A Weighting Scheme for Open Information Extraction

Yuval Merhav
Illinois Institute of Technology
Chicago, IL USA
yuval@ir.iit.edu

Abstract

We study[1] the problem of extracting all possible relations among named entities from unstructured text, a task known as Open Information Extraction (Open IE). A state-of-the-art Open IE system consists of natural language processing tools to identify entities and extract sentences that relate such entities, followed by using text clustering to identify the relations among co-occurring entity pairs. In particular, we study how the current weighting scheme used for Open IE affects the clustering results and propose a term weighting scheme that significantly improves on the state-of-the-art in the task of relation extraction both when used in conjunction with the standard $tf \cdot idf$ scheme, and also when used as a pruning filter.

1 Introduction

The extraction of structured information from text is a long-standing challenge in Natural Language Processing which has been re-invigorated with the ever-increasing availability of user-generated textual content online. The large-scale extraction of unknown relations has been termed as Open Information Extraction (Open IE) (Banko et al., 2007) (also referred to as Open Relationship Extraction, Relation Extraction, or Relation Discovery). Many challenges exist in developing an Open IE solution, such as recognizing and disambiguating entities in a multi-document setting, and identifying all so-called *relational* terms in the sentences connecting pairs of entities. Relational terms are words (usually one or two) that describe a relation between entities (for instance, terms like "running mate", "opponent", "governor of" are relational terms).

One approach for Open IE is based on clustering of entity pairs to produce relations, as introduced by Hasegawa et al. (Hasegawa et al., 2004). Their and follow-up works (e.g., (Mesquita et al., 2010)) extract terms in a small window between two named entities to build the context vector of each entity pair, and then apply a clustering algorithm to cluster together entity pairs that share the same relation (e.g., `Google–Youtube` and `Google–Motorola Mobility` in a cluster about the "acquired" relation). Contexts of entity pairs are represented using the vector space model. The state-of-the-art in clustering-based Open IE assigns weights to the terms according to the standard $tf \cdot idf$ scheme.

Motivation. Intuitively, the justification for using idf is that a term appearing in many documents (i.e., many contexts in our setting) would not be a good *discriminator* (Robertson, 2004), and thus should weigh proportionally less than other, more *rare* terms. For the task of relation extraction however, we are interested specifically in terms that describe relations. In our settings, a single document is a context vector of one entity pair, generated from all articles discussing this pair, which means that the fewer entity pairs a term appears in, the higher its idf score would be. Consequently, it is not necessarily the case that terms that are associated with high idf weights would be good relation discriminators. On the other hand, popular relational terms that ap-

[1] This thesis proposal has been accepted for publication in (Merhav et al., 2012).

Proceedings of the NAACL HLT 2012 Student Research Workshop, pages 60–65,
Montréal, Canada, June 3-8, 2012.

ply to many entity pairs would have relatively lower idf weights.

It is natural to expect that the relations extracted by an Open IE system are strongly correlated with a given context. For instance, marriage is a relation between two persons and thus belongs to the domain PER–PER. We exploit this observation to boost the weight of relational terms associated with marriage (e.g., "wife", "spouse", etc.) in those entity pairs where the domain is also PER–PER. The more dominant a term in a given domain compared to other domains, the higher its boosting score would be.

Our work resembles the work on selectional preferences (Resnik, 1996). Selectional preferences are semantic constraints on arguments (e.g. a verb like "eat" prefers as object edible things).

2 Related Work

Different approaches for Open IE have been proposed in the literature, such as bootstrapping (e.g., (Zhu et al., 2009) (Bunescu and Mooney, 2007)), self or distant supervision (e.g., (Banko et al., 2007) (Mintz et al., 2009)) and rule based (e.g., (Fader et al., 2011)). In this work we focus on unsupervised approaches.

