labeled examples, a variety of learning methods can be applied to learn C_j. As it is costly to annotate a large set of examples with respect to each category, we adopt a distantly supervised approach. For each target semantic type $\ell \in L$, we obtain relevant entity names from Freebase. Following previous researchers (Ritter et al., 2011; Ling and Weld, 2012), we manually identified Freebase tags that correspond to each category ℓ_j. Entity names were then sampled from the selected Freebase lists, forming a set of positively labeled examples for classifier c_j. In our experiments, we consider a random sample of examples positively associated with the remaining classes, $\ell_i \neq \ell_j$ as negative examples for c_j.

There are several limitations of distant supervision. We expect the distribution of entity names indexed in FB to be different from entity names that are mentioned on social media. In particular, long tail entities are under-represented in Freebase. In addition, the labels obtained using distant supervision may be noisy. (Mainly, negative labels may be false due to FB incompleteness.) Nevertheless, this approach enables one to collect a large number of auto-labeled examples with minimal effort, and is therefore highly adaptive.

4 Experimental setup

For each target type $\ell \in L$, we sampled 900 entity names from relevant Freebase categories. In order to assess and correct possible representation bias, we further sampled 100 additional entity names per category from manually identified Web lists. We found that 18% of the latter names were missing from Freebase. Overall, the constructed training dataset includes about 1,000 unique names per target category. In learning a binary classifier c_j, we uniformly sample an equal number of entity names (1,000) associated with the other classes as negative examples.

We used Google API to perform Web-scale search.[4] Following tuning experiments, in which we examined cross-validation results using the training data, we set the number of top search results modeled per entity name to $k = 15$. We found performance to be relatively insensitive to value of k, as long as at least 5 search results were modeled.

Arguably, search engines inflect bias over the produced rankings. We experimented with shuffling of the top 50 search results (prior to selecting the top 15 results) and found performance to be robust with respect to ranking variance. Furthermore, in our experiments, we ignore exact ranking information, assigning equal importance to each of the selected search results.

We report the results of a strict yet realistic learning setting, in which the classifiers trained using the examples obtained from Freebase and specialized Web lists were applied across domains to a test dataset, which includes entity names found on twitter. We experimented with several classification algorithms using Weka (Hall et al., 2009), and report our results using SVM, which was found to perform best. In the following section we first describe the test dataset, and then turn to discuss the experimental results.

5 Gold labeled evaluation dataset

For evaluation purposes, we manually constructed a gold-labeled dataset of entity names mentioned

[4] https://developers.google.com/custom-search

	N	Agreement	Out-of-FB: ratio / examples	
Music artist (MA)	193	0.69	6.2%	*suenalo, shakemode, testing tomorrow*
Company (CO)	155	0.81	6.5%	*indigenous, evergreen Subaru, nex-tech*
Facility (F)	123	0.73	23.6%	*the tall ship silva, belles mansion, knighttime billiards*
Geo location (GL)	245	0.83	2.9%	*robinhoods bay, long island mac Arthur airport, Cromwell field*
Movie (M)	121	0.82	0%	
Product (PR)	141	0.75	11.3%	*Avast AntiVirus 4 8, Air Music Jump, vanilla vodka, Bugatti V*
Sports team (ST)	38	1.00	13.2%	*Marlboro Ducati, Ryerson Quidditch Team, AIS U21, WB Wildcats*
TV show (TV)	70	0.90	2.9%	
Person (P)	356	0.86	8.4%	*denise calaman, Eduardo surita*

Table 2: Statistics of the gold-labeled twitter dataset detailed by semantic type: no. of labels, inter-annotator agreement, and ratio and examples of entity names for which no match in FB was found.

	P	R	F1	cov.	acc.
FB mapping	*.54*	*.60*	*.57*	*.80*	*.31*
Single-source:					
FB	.53	**.65**	**.58**	.88	.24
WS: snippet	.53	.60	.56	1.0	.22
WS: title	.46	.59	.52	1.0	.15
WS: URL	.45	**.63**	.53	1.0	.13
WS (all)	**.55**	**.63**	**.59**	1.0	.24
Multi-source:					
WS + FB	**.63**	**.63**	**.63**	1.0	**.32**
WS + RV	**.57**	**.61**	**.59**	1.0	.27
WS + FB + RV	**.65**	**.65**	**.65**	1.0	**.35**

Table 3: Instance-level performance using various information sources and features. Results that improve over the baseline are bold faced.

in tweets. We considered a corpus of 2,400 tweets collected by Ritter *et al.* (2011). They have identified the name mentions in this corpus, and annotated these mentions with their contextualized meaning with respect to the following set of types: *music artist*, *company*, *facility*, *geolocation*, *movie*, *product*, *sportsteam*, *tv-show* and *person*. These types were found to be most popular in the given tweets. A similar set of categories has been used in other works on social media (Gattani et al., 2013; Derczynski et al., 2015).

