Named Entity Extraction Using Information Distance

Sangameshwar Patil SANGAMESHWAR.PATIL@TCS.COM
Sachin Pawar SACHIN7.P@TCS.COM
Girish K. Palshikar GK.PALSHIKAR@TCS.COM
Tata Research Development and Design Centre, 54B, Hadapsar Industrial Estate, Pune 411013, India

Abstract

Named Entity extraction (NEX) problem consists of automatically constructing a gazette containing instances for each NE of interest. NEX is important for domains which lack a corpus with tagged NEs. In this paper, we propose a new unsupervised (bootstrapping) NEX technique, based on a new variant of the Multiword Expression Distance (MED) (Bu et al., 2010) and information distance (Bennett et al., 1998). Efficacy of our method is shown using comparison with BASILISK and PMI in agriculture domain. Our method discovered 8 new diseases which are not found in Wikipedia.

1. Introduction

The problem of information extraction for agriculture is particularly important as well as challenging due to non-availability of any tagged corpus. Several domain-specific *named entities (NE)* occur in the documents (such as news) related to the agriculture domain: CROP (names of the crop including varieties), DISEASE (names of the crop diseases and disease causing agents such as bacteria, viruses, fungi, insects etc.) and CHEMICAL_TREATMENT (names of pesticides, insecticides, fungicides etc.). *NE extraction (NEX)* problem consists of automatically constructing a gazette containing example instances for each NE of interest.

2. Information Distance for NE

(Bu et al., 2010) presented a variant of the information distance, named Multiword Expression Distance (MED), which measures the distance between an n-gram and its semantics. In this paper, we use a variant of MED to perform NEX. Let $\mathbf{D}$ be a given untagged corpus of sentences. Let K be a given constant indicating the window size (e.g., $K = 3$). Let g be a given candidate phrase. The *context* of g and a given word w, denoted $\phi_K(g, w)$, is the set of all sentences in $\mathbf{D}$ which contain both g (as an n-gram) and w and w occurs within a window of size K around g in that sentence. The *semantics* of g and a given word w, denoted $\mu(g, w)$, is the set of all sentences in $\mathbf{D}$ which contain both g (as an n-gram) and w, though g and w need not be within a window of size K in the sentence. Clearly, $\phi_K(g, w) \subseteq \mu(g, w)$. Then we define the *distance* between g and a given word w as follows:

$$MED0_{D,K}(g, w) = log|\mu(g, w)| - log|\phi_K(g, w)|$$

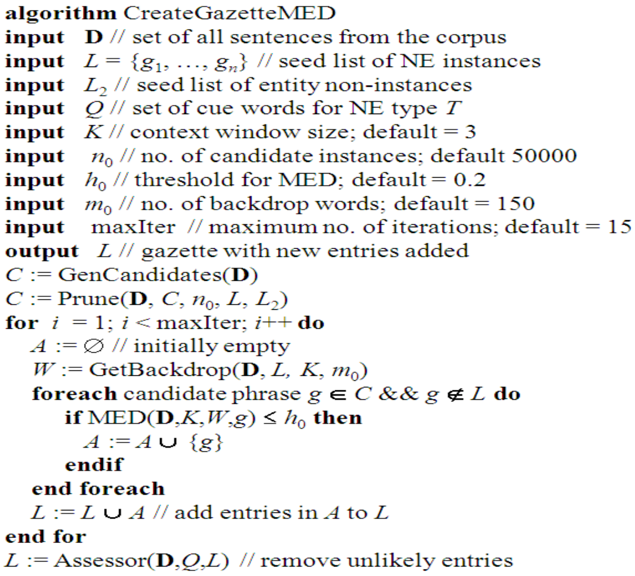
Let $W = \{w_1, w_2, \ldots, w_m\}$ be a given finite, nonempty set of m words. $MED_{D,K}(g, W)$ is defined as the average of the $MED0$ distance between g and each word in W.

3. Unsupervised Gazette Creation Using MED

In unsupervised gazette creation, we are given (i) an untagged corpus $\mathbf{D}$ of documents; and (ii) a seed list L containing known examples of a particular NE type T. The goal is to create a gazette containing other instances of the NE type T that occur in $\mathbf{D}$.

The **algorithm** *CreateGazetteMED* (Fig 1) starts with an initial seed list L of instances of a particular NE type T and the corpus $\mathbf{D}$. The **pre-processing step** consisting of *GenCandidates* (to extract noun phrases) and *Prune* produces a list C of candidate phrases. The function *Prune* consists of a Maximum Entropy classifier used in self-training iterative mode and bootstrapped from the seed instances in L. Then in each iteration, the algorithm *GetBackdrop* creates the set W of backdrop words for T using L. *GetBackDrop* identifies a set of context words (i.e. "backdrop") which characterize the NE of a