Fully unsupervised Open IE systems are mainly based on clustering of entity pair contexts to produce clusters of entity pairs that share the same relations, as introduced by Hasegawa et al. (Hasegawa et al., 2004) (this is the system we use in this work as our baseline). Hasegawa et al. used word unigrams weighted by $tf \cdot idf$ to build the context vectors and applied Hierarchical Agglomerative Clustering (HAC) with complete linkage deployed on a 1995 New York Times corpus. Mesquita et al. extended this work by using other features such as part of speech patterns (Mesquita et al., 2010). To reduce noise in the feature space, a common problem with text mining, known feature selection and ranking methods for clustering have been applied (Chen et al., 2005; Rosenfeld and Feldman, 2007). Both works used the K-Means clustering algorithm with the stability-based criterion to automatically estimate the number of clusters.

This work extends all previous clustering works by utilizing domain frequency as a novel weighting scheme for clustering entity pairs. The idea of domain frequency was first proposed for predicting entities which are erroneously typed by NER systems (Merhav et al., 2010).

3 Data and Evaluation

This work was implemented on top of the SONEX system (Mesquita et al., 2010), deployed on the ICWSM 2009 Spinn3r corpus (Burton et al., 2009), focusing on posts in English (25 million out of 44 million in total), collected between August 1st, 2008 and October 1st, 2008. The system uses the Illinois Entity Tagger (Ratinov and Roth, 2009) and Orthomatcher from the GATE framework[2] for within-a-document co-reference resolution.

Evaluating Open IE systems is a difficult problem. Mesquita et al. evaluated SONEX by automatically matching a sample of the entity pairs their system identified from the Spinn3r corpus against a publicly available curated database[3]. Their approach generated two datasets: INTER and 10PERC. INTER contains the intersection pairs only (i.e., intersection pairs are those from Spinn3r and Freebase that match both entity names and types exactly), while 10PERC contains 10% of the total pairs SONEX identified, including the intersection pairs. We extended these two datasets by adding more entity pairs and relations. We call the resulting datasets INTER (395 entity pairs and 20 different relations) and NOISY (contains INTER plus approximately 30,000 entity pairs as compared to the 13,000 pairs in 10PERC).

We evaluate our system by reporting f-measure numbers for our system running on INTER and NOISY against the ground truth, using similar settings used by (Hasegawa et al., 2004) and (Mesquita et al., 2010). These include word unigrams as features, HAC with average link (outperformed single and complete link), and $tf \cdot idf$ and cosine similarity as the baseline.

4 Weighting Scheme

Identifying the relationship (if any) between entities e_1, e_2 is done by analyzing the sentences that mention e_1 and e_2 together. An *entity pair* is defined by two entities e_1 and e_2 together with the *context* in

[2] http://gate.ac.uk/

[3] http://www.freebase.com

which they co-occur. For our purposes, the context can be any textual feature that allows the identification of the relationship for the given pair. The contexts of entity pairs are represented using the vector space model with the common $tf \cdot idf$ weighting scheme. More precisely, for each term t in the context of an entity pair, tf is the frequency of the term in the context, while

$$idf = \log\left(\frac{|D|}{|d : t \in d|}\right),$$

where $|D|$ is the total number of entity pairs, and $|d : t \in d|$ is the number of entity pairs containing term t. The standard cosine similarity is used to compute the similarity between context vectors during clustering.

4.1 Domain Frequency

We start with a motivating example before diving into the details about how we compute *domain frequency*. We initially built our system with the traditional $tf \cdot idf$ and were unsatisfied with the results. Consequently, we examined the data to find a better way to score terms and filter noise. For example, we noticed that the pair *Youtube[ORG] – Google[ORG]* (associated with the "Acquired by" relation) was not clustered correctly. In Table 1 we listed all the Unigram features we extracted for the pair from the entire collection sorted by their domain frequency score for `ORG–ORG` (recall that these are the intervening features between the pair for each co-occurrence in the entire dataset). For clarity the terms were not stemmed.

Clearly, most terms are irrelevant which make it difficult to cluster the pair correctly. We listed in bold all terms that we think are useful. Besides "belongs", all these terms have high domain frequency scores. However, most of these terms do not have high idf scores. Term frequencies within a pair are also not helpful in many cases since many pairs are mentioned only a few times in the text. Next, we define the domain frequency score (Merhav et al., 2010).

