Incorporating Global Information into Named Entity Recognition Systems using Relational Context

Yuval Merhav
Illinois Institute of Technology
yuval@ir.iit.edu

Filipe Mesquita
University of Alberta, Canada
mesquita@cs.ualberta.ca

Denilson Barbosa
University of Alberta, Canada
denilson@cs.ualberta.ca

Wai Gen Yee
Illinois Institute of Technology
yee@iit.edu

Ophir Frieder
Georgetown University
ophir@cs.georgetown.edu

ABSTRACT

The state-of-the-art in Named Entity Recognition relies on a combination of local features of the text and global knowledge to determine the types of the recognized entities. This is problematic in some cases, resulting in entities being classified as belonging to the wrong *type*. We show that using global information about the *corpus* improves the accuracy of type identification. We explore the notion of a *global* domain frequency that relates relation-identifying *terms* with pairs of entity types which are used in that relation. We use this to identify entities whose types are not compatible with the terms they co-occur in the text. Our results on a large corpus of social media content allows the identification of mistyped entities with 70% accuracy.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Experimentation, Performance

1. INTRODUCTION

Named Entity Recognition (NER) is an important task for many Information Retrieval applications. One sub-task of particular importance is *type identification*: assigning meaningful types (e.g., Person, Organization, Location, etc.) to the extracted entities. The state-of-the-art NER systems rely on a mix of *local* information (statistics and results of lexical analysis) about small portions of the corpora and external knowledge (usually obtained through learning on training data) to perform type identification. For example, LBJ [3] analyzes the corpus in fixed-sized text windows ignoring document boundaries and relies on two sources of external information: high-precision lists of named entities (gazetteers) and clusters of commonly used words in different contexts. While the use of external information has been shown to improve accuracy over purely local methods, they are limited to the knowledge contained in a small collection of documents or tokens for every entity type assignment decision they make. Inevitably, these methods eventually assign incorrect types to the entities they extract. As an example, consider the snippet "[MISC Jewish] by birth, [ORG Alamo] married [PER Edith Opal] who was also [MISC Jewish]" which is tagged with LBJ. As one can see, Alamo is correctly identified as an entity, but assigned type ORG (for Organization) instead of PER (for Person). As Alamo appears in LBJ's gazetteers as an organization, it is likely that this is the reason LBJ labeled Alamo incorrectly.

To overcome the limitation, we propose a scoring feature based on global information (extracted from the entire corpus) for improving the assignment accuracy of entity types. The feature is designed to be used as a supplement in different NER systems and other tasks. This work is motivated by our social network extractor system SONEX [2], that extracts latent social networks from social media text (in [2] we report results on the ICWSM'09 Spinn3r Blog dataset with 44 millions posts [1]). SONEX works by extracting named entities with LBJ, and individual sentences with LingPipe[1]. Then it identifies relations that associate *pairs of entities* (e.g., Alamo and Edith Opal) by clustering those sentences using a variety of features. In the example above, SONEX identifies that Alamo and Edith are *married* to each other, and thus assigns this term to the pair. In this work, we show how to automatically use the results of this relation extraction in SONEX to identify entities which are incorrectly typed (Alamo in the example).

2. DOMAIN FREQUENCY

It is natural to expect that the relations extracted in SONEX are strongly correlated with a given context. For instance, marriage is a relation between Persons, and thus, belongs to the domain $\mathrm{PER}-\mathrm{PER}$. We exploit this observation to identify mistyped entities. Starting from the social network extracted by SONEX, which we call the dataset in the sequel, we proceed as follows. For each relation-identifying term in the dataset (excluding stop words), we compute its relative frequency in every possible domain. We refer to the frequency of a term t in a given domain i as the term's *Domain Frequency* (DF) score, and refer to it as $df_i(t)$.

We normalize the domain frequency of every term across all domains associated with the term. More precisely, let t be a term and let $i_1, \ldots, i_n$ be all possible domains of pairs of entity types; let $f_i(t), \ldots, f_n(t)$ be the frequencies in which that term identifies a relation of a pair of entities in each of the domains. Then,

$$df_i(t) = \frac{f_i(t)}{\sum_{1 \le j \le n} f_j(t)}.$$

The number of possible domains is the square of the number of types the NER system identifies. LBJ offers the following types: PER (Person), ORG (Organization), LOC (Location), and MISC

[1] http://alias-i.com/lingpipe

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

PER − PER	0.8409
ORG − ORG	0.0681
ORG − PER	0.0227

Table 1: Top-3 Domain Frequencies for "married" in the Spinn3r social network extracted with SONEX.

(miscellaneous)[2]. Table 1 shows the most significant relative DF scores for the term married across different domains. As expected, the DF score for the PER to PER domain is significantly larger than all other domains. If a term does not appear in a certain domain, its DF score for the domain is zero (e.g., "married" does not appear in the LOC to LOC domain, and hence, it does not appear in Table 1). The entire list of terms associated with their DF scores is available by request. This list can be used as an external knowledge source in different NER systems, in various dataset domains.