In order to use this resource for the evaluation of multi-label named entity typing, we have annotated the entity names in the corpus with their full set of types, using the same target category set. Labeling was performed by a graduate student, who was allowed to use any resource available, including Web searches, to determine whether an entity belonged to each of the target classes.

A random subset of 100 entities has been co-annotated by the first author in order to assess inter-agreement rates. Cohen's kappa agreement scores with respect to each of the target classes are detailed in Table 2. As shown, agreement ranges between 0.69–1, denoting substantial to perfect agreement (Landis and Koch, 1977). Interestingly, the lowest agreement was observed for the *music artist* category. We found that disagreements mainly occurred for short, and therefore highly ambiguous, entity names; e.g., 'Justin' may or may not refer to the music artist "justin bieber", and 'MAC' is possibly an alias for "The Mac Band". Somewhat more expectedly, moderate agreement was observed for the *product* and *facility* categories. It is not clear, for example, if 'Twitter' and 'Facebook' are services or products.

The resulting dataset includes 965 distinct entity names and 1,442 label assignments overall. About 27% of the entity names were assigned two classes, and 11% of the entity names were labeled with three types or more. The corpus includes misspelled, abbreviated, and 'long tail' entity names, such as local restaurants and hotels. Table 2 details the total number, as well as concrete examples, of names per category for which no matching entry was found in Freebase ('Out-of-FB'), despite using a proximate alignment procedure.

This dataset may be first to include gold mappings of entity names into a set of types of interest in the social media domain. We will make the dataset freely available to the research community.

6 Results

Table 3 details our test results on the gold-labeled twitter dataset. We report the following evaluation measures: (i) example-level macro average precision, recall, and F1: these measures are first computed with respect to the set of types assigned to each entity name, and are then averaged over all entity names in the dataset; (ii) coverage: the ratio of entity names for which type predictions were generated. (iii) accuracy: the ratio of names perfectly classified, i.e., having all their types correctly predicted, with no incorrect types assigned to them. The table reports multiple feature configurations, varying the information sources and fea-

		FB Mapping			MS-NET			
Target type	N	P	R	F1	P	R	F1	ΔF1
Music-artist	193	.48	.63	.54	.43	**.69**	.53	*-1.9%*
Company	155	.54	.43	.48	**.61**	**.46**	**.52**	**8.3%**
Facility	123	.46	.33	.38	**.61**	**.55**	**.58**	**52.6%**
Geo-location	245	.69	.72	.70	**.83**	.67	**.74**	**5.7%**
Movie	121	.64	.42	.51	.60	.41	.49	*-3.9%*
Product	141	.49	.40	.44	**.53**	**.50**	**.51**	**15.9%**
Sportsteam	38	.28	.63	.39	**.38**	**.79**	**.51**	**30.8%**
Tv-show	70	.59	.50	.54	.43	**.53**	.47	*-13.0%*
Person	356	.87	.78	.82	.82	.67	.74	*-9.8%*
Macro average		.56	.54	.55	**.58**	**.59**	**.58**	**5.5%**

Table 4: Classifier performance by type. Results that improve over the baseline are bold faced.

Music-artist	**Company**	**Facility**
FreebaseType./music/artist	FreebaseType./business/business_operation	snippet.review.0
url.artist	FreebaseType./organization/organization	snippet.locat.0
title.music.0	snippet.compani.0	FreebaseType./architecture/structure
snippet.music.0	url.compani	url.tripadvisor
url.music	FreebaseType./business/employer	FreebaseType./projects/project_focus
snippet.album.0	title.compani.0	FreebaseType./location/location
snippet.song.0	FreebaseType./business/consumer_company	snippet.rate.0
url.fm	url.stock	FreebaseType./architecture/building
title.listen.0	snippet.min.0	url.yelp
title.concert.0	url.invest	url.biz
Geo-location	**Movie**	**Product**
FreebaseType./location/location	FreebaseType./film/film	FreebaseType./business/consumer_product
FreebaseType./location/statistical_region	snippet.direct.0	url.product
FreebaseType./location/dated_location	snippet.film.0	url.facebook
FreebaseType./location/citytown	url.titl	snippet.camera.0
snippet.weather.0	snippet.movi.0	title.review.0
title.weather.0	url.imdb	title.facebook.0
snippet.forecast.0	title.imdb.0	snippet.profil.0
snippet.locat.0	title.imdb.2	title.review.1
title.forecast.0	url.movi	snippet.like.0
url.weather	title.movi.0	title.spec.0
Sport-team	**Tv-show**	**Person**
FreebaseType./sports/sports_team	FreebaseType./tv/tv_program	FreebaseType./people/person
url.team	title.tv.0	snippet.born.0
snippet.team.0	snippet.episod.0	FreebaseType./people/deceased_person
snippet.club.0	title.tv.1	snippet.born.1
snippet.footbal.0	url.tv	snippet.born.2
FreebaseType./soccer/football_team	title.seri.0	snippet.profession.-2
snippet.leagu.0	snippet.seri.0	snippet.profil.0
title.footbal.0	title.seri.2	snippet.review.0
snippet.fixtur.0	snippet.tv.0	snippet.linkedin.0
snippet.statist.0	snippet.episod.1	url.peopl

Table 5: Top weighted features per category

ture types modeled.