Definition. Let P be the set of entity pairs, let T be the set of all entity types, and let $D = T \times T$ be the set of all possible relation domains. The *domain frequency* (df) of a term t, appearing in the context of some entity pair in P, in a given relation domain $i \in D$, denoted $df_i(t)$, is defined as

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

where $f_i(t)$ is the frequency with which term t appears in the context of entity pairs of domain $i \in D$, and n is the number of domains in D. When computing the df score for a given term, it is preferred to consider each pair only once. For example, *"Google[ORG] acquired Youtube[ORG]"* would be counted only once (for "acquired" in the `ORG–ORG` domain) even if this pair and context appear many times in the collection. By doing so we eliminate the problem of duplicates (common on the web).

Unlike the idf score, which is a *global* measure of the discriminating power of a term, the df score is domain-specific. Thus, intuitively, the df score would favour specific relational terms (e.g., "wife" which is specific to personal relations) as opposed to generic ones (e.g., "member of" which applies to several domains). To validate this hypothesis, we computed the df scores of several relational terms found in the clusters the system produced on the main Spinn3r corpus.

Figure 1 shows the relative df scores of 4 relational terms (**mayor**, **wife**, **CEO**, and **coach**) which illustrate well the strengths of the df score. We can see that for the majority of terms (Figure 1(a)–(c)), there is a single domain for which the term has a clearly dominant df score: `LOC–PER` for **mayor**, `PER–PER` for **wife**, and `ORG–PER` for **CEO**.

Dependency on NER Types. Looking again at Figure 1, there is one case in which the df score does not seem to discriminate a reasonable domain. For **coach**, the dominant domain is `LOC–PER`, which can be explained by the common use of the city (or state) name as a proxy for a team as in the sentence "Syracuse football coach Greg Robinson". Note, however, that the problem in this case is the difficulty for the NER to determine that "Syracuse" refers to the university. These are some examples of correctly identified pairs in the **coach** relation but in which the NER types are misleading:

- `LOC–PER` domain: (England, Fabio Capello); (Croatia, Slaven Bilic); (Sunderland, Roy Keane).

Table 1: Unigram features for the pair *Youtube[ORG] – Google[ORG]* with idf and df (ORG–ORG) scores

Term	idf	df **(ORG–ORG)**	**Term**	idf	df **(ORG–ORG)**
ubiquitious	11.6	1.00	blogs	6.4	0.14
sale	5.9	0.80	services	5.9	0.13
parent	6.8	0.78	instead	4.0	0.12
uploader	10.5	0.66	free	5.0	0.12
purchase	6.3	0.62	similar	5.7	0.12
add	6.1	0.33	recently	4.2	0.12
traffic	7.0	0.55	disappointing	8.2	0.12
downloader	10.9	0.50	dominate	6.4	0.11
dailymotion	9.5	0.50	hosted	5.6	0.10
bought	5.2	0.49	hmmm	9.3	0.10
buying	5.8	0.47	giant	5.4	< 0.1
integrated	7.3	0.44	various	5.7	< 0.1
partnership	6.7	0.42	revealed	5.2	< 0.1
pipped	8.9	0.37	experiencing	7.7	< 0.1
embedded	7.6	0.36	fifth	6.5	< 0.1
add	6.1	0.33	implication	8.5	< 0.1
acquired	5.6	0.33	owner	6.0	< 0.1
channel	6.3	0.28	corporate	6.4	< 0.1
web	5.8	0.26	comments	5.2	< 0.1
video	4.9	0.24	according	4.5	< 0.1
sellout	9,2	0.23	resources	6.9	< 0.1
revenues	8.6	0.21	grounds	7.8	< 0.1
account	6.0	0.18	poked	6.9	< 0.1
evading	9.8	0.16	**belongs**	6.2	< 0.1
eclipsed	7.8	0.16	authors	7.4	< 0.1
company	4.7	0.15	hooked	7.1	< 0.1

- `MISC–PER` domain: (Titans, Jeff Fisher); (Jets, Eric Mangini); (Texans, Gary Kubiak).

