UNSUPERVISED ENTITY CLASSIFICATION WITH WIKIPEDIA AND WORDNET

Tomáš Kliegr

UEP Prague, Knowledge Engineering Group, Czech Republic

ABSTRACT

The task of classifying entities appearing in textual annotations to an arbitrary set of classes has not been extensively researched, yet it is useful in multimedia retrieval. We proposed an unsupervised algorithm, which expresses entities and classes as Wordnet synsets and uses Lin measure to classify them. Real-time hypernym discovery from Wikipedia is used to map uncommon entities to Wordnet. Further, this paper investigates the possibility to improve the performance by utilizing the global context with simulated annealing.

1. INTRODUCTION

Analysis of textual annotations attached to various objects can provide useful information complementary to results of the analysis of the object itself. Annotations are typically short, but very informative due to use of Named Entities (NE). Although NEs have high information content, background knowledge is needed to resolve their meaning.

Named Entity Recognition (NER) is a long established discipline which aims at classifying NEs to a predefined set of classes[1]). Large labeled corpora available for this task are exploited by NER systems to learn statistical classification models. However, this approach cannot be utilized in a generic entity classification due to the data acquisition bottleneck [1].

This paper present a framework for unsupervised classification of named entities that utilizes background knowledge extracted from Wikipedia to overcome data sparsity and Wordnet similarity to perform classification (Section 3). We discuss ongoing work leading to improved results (Section 4).

2. RELATED RESEARCH

Treating classes as word categories makes entity classification a WSD problem [2], and thus a range of WSD algorithms can be directly applied. However, many WSD algorithms including [2] are supervised, which is not desirable in entity classification [1]. Further, WSD algorithms are typically constrained to finding a maximizing combination of word senses only within a local context, which is due to a combinatorial explosion typically limited to a window of several words before and after the entity. This is less suitable for textual annotations of objects, which are often too short to contain a usable local context. However, we noticed in our past work [3, 4] that object annotations tend to have common global context within the collection.

Our proposal is more similar to the recent work [1], who propose unsupervised classification algorithm that uses context vectors automatically extracted from text to represent both entities and classes and assigns entity to a class with which it has the highest similarity. Paper [1] shows that using pseudosyntactic dependencies is superior to word windows.

3. OUR FRAMEWORK

In our previous work, we have addressed the problem of classifying entities to an arbitrary set of classes by introducing a framework utilizing two algorithms Targeted Hypernym Discovery (THD) and Semantic Concept Mapping (SCM).

Semantic Concept Mapping is an unsupervised algorithm, which classifies each entity occurring in the annotation to one class; both entities and classes need to be expressed as Wordnet Synsets. The winning class has the highest Lin similarity sim_L with the entity.

$$sim_L(c_1, c_2) = \frac{2 * log\, p(lso(c_1, c_2))}{log\, p(c_1) + log\, p(c_2)} \tag{1}$$

The lowest common subsumer from the hierarchy is returned by lso, value $-log(p(c))$ is information content, $p(c)$ denotes the probability of encountering an instance of concept c. When entity is not present in Wordnet, Targeted Hypernym Discovery (THD) is used to provide a hypernym for the entity.

Targeted Hypernym Discovery builds upon the large body of available work on discovery of hypernyms with lexico-syntactic patterns from text. It is called *targeted*, because it does not extract all word-hypernym pairs like most other approaches, but only the most likely hypernym from the most suitable document. In our implementation[2], we use the GATE NLP text engineering framework[3] (see Figure 1) to extract the first hypernym from the Wikipedia article defining the entity. In our earlier work [5], we found Wikipedia to be the perfect and sustainable resource for hypernym discovery.

This paper would not be here without Krishna and Jan

[1] Typically PERSON, LOCATION, ORGANIZATION, MISC in the CONLL task: www.cnts.ua.ac.be/conll/

[2] http://nb.vse.cz/~klit01/hypernym_discovery/

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

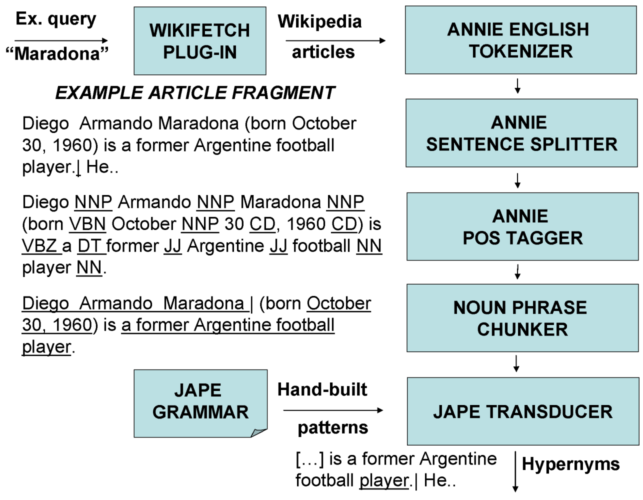
Ex. query
"Maradona"
WIKIFETCH PLUG-IN
Wikipedia articles
ANNIE ENGLISH TOKENIZER
EXAMPLE ARTICLE FRAGMENT
Diego Armando Maradona (born October 30, 1960) is a former Argentine football player.| He..
ANNIE SENTENCE SPLITTER
Diego NNP Armando NNP Maradona NNP (born VBN October NNP 30 CD, 1960 CD) is VBZ a DT former JJ Argentine JJ football NN player NN.
ANNIE POS TAGGER
Diego Armando Maradona | (born October 30, 1960) is a former Argentine football player.
NOUN PHRASE CHUNKER
JAPE GRAMMAR
Hand-built patterns
JAPE TRANSDUCER
[…] is a former Argentine football player.| He..
Hypernyms

