Named Entity Extraction and Disambiguation: The Reinforcement Effect

Mena B. Habib
Faculty of EEMCS, University of Twente
Enschede, The Netherlands
m.b.habib@ewi.utwente.nl

Maurice van Keulen
Faculty of EEMCS, University of Twente
Enschede, The Netherlands
m.vankeulen@ewi.utwente.nl

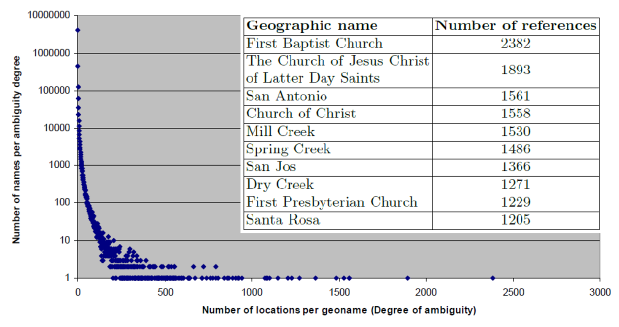
ABSTRACT

Named entity extraction and disambiguation have received much attention in recent years. Typical fields addressing these topics are information retrieval, natural language processing, and semantic web. Although these topics are highly dependent, almost no existing works examine this dependency. It is the aim of this paper to examine the dependency and show how one affects the other, and vice versa. We conducted experiments with a set of descriptions of holiday homes with the aim to extract and disambiguate *toponyms* as a representative example of named entities. We experimented with three approaches for disambiguation with the purpose to infer the country of the holiday home. We examined how the effectiveness of extraction influences the effectiveness of disambiguation, and reciprocally, how filtering out ambiguous names (an activity that depends on the disambiguation process) improves the effectiveness of extraction. Since this, in turn, may improve the effectiveness of disambiguation again, it shows that extraction and disambiguation may reinforce each other.

1. INTRODUCTION

In natural language, *toponyms*, i.e., names for locations, are used to refer to these locations without having to mention the actual geographic coordinates. The process of *toponym extraction* (a.k.a. toponym recognition) is a subtask of information extraction that aims to identify location names in natural text. This process has become a basic step of many systems for Information Extraction (*IE*), Information Retrieval (*IR*), Question Answering (*QA*), and in systems combining these, such as [1].

Toponym disambiguation (a.k.a. toponym resolution) is the task of determining which real location is referred to by a certain instance of a name. Toponyms, as with named entities in general, are highly ambiguous. For example, according to GeoNames,[1] the toponym "Paris" refers to more than sixty different geographic places around the world besides the capital of France. Figure 1 shows the top ten of the most ambiguous geographic names. It also shows the long tail distribution of toponym ambiguity. From this figure, it can be observed that around 46% of toponyms have two or more, 35% three or more, and 29% four or more references.



Geographic name	Number of references
First Baptist Church	2382
The Church of Jesus Christ of Latter Day Saints	1893
San Antonio	1561
Church of Christ	1558
Mill Creek	1530
Spring Creek	1486
San Jos	1366
Dry Creek	1271
First Presbyterian Church	1229
Santa Rosa	1205

Figure 1: Toponym ambiguity in GeoNames: top-10 and long tail.

In natural language, humans rely on the *context* disambiguate a toponym. Note that in human communication, the context used for disambiguation is broad: not only the surrounding text matters, but also the author and recipient, their background knowledge, the activity they are currently involved in, even the information the author has about the background knowledge of the recipient, and much more.

Although entity extraction and disambiguation are highly dependent, almost all efforts focus on improving the effectiveness of either one but not both. Hence, almost none examine their interdependency. It is the aim of this paper to examine exactly this. We studied not only the positive and the negative effect of the extraction process on the disambiguation process, but also the potential of using the result of disambiguation to improve extraction. We call this potential for mutual improvement, the *reinforcement effect* (see Figure 2).

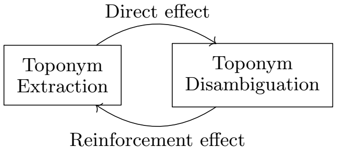


Figure 2: The reinforcement effect between the toponym extraction and disambiguation processes.

[1] www.geonames.org

To examine the reinforcement effect, we conducted experiments on a collection of holiday home descriptions from the Eurocottage[2] portal. These descriptions contain general information about the holiday home including its location and its neighborhood (See Figure 5 for an example).