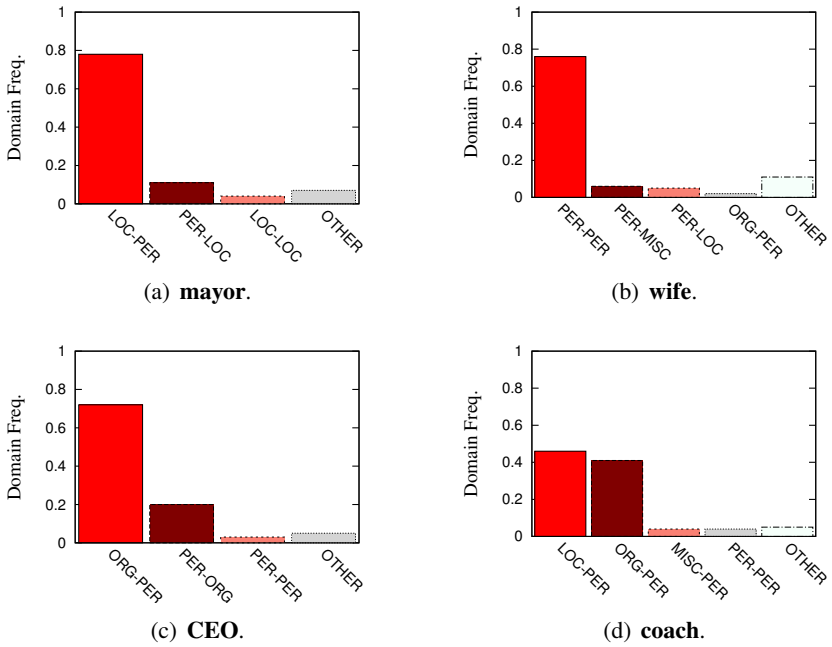
4.2 Using the df Score

We use the df score for two purposes in our work. First, for clustering, we compute the weights of the terms inside all vectors using the product $tf \cdot idf \cdot df$. Second, we also use the df score as a filtering tool, by removing terms from vectors whenever their df scores lower than a threshold. Going back to the *Youtube[ORG] – Google[ORG]* example in Table 1, we can see that minimum df filtering helps with removing many noisy terms. We also use maximum idf filtering which helps with removing terms that have high df scores only because they are rare and appear only within one domain (e.g., ubiquitious (misspelled in source) and uploader in this example).

As we shall see in the experimental evaluation, even in the presence of incorrect type assignments made by the NER tool, the use of df scores improves the accuracy significantly. It is also worth mentioning that computing the df scores can be done fairly efficiently, and as soon as all entity pairs are extracted.

5 Results

We now report the results on INTER and NOISY. Our baseline run is similar to the systems published by Hasegawa et al. (Hasegawa et al., 2004) and Mesquita et al. (Mesquita et al., 2010); that is HAC with average link using $tf \cdot idf$ and cosine similarity, and stemmed word unigrams (excluding stop words) as features extracted using a window size of five words between pair of entities. Figure 2 shows that by integrating domain frequency



(a) **mayor**. (b) **wife**.

(c) **CEO**. (d) **coach**.

Figure 1: Domain Frequency examples.

(df) we significantly outperformed this baseline on both datasets (INTER: F-1 score of 0.87 compared to 0.75; NOISY: F-1 score of 0.72 compared to 0.65). In addition, filtering terms by minimum df and maximum idf thresholds improved the results further on INTER. These results are promising since a major challenge in text clustering is reducing the noise in the data.

We also see a substantial decrease of the results on NOISY compared to INTER. Such a decrease is, of course, expected: NOISY contains not only thousands more entity pairs than INTER, but also hundreds (if not thousands) more *relations* as well, making the clustering task harder in practice.