Fig. 1. Targeted Hypernym Discovery

3.1. Use case: SCM/THD in Image Classification

So far, we have performed experiments with THD in image relevance feedback [4] and image classification [3]. Our algorithm proceeds in a similar way as a human would if presented an image annotation, a pool of possible concepts C_{tc}, and asked to express what is probably on the image using only the concepts provided: first identify the likely objects on the image by parsing the annotation for entities (noun phrases). If entity is not known, look it up in the Wikipedia. For each entity, select the class with highest semantic similarity.

```
INPUT: Annotation $ANOT$, set of concepts $C_{tc}$
OUTPUT: set of concepts $T$, $T \subseteq C_{tc}$

NP:= extractNounphrases(ANOT)
for all noun phrases np in NP do
  syn:= mapToWordnetSynsetWithTHD(np)
  maxSim:= 0, maxSimConc:= $\{\}$
  for all c in $C_{tc}$ do
    sim := wordnetSim(syn, c)
    if sim $>$ maxSim then
      maxSim:=sim, maxSimConc:=s
    end if
  end for
  $T := T \cup maxSimConc$
end for
```

Performance of SCM/THD alone was mediocre with accuracy of 27%, but combining its results with image classifier (KAA) resulted into the accuracy of 55% (relative improvement of 49% and 31% over the text/image-only baselines[4].

4. ONGOING WORK: ADAPTED E-LESK

Close analysis of the experimental results showed that the misclassification error in SCM/THD can be attributed to the first-sense assumption, due to which the system maps the first hypernym found to its first Wordnet synset, and particularly to the poor performance of the Lin measure on Wordnet.

We suggest to simultaneously address both these problems with a variation of the Lesk Algorithm [6], which uses simulated annealing to find combination of word senses that maximizes the overall similarity of dictionary definitions of words in the sentence. Instead of dictionary, we plan to use Wikipedia as a source definitions for both classes and the entities. The amount of data will be further increased by involving hypernyms discovered by THD. We intent to evaluate the performance of this approach on Fine-Grained Senseval Task.

5. CONCLUSIONS

Most NER systems use supervised techniques. However, as noted in [1], unsupervised algorithms are needed when the set of classes is larger and flexible. There is not much existing work in this area [2] as most of the research has been focusing on the typical NER task. We have proposed and implemented an unsupervised entity classification system. Further work will focus on substituting the currently used Lin measure, which uses Wordnet relations, with a variation of Lesk measure applied on definitions obtained from Wikipedia.

6. REFERENCES

[1] Philipp Cimiano and Johanna Völker, "Towards large-scale, open-domain and ontology-based named entity classification," in *RANLP*, 2005, pp. 166–172.

[2] Michael Fleischman and Eduard Hovy, "Fine grained classification of named entities," in *COLING*. 2002, ACL.

[3] Tomáš Kliegr, Krishna Chandramouli, Jan Nemrava, Vojtěch Svátek, and Ebroul Izquierdo, "Combining captions and visual analysis for image concept classification," in *MDM/KDD'08*. 2008, ACM, To appear.

[4] Krishna Chandramouli, Tomáš Kliegr, Jan Nemrava, Vojtěch Svátek, and Ebroul Isquierdo, "Query refinement and user relevance feedback for contextualized image retrieval," in *VIE 08*, 2008, To appear.

[5] Tomáš Kliegr, Krishna Chandramouli, Jan Nemrava, Vojtěch Svátek, and Ebroul Izquierdo, "Wikipedia as the premiere source for targeted hypernym discovery," in *WBBT/ECML'08*, 2008, To appear.

[6] Jim Cowie, Joe Guthrie, and Louise Guthrie, "Lexical disambiguation using simulated annealing," in *COLING*, Morristown, NJ, USA, 1992, pp. 359–365, ACM.

[4] *Text-only*: concept with the highest confidence was selected as the image label; *image-only (KAA)*:the class associated with segment with the highest ratio between the area of the segmented region and the whole image [3].