The task we focus on is to extract toponyms from the description and use them to infer the country where the holiday property is located. We use country inference as a way to disambiguate the extracted toponyms. A set of heuristics have been developed to extract toponyms from the text. Three different approaches for toponym disambiguation are compared. We investigate how the effectiveness of disambiguation is affected by the effectiveness of extraction by comparing with results based on manually extracted toponyms. We investigate the reverse measuring the effectiveness of extraction when filtering out those toponyms found to be highly ambiguous, and in turn, measure the effectiveness of disambiguation based on this filtered set of toponyms.

The rest of the paper is organized as follows. Section 2 presents related work on named entity extraction and disambiguation. The approaches we used for toponym extraction and disambiguation are described in Section 3. In Section 4, we describe the experimental setup, present its results, and discuss some observations and their consequences. Finally, conclusions and future work are presented in Section 5.

2. RELATED WORK

Named entity extraction (NEE) and disambiguation (NED) are two areas of research that are well-covered in literature. Many approaches were developed for each. NEE research focuses on improving the precision and recall of extracting all entity names from unstructured natural text. NED research focuses on improving the precision and recall of the entities these names refer to. As mentioned earlier, we focus on toponyms as a subcategory of named entities. Is this section, we briefly survey a few major approaches for toponym extraction and disambiguation.

NEE is a subtask of IE that aims to annotate phrases in text with its entity type such as names (e.g., person, organization or location name), or numeric expressions (e.g., time, date, money or percentage). The term 'named entity recognition (extraction)' was first mentioned in 1996 at the Sixth Message Understanding Conference (MUC-6) [2], however the field started much earlier. The vast majority of proposed approaches for NEE fall in two categories: handmade rule-based systems and supervised learning-based systems.

One of the earliest rule-based system is FASTUS [3]. It is a nondeterministic finite state automaton text understanding system used for IE. In the first stage of its processing, names and other fixed form expressions are recognized by employing specialized microgrammars for short, multi-word fixed phrases and proper names. Another approach for NEE is matching against pre-specified gazetteers such as done in LaSIE [4, 5]. It looks for single and multi-word matches in multiple domain-specific full name (locations, organizations, etc.) and keyword lists (company designators, person first names, etc.). It supports hand-coded grammar rules that make use of part of speech tags, semantic tags added in the gazetteer lookup stage, and if necessary the lexical items themselves.

The idea behind supervised learning is to discover discriminative features of named entities by applying machine learning on positive and negative examples taken from large collections of annotated texts. The aim is to automatically generate rules that recognize instances of a certain category entity type based on their features. Supervised learning techniques applied in NEE include Hidden Markov Models [6], Decision Trees [7], Maximum Entropy Models [8], Support Vector Machines [9], and Conditional Random Fields [10].

According to [11], there are different kinds of toponym ambiguity. One type is structural ambiguity, where the structure of the tokens forming the name are ambiguous (e.g., is the word "Lake" part of the toponym "Lake Como" or not?). Another type of ambiguity is semantic ambiguity, where the type of the entity being referred to is ambiguous (e.g., is "Paris" a toponym or a girl's name?). A third form of toponym ambiguity is reference ambiguity, where it is unclear to which of several alternatives the toponym actually refers (e.g., does "London" refer to "London, UK" or to "London, Ontario, Canada"?). In this paper, we focus on reference ambiguity.

Toponym disambiguation or resolution is a form of Word Sense Disambiguation (WSD). According to [12], existing methods for toponym disambiguation can be classified into three categories: (i) map-based: methods that use an explicit representation of places on a map; (ii) knowledge-based: methods that use external knowledge sources such as gazetteers, ontologies, or Wikipedia; and (iii) data-driven or supervised: methods that are based on machine learning techniques. An example of a map-based approach is [13], which aggregates all references for all toponyms in the text onto a grid with weights representing the number of times they appear. References with a distance more than two times the standard deviation away from the centroid of the name are discarded.

Knowledge based approaches are based on the hypothesis that toponyms appearing together in text are related to each other, and that this relation can be extracted from gazetteers and knowledge bases like Wikipedia. Following this hypothesis, [14] used a toponym's local linguistic context to determine the toponym type (e.g., river, mountain, city) and then filtered out irrelevant references by this type. Another example of a knowledge-based approach is [15] which uses Wikipedia to generate co-occurrence models for toponym disambiguation.

Supervised approaches use machine learning techniques for disambiguation. [16] trained a naive Bayes classifier on toponyms with disambiguating cues such as "Nashville, Tennessee" or "Springfield, Massachusetts", and tested it on texts without these clues. Similarly, [17] used Hidden Markov Models to annotate toponyms and then applied Support Vector Machines to rank possible disambiguations.

In this paper, as toponyms training examples are not available in our data set, we chose to use handcrafted rules for extraction as suggested in [18]. We used a representative example of each of the three categories for our toponym disambiguation. This is described in the following section.