Appearing in Proceedings of the 1^{st} *Indian Workshop on Machine Learning*, IIT Kanpur, India, 2013. Copyright 2013 by the author(s).



```
algorithm CreateGazetteMED
input   D // set of all sentences from the corpus
input   $L = \{g_1, \ldots, g_n\}$ // seed list of NE instances
input   $L_2$ // seed list of entity non-instances
input   $Q$ // set of cue words for NE type $T$
input   $K$ // context window size; default = 3
input   $n_0$ // no. of candidate instances; default 50000
input   $h_0$ // threshold for MED; default = 0.2
input   $m_0$ // no. of backdrop words; default = 150
input   maxIter // maximum no. of iterations; default = 15
output  $L$ // gazette with new entries added
$C$ := GenCandidates(D)
$C$ := Prune(D, $C$, $n_0$, $L$, $L_2$)
for $i$ = 1; $i$ < maxIter; $i$++ do
    $A := \varnothing$ // initially empty
    $W$ := GetBackdrop(D, $L$, $K$, $m_0$)
    foreach candidate phrase $g \in C$ && $g \notin L$ do
        if MED(D,$K$,$W$,$g$) $\le h_0$ then
            $A := A \cup \{g\}$
        endif
    end foreach
    $L := L \cup A$ // add entries in $A$ to $L$
end for
$L$ := Assessor(D,$Q$,$L$) // remove unlikely entries
```

Figure 1. Algorithm *CreateGazetteMED*.

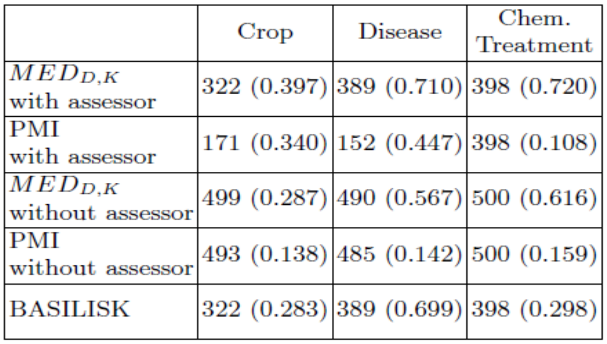
given type. Each backdrop word is weighed using a relevance score by taking into account its frequency, fraction of seeds for which this word is relevant as a backdrop and its entropy with respect to different NE types. *CreateGazetteMED* then uses the modified MED to measure the similarity of each candidate phrase $g \in C$ with W; adding a pre-specified number of candidate phrases with "high" (above a threshold) similarity with W to a temporary set A. At the end of *maxIter*, final set of candidates in L is pruned by the post-processing step. The **post-processing step** consists of an *assessor* that uses a set of cue words for the NE type T and performs a statistical hypothesis test (called *proportion test*). It improves gazette quality by identifying (and removing) those entries in the candidate gazette which are very unlikely to be true instances of NE type T.

4. Experimental Evaluation

The corpus consists of 30533 agriculture news articles in English containing 999168 sentences and approximately 19 million words. Some of the seeds used for each NE type are as follows:

`CROP: wheat, cotton, corn, soybean, strawberry`
`DISEASE:sheath blight, wilt, leaf spot, scab`
`CHEMICAL_TREATMENT:di-syston, metalaxyl, keyplex`

Starting with the candidate list C and the initial seed list for T, the algorithm *CreateGazetteMED* iteratively created the final set of 500 candidates based on $MED_{D,K}$. The *assessor* is used to further prune this list to create the final gazette for each NE type.



	Crop	Disease	Chem. Treatment
$MED_{D,K}$ with assessor	322 (0.397)	389 (0.710)	398 (0.720)
PMI with assessor	171 (0.340)	152 (0.447)	398 (0.108)
$MED_{D,K}$ without assessor	499 (0.287)	490 (0.567)	500 (0.616)
PMI without assessor	493 (0.138)	485 (0.142)	500 (0.159)
BASILISK	322 (0.283)	389 (0.699)	398 (0.298)

Figure 2. Number of entries (& precision) in the final gazette.

Gazette sizes for each NE type are shown in Fig 2. Assessor improves precision for all NE types for both measures $MED_{D,K}$ and PMI. We compare the proposed algorithm with BASILISK (Thelen & Riloff, 2002). Also, to gauge the effectiveness of $MED_{D,K}$ as a proximity measure, we compare it with PMI. To highlight effectiveness of the gazettes created, we compared our DISEASE gazette with wikipedia. It was quite encouraging to find that, our gazette, though created on a limited size corpus, contained diseases/pathogens not present in Wikipedia.[1] Some of these are - `limb rot, grape colaspis, black shank, glume blotch, seed corn maggot, mexican rice borer, hard lock` .

5. Conclusions

In this paper, we proposed a new unsupervised (bootstrapping) NEX technique for automatically creating gazettes of domain-specific named entities. It is based on a new variant of the Multiword Expression Distance (MED) (Bu et al., 2010). We also compared the effectiveness of the proposed method with PMI, BASILISK (Thelen & Riloff, 2002)

References

Bennett, C.H., Gacs, P., Li, M., Vitanyi, P.M.B., and Zurek, W.H. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

Bu, F., Zhu, X., and Li, M. Measuring the noncompositionality of multiword expressions. In *COLING*, 2010.

Thelen, M. and Riloff, E. A bootstrapping method for learning seman-tic lexicons using extraction pattern contexts. In *EMNLP*, 2002.

[1] Verified on 30th January, 2013