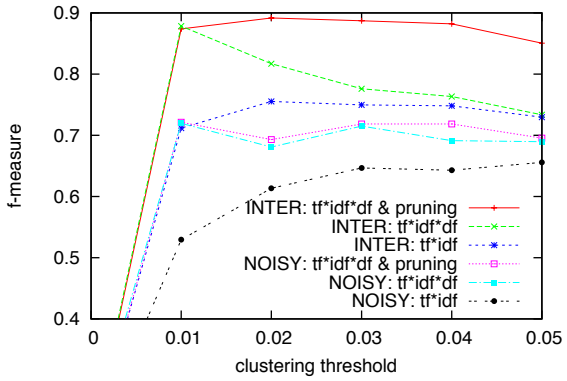


Figure 2: $tf \times idf$ Vs. $tf \times idf \times df$ with and without minimum df and maximum idf pruning on INTER and NOISY. All results consistently dropped for clustering thresholds larger than 0.05.

6 Conclusion and Future Research Directions

We utilized the Domain Frequency (df) score as a term-weighting score designed for identifying relational terms for Open IE. We believe that df can be utilized in various of applications, with the advantage that in practice, for many such applications, the list of terms and scores can be used off-the-shelf with no further effort. One such application is Named Entity Recognition (NER) – df helps in identifying relational patterns that are associated with a certain domain (e.g., `PER–PER`). If the list of words and phrases associated with their df scores is generated using an external dataset annotated with entities, it can be applied to improve results in other, more difficult domains, where the performance of the NER is poor.

It is also appealing that the df score is probabilistic, and as such, it is, for the most part, language independent. Obviously, not all languages

have the same structure as English and some adjustments should be made. For example, df exploits the fact that relational verbs are usually placed between two entities in a sentence, which may not be always the case in other languages (e.g., German). Investigating how df can be extended and utilized in a multi-lingual environment is an interesting future direction.

7 Acknowledgements

The author would like to thank Professor Denilson Barbosa from the University of Alberta, Professor David Grossman and Gady Agam from Illinois Institute of Technology, and Professor Ophir Frieder from Georgetown University. All provided great help in forming the ideas that led to this work.

References

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In Manuela M. Veloso, editor, *IJCAI*, pages 2670–2676.

Razvan C. Bunescu and Raymond J. Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *ACL*.

K. Burton, A. Java, and I. Soboroff. 2009. The icwsm 2009 spinn3r dataset. In *Proceedings of the Annual Conference on Weblogs and Social Media*.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2005. Unsupervised feature selection for relation extraction. In *IJCNLP-05: The 2nd International Joint Conference on Natural Language Processing*. Springer.

Anthony Fader, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. Manuscript submitted for publication. University of Washington.

Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. 2004. Discovering relations among named entities from large corpora. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 415, Morristown, NJ, USA. Association for Computational Linguistics.

Yuval Merhav, Filipe Mesquita, Denilson Barbosa, Wai Gen Yee, and Ophir Frieder. 2010. Incorporating global information into named entity recognition systems using relational context. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 883–884, New York, NY, USA. ACM.

Yuval Merhav, Filipe Mesquita, Denilson Barbosa, Wai Gen Yee, and Ophir Frieder. 2012. Extracting information networks from the blogosphere. *ACM Transactions on the Web (TWEB)*. Accepted 2012.

Filipe Mesquita, Yuval Merhav, and Denilson Barbosa. 2010. Extracting information networks from the blogosphere: State-of-the-art and challenges. In *4th Int'l AAAI Conference on Weblogs and Social Media–Data Challenge*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 1003–1011, Morristown, NJ, USA. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, Morristown, NJ, USA. Association for Computational Linguistics.

Philip Resnik. 1996. Selectional constraints: an information-theoretic model and its computational realization.

Stephen Robertson. 2004. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:2004.

Benjamin Rosenfeld and Ronen Feldman. 2007. Clustering for unsupervised relation identification. In *CIKM '07*, pages 411–418, New York, NY, USA. ACM.

Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. 2009. Statsnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 101–110, New York, NY, USA. ACM.