[2] http://www.eurocottage.com

```
( ({Token,!Token.string==":",!Token.kind=="number",!Token.string==".",!Split})
  ( ({Token.orth == upperInitial,!Lookup.majorType=="date"})[1,2]
    ( ({Token.string == "-"})[0,1] )
    ({Token.orth == upperInitial,!Lookup.majorType=="date"})[0,2]
):Toponym )
```

Extraction Rule 1

```
( ({Split})
  ( ({Token.orth == upperInitial,!Lookup.majorType=="date"})
    ({Token.string == "-"})[0,1]
    ({Token.orth == upperInitial,!Lookup.majorType=="date"})[1,2]
):Toponym )
```

Extraction Rule 2

```
( ({Token,!Token.string==":",!Token.kind=="number",!Token.string==".",!Split})
  ( ({Token.orth == upperInitial,!Lookup.majorType=="date"})[1,2]
    ( ({Token.string == "-"})[0,1]
      | ({Token.orth == lowercase, Token.string!="and",Token.length<=3})[0,1]
    )
    ({Token.orth == upperInitial,!Lookup.majorType=="date"})[1,2]
):Toponym )
```

Extraction Rule 3

```
( ({Token.string= "(of|from|at|to|near)"})
  ( ({Token.orth == upperInitial,!Lookup.majorType=="date"})
    ({Token.string == "-"})[0,1]
    ({Token.orth == upperInitial,!Lookup.majorType=="date"})[1,2]
):Toponym )
```

Extraction Rule 4

```
 ( ( ({Token,Token.string==":"})
     | ({Token,Token.string=="."})
     | ({Split})
   )
   ( ({Token.orth == upperInitial,!Lookup.majorType=="date"})[1,2]
     ( ({Token.string == "-"})[0,1]
       | ({Token.orth == lowercase, Token.string!="and",Token.length<=3})[0,1]
     )
     ({Token.orth == upperInitial,!Lookup.majorType=="date"})[1,2]
 ):Toponym )
```

Extraction Rule 5

```
( ({Token.string= "(Ï)́"})
  ( ({Token.orth == upperInitial,!Lookup.majorType=="date"})
    ({Token.string == "-"})[0,1]
    ({Token.orth == upperInitial,!Lookup.majorType=="date"})[1,2]
  ):Toponym
  ({Token.string= "(Ï)́"})
)
```

Extraction Rule 6

Figure 3: JAPE rules for Toponym Extraction.

3. EXPERIMENTAL SETUP

3.1 Toponym extraction

3.1.1 Extraction rules

We use GATE [19] for toponym extraction. As toponym training examples are not available in our data set, we preferred to develop handcrafted rules for extraction as suggested in [18]. The rules are specified in GATE's JAPE-language. They are based on heuristics on the orthography features of tokens and other annotations. Figure 3 contains the toponym extraction rules used in our experiments.

JAPE is a Java Annotation Patterns Engine. JAPE provides nite state transduction over annotations based on regular expressions. A JAPE grammar consists of a set of phases, each of which consists of a set of pattern/action rules. The rules always have two sides: Left Hand Side (LHS) and Right Hand Side (RHS). The LHS of the rule contains the annotation pattern; it may contain regular expression operators (e.g. *, ?, +). The RHS outlines the action to be taken on the detected pattern and consists of annotation manipulation statements. Annotations matched on the LHS of a rule are referred to in the RHS by means of labels. What is shown in Figure 3 is the LHS part of our set of rules.

3.1.2 Entity matching

We use the GeoNames geographical database for entity matching. It consists of 7.5 million unique entities of which 2.8 million are populated places with in total 5.5 million alternative names. All entities are categorized into 9 classes defining the type of place (e.g., country, region, lake, city, road). Figure 4 shows the coverage of GeoNames as a map drawn by placing a point at the coordinates of each entity.

3.2 Toponym Disambiguation

We compare three approaches for toponym disambiguation, one representative example for each of the categories described in Section 2. All require the text to contain toponym annotations. Hence, disambiguation can be seen as a classification problem assigning the toponyms to their most probable country. The notation we used for describing the approaches can be found in Table 1.

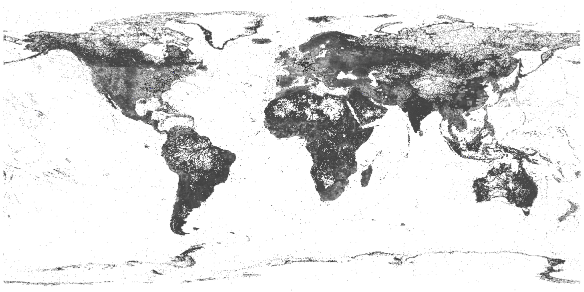


Figure 4: The world map drawn with the GeoNames longitudes and latitudes.