Mapping baseline. As baseline, we consider the plausible strategy of rule-based mapping, aligning $S^F(e)$, the set of Freebase tags extracted for name e (Alg. 1), with the target labels L. We manually determined a set of alignment rules for the purposes of this study, utilizing and expanding rules made previously available by other researchers (Ling and Weld, 2012). As shown, manual mapping applies to 80% of the entity names in the twitter dataset, for which at least one possibly matching entry in Freebase was found. The resulting instance-level precision and recall are .54 and .60. While manual rules tend to be precise, a main source of noise lies in the proximate matching of string e against FB entities. And, hand-picking FB categories for alignment hurts recall.

Single-source learning results. We first discuss our experimental results using each information source separately. Encoding the set of Freebase categories associated with e as features in the learning framework ('FB') substantially improves over the manual alignment rules in terms of coverage (.88 vs. .80) and recall (.65 vs. .60). While both methods use the same resource, the mapping rules rely on fewer high-quality FB categories.

Learning using the features derived from Web search results ('WS') gives perfect coverage. Performance using either snippet, title or URL information is lower compared with the KB-lookup baseline, however when all of the Web search features are modeled, learning outperforms the mapping baseline across all macro-level measures.

Overall, we find that learning using Freebase tags and Web search results as alternative information sources gives comparable performance–instance-level macro-F1 is .58 vs. .59, respectively.

Finally, classification results using ReVerb features ('RV') were poor, and were therefore omitted from the table. The coverage of these features was low, i.e., many of the example entity names were found no matching ReVerb assertions. In addition, as mentioned before, ReVerb patterns are noisy in that they do not distinguish between general noun phrases and named entities. Nevertheless, we found these features to be useful in hybrid configurations, as discussed next.

Multi-source learning. We now review named entity typing results using multi-source learning. The modeling of both Web search and FB features improves on either one of the individual sources, reaching macro-$F1$ performance of .63. As shown, this gain is due to a large improvement in macro-precision, reflected also in higher entity-level accuracy (.32). We thus observe that Web search results and Freebase tags are complimentary in that integrating the two perspectives serves to eliminate noise, while maintaining high recall.

The best-performing system models all of the three information sources ('WS+FB+RV'): macro-precision, recall and $F1$ all measure .65, improving 20.4%, 8.3% and 14% over the mapping baseline, respectively. Similarly, accuracy improves 13% over the baseline. In what follows, we set MS-NET to this best-performing configuration.

Performance by type. Table 4 details the performance of each of the binary classifiers that comprise MS-NET, comparing it against the rule-based mapping baseline. The best results per category are shown in boldface. Table 5 further shows the features with the highest information gain per category.

As shown in Table 4, a boost in performance was achieved using MS-NET for the categories *facility* ($\Delta F1$=52.6%), *sportsteam* (30.8%) and *product* (15.9%). These types apply to a variety of long tail entities, like local restaurants, or amateur sports teams, which are not indexed in Freebase but can be found on the Web (see dataset statistics in Table 2). Moderate performance gains were obtained for the categories *company* (8.3%) and *geolocation* (5.6%).

MS-NET failed to improve on the manual alignment rules for highly-granular categories like *TV-show* and *movie*. As indicated in Table 2, almost all of the name mentions of these categories in our twitter evaluation dataset are included in Freebase. Automatic imports from external data sources like Wikipedia ensure that FB information in these areas is complete and up-to-date.

Finally, learning failed to improve on person name typing. We found one reason for this to be high ambiguity of person names, resulting sometimes in bias of the search results towards more popular entities with the same name. Consider, for example, the celebrity names *Paris* or *MAC*–in these cases, mentions of the *city* and *product*, respectively, dominate the search results. Sampling of the search results may correct possible bias towards highly popular entities (Anagnostopoulos et al., 2006). Lastly, some of the annotated person names in our dataset refer to non-public figures that have neither KB-, nor Web presence (Gattani et al., 2013).

7 Conclusion

We have presented and evaluated a distantly supervised multi-source learning approach for semantic named entity typing. We believe this work to be the first to model KB information, Web-scale distributional evidence and Web search results as information sources in a joint framework, and the first to evaluate the task of named entity typing in the social media domain.