2.1 Detecting Incorrectly Typed Entities

Our premise is that a pair of entities (E_1, E_2) from a given domain $d_1 = T_1 \times T_2$ contains at least one mistyped entity if there is a relationship-identifying term t that connects them in the dataset, *and* the term's domain frequency for d_1 is "significantly" lower than that of another different domain d_2. Our hypothesis is that we can detect mistyped entities by comparing such domain frequencies. We validate it as follows.

Setup. Since the ICWSM'09 dataset does not include labels for named entities, we identify all type errors through manual evaluation. The complete evaluation we performed is available by request. We compute the list of terms in the dataset and their associated DF scores. Then, we obtain two random entity sets for our evaluation. The first, RANDOM contains 70 entity pairs (thus 140 typed entities) randomly chosen from the dataset. The second, SUSPICIOUS, consists of 326 entity pairs for which the *gap* in domain frequencies (highest to lowest) is higher a threshold.

More precisely, SUSPICIOUS is obtained as follows. For each term t in the dataset, we compute the difference between its highest and the lowest DF scores. We keep those terms for which this difference is larger than 0.5 (empirically tested to produce high precision with reasonable recall), resulting in 482 unique terms (i.e., relations). From these, we randomly chose 30 to obtain the entity pairs in our tests. For each term, we gathered up to 15 entity pairs, resulting in a total of 326 unique entity pairs (some terms were associated with less than 15 pairs), that generated 375 entities for the SUSPICIOUS evaluation (for most of the entity pairs, only one entity out of the two in a domain is identified as a mistake).

Hypothesis. Our hypothesis is that pairs in the SUSPICIOUS set are more likely to contain at least one mistyped entity. This is justified by the following observation: out of the 375 entities, 269 (or 72%) are of type ORG, 83 (22%) are of type LOC, and 23 (6%) are of type PER. This distribution is very different than the one that takes into account all entities in the dataset: 12% for ORG, 43% for LOC, and 45% for PER. The type ORG, which appears in only 12% of all entities, appears in 72% of the entities in the SUSPICIOUS set, implying a higher error rate for organizations compared to other entity types.

Table 2 validates this hypothesis. In the table, precision is the number of entities with correct types (we did not consider the correctness of the entity boundary) divided by the total number of entities in each set. Weighted is the weighted average of the precision for the three types. Observe that in the RANDOM set entities are correctly typed 62% of the times, whereas in the SUSPICIOUS set this happens only 30% of the time. This 30% reflects the entities we incorrectly identified as mis-typed. Also, observe that LBJ is much more accurate in correctly identifying persons and locations compared to organizations.

[2]We do not consider the MISC type in our experiments. They are too generic, making it hard to evaluate whether the entities assigned to this type are misclassified by LBJ.

	RANDOM	SUSPICIOUS
PER	0.79	0.43
LOC	0.70	0.44
ORG	0.43	0.25
Weighted	0.62	0.30

Table 2: Correctness of the entity types in the 2 evaluation sets

Method. The significant observation from our experiment is that the difference in domain frequency scores may be an effective way of identifying mis-typed entities. It effectively yields a very simple and automatic procedure for detecting incorrect type assignments which has 70% precision. Given the much lower rate mistyping rate of just 38% for the RANDOM set, these results are promising.

3. CONCLUSION

We proposed the use of Domain Frequency scores to predict entities which are erroneously typed by NER systems. This measure can be readily incorporated into existing NER systems with ease. DF exploits terms between pairs of entities to estimate the likelihood of a term to appear between given entity types. DF relies on global (corpus-wide) information as is thus sensitive to the domain at hand. We showed experimentally that the difference in DF scores for a given term serves as a good indicator that the entities associated through that term are incorrectly typed, and that this simple rule was able to detect an entity with incorrect type in 70% of the cases, and that this rate is much higher than that of a random sample of the dataset.

We are investigating ways in which to use the DF scores to further improve not only entity type identification but also the extraction of the relations among the entities. We envision a mutual refinement scheme in which both tasks go hand-in-hand.

4. ACKNOWLEDGEMENTS

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Alberta Ingenuity Fund.

5. REFERENCES

[1] K. Burton, A. Java, and I. Soboroff. The icwsm 2009 spinn3r dataset. In *ICWSM '09: Proceedings of the 3rd Int'l AAAI Conference on Weblogs and Social Media*, 2009.

[2] F. Mesquita, Y. Merhav, and D. Barbosa. Extracting information networks from the blogosphere: State-of-the-art and challenges. In *ICWSM '10: Proceedings of the 4th Int'l AAAI Conference on Weblogs and Social Media*, 2010.

[3] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *CoNLL '09: Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 147–155, Morristown, NJ, USA, 2009. Association for Computational Linguistics.