Our results indicate these various sources to be complimentary–their combination yields best overall performance with respect to both precision and recall. In particular, multi-source learning boosts typing performance for semantic types that apply to out-of-KB entity names, which are prevalent in social media.

One may easily incorporate additional information sources in this general framework–such as additional KBs, lexicons, additional search engines' results, and more elaborate representations of distributional semantics–so as to increase coverage and downweight source-specific bias. Learning-wise, multi-label schemes that model inter-label dependencies may prove useful (Madjarova et al., 2012). Considering the variance in performance across categories, we suggest to identify the best configuration of information sources per target semantic type, possibly using meta-learning.

References

Aris Anagnostopoulos, Andrei Z. Broder, and David Carmel. 2006. Sampling search-engine results. *World Wide Web*, 9(4).

Sőren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. 2007. Open information extraction from the web. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Sfreebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD international conference on Management of data (SIGMOD)*.

Antoine Bordes, Jason Weston, and Nicolas Usunier. 2014. Open question answering with weakly supervised embedding models. In *Proceedings of ECML-PKDD*. Springer.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*.

David Carmel, Ming-Wei Chang, Evgeniy Gabrilovich, Bo-June (Paul) Hsu, and Kuansan Wang. 2014. Erd'14: Entity recognition and disambiguation challenge. *ACM SIGIR Forum*, 48(2).

Kai-Wei Chang, Wen tau Yih, Bishan Yang, and Christopher Meek. 2014. Typed tensor decomposition of knowledge bases for relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marco Cornolti, Paolo Ferragina, Massimiliano Ciaramita, Stefan Rued, and Hinrich Schuetze. 2014. The smaph system for query entity recognition and disambiguation. In *ERD 2014: Entity Recognition and Disambiguation Challenge. SIGIR Forum.*

Qing Cui, Bin Gao, Jiang Bian, Siyu Qiu, Hanjun Dai, and Tie-Yan Liu. 2015. Knet: A general framework for learning word embedding using morphological knowledge. *ACM Transactions on Information Systems (TOIS)*, 34(1).

Bhavana Bharat Dalvi, Einat Minkov, Partha Pratim Talukdar, and William W. Cohen. 2015. Automatic gloss finding for a knowledge base using ontological constraints. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphal Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, 51:32–49.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yuan Fang and Ming-Wei Chang. 2014. Entity linking on microblogs with spatial and temporal signals. *Transactions of the Association for Computational Linguistics*, 2.

Abhishek Gattani, Digvijay S. Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Arnand Rajaraman, Venky Harinarayan, and AnHai Doan. 2013. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *Proceedings of the VLDB Endowment*, 6(11):1126–1137.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations*, 11(1).

Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 385–396.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Thomas Lin, Mausam, and Oren Etzioni. 2012. No noun phrase left behind: Detecting and typing unlinkable entities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL)*.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI)*.

Gjorgji Madjarova, Dragi Kocevb, Dejan Gjorgjevikja, and Sašo Džeroskib. 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9).

Marius Pasca. 2004. Acquisition of categorized named entities for web search. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*.

Evgenia Wasserman Pritsker, William W. Cohen, and Einat Minkov. 2015. Learning to identify the best contexts for knowledge-based wsd. In *Proceedings* *of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier chains for multi-label classification. In W. Buntine, M. Grobelnik, D. Mladenic, and J. Shawe-Taylor, editors, *Lecture Notes in Artificial Intelligence*. Springer.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *EMNLP*.

Xin Rong, Zhe Chen, Qiaozhu Mei, and Eytan Adar. 2016. Egoset: Exploiting word ego-networks and user-generated ontology for multifaceted set expansion. In *ACM International Conference on Web Search and Data Mining (WSDM)*.

Lev Ratinov Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Stefan Rüd, Massimiliano Ciaramita, Jens Müller, and Hinrich Schütze. 2011. Piggyback: using search engines for robust cross-domain named entity recognition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2006. Building bridges for web query classification. In *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *Proceedings of the International World Wide Web Conference (WWW)*.

Prashant Ullegaddi and Vasudeva Varma. 2011. Learning to rank categories for web queries. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*.

Chi Wang, Kaushik Chakrabarti, Tao Cheng, and Surajit Chaudhuri. 2012. Targeted disambiguation of ad-hoc, homogeneous sets of named entities. In *Proceedings of the International World Wide Web Conference (WWW)*.

Robert West, Evgeniy Gabrilovich Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge base completion via search-based question answering. In *Proceedings of the international conference on World wide web (WWW)*.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2015. Corpus-level fine-grained entity typing using contextual information. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ikuya Yamada, Hideaki Takeda, and Yoshiyasu Takefuji. 2015. Enhancing named entity recognition in messages using entity linking. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.