3.2.1 Bayes Approach

This is a supervised learning approach for toponym disambiguation based on Naive Bayes (NB) theory. NB is a probabilistic approach to text classification. It uses the joint probabilities of terms and categories to estimate the probabilities of categories given a document [20]. It is naive in the sense that it makes the assumption that all terms are conditionally independent of each other given a category. Because of this independence assumption, the parameters for each term can be learned separately which simplifies and speeds up computations compared to non-naive Bayes classifiers. Toponym disambiguation can be seen as a text classification problem where extracted toponyms are considered as terms and the country associated with the text as a class.

There are two common event models for NB text classification: the multinomial and multivariate Bernoulli model [21]. Here, we use the multinomial model as suggested by the same reference. In both models, classification of toponyms is performed by applying Bayes rule:

$$P(C = c_j \mid d_i) = \frac{P(d_i \mid c_j)P(c_j)}{P(d_i)} \tag{1}$$

D	:the set of all documents. $D = \{d_l \in D \mid l = 1 \ldots n\}$
T	:the set of toponyms appearing in the document d. $T = \{t_i \in d \mid i = 1 \ldots m\}$
G	:GeoNames gazetteer. $G = \{r_{ix} \mid r_{ix}$ is geographical location$\}$ Where i is the toponym index and x is the reference index. Each reference r_{ix} is represented by a set of characteristics: its country, longitude, latitude, and its class. r_{ix} is a reference for t_i, if t_i is string-wise equal to r_{ix} or one of its alternatives.
$R(t_i)$	:the set of references for toponym t_i. $R(t_i) = \{r_{ix} \in G \mid t_i$ is string-wise equal to r_{ix} or to one of its alternatives$\}$
R	:the set of all sets $R(t_i)$. $\forall t_i \in T$.
C_i	:the set of countries of $R(t_i)$. $C_i = \{c_{ix} \mid c_{ix}$ is the country of the reference $r_{ix}\}$

Table 1: Notation used for describing the toponym disambiguation approaches

where d_i is a test document (as a list of extracted toponyms) and c_j is a country. We assign that country c_j to d_i that has the highest $P(C = c_j \mid d_i)$, i.e., the highest posterior probability of country c_j given test document d_i. To be able to calculate $P(C = c_j \mid d_i)$, the prior probability $P(c_j)$ and the likelihood $P(d_i \mid c_j)$ have to be estimated from a training set. Note that the evidence $P(d_i)$ is the same for each country, so we can eliminate it from the computation. The prior probability for countries, $P(c_j)$, can be estimated as follows:

$$P(c_j) = \frac{\sum_{i=1}^{N} y(d_i, c_j)}{N} \qquad (2)$$

where N is the number of training documents and $y(d_i, c_j)$ is defined as:

$$y(d_i, c_j) = \begin{cases} 1 & \text{if } d_i \in c_j \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

So, the prior probability of country c_j is estimated by the fraction of documents in the training set belonging to c_j. $P(d_i \mid c_j)$ parameters are estimated using the multinomial model. In this model, a document d_i is a sequence of extracted toponyms. The Naive Bayes assumption is that the probability of each toponym is independent of its context, position, and length of the document. So, each document d_i is drawn from a multinomial distribution of toponyms with a number of independent trials equal to the length of d_i. The likelihood probability of a document d_i given its country c_j can hence be approximated as:

$$P(d_i \mid c_j) = P(t_1, t_2, \ldots, t_n \mid c_j) \approx \prod_{k=1}^{n} P(t_k \mid c_j) \qquad (4)$$

where n is the number of toponyms in document d_i, and t_k is the k^{th} toponym occurring in d_i. Thus, the estimation of $P(d_i \mid c_j)$ is reduced to estimating each $P(t_k \mid c_j)$ independently. $P(t_k \mid c_j)$ can be estimated with Laplacian smoothing:

$$P(t_k \mid c_j) = \frac{\Theta + tf_{kj}}{(\Theta \times |T|) + \sum_{l=1}^{|T|} tf_{lj}} \qquad (5)$$

where tf_{kj} is the term frequency of toponym t_k belonging to country c_j. The summation term in the denominator stands for the total number of toponym occurrences belonging to c_j. Θ in the numerator and $\Theta \times |T|$ in the denominator are used to avoid zero probabilities. Θ is set to 0.0001 according to [22].

Using this approach, all the Bayes parameters for classifying a test document to its associated country, which in a sense disambiguates its toponyms, can be estimated using a training set.

3.2.2 *Popularity Approach*

This is an unsupervised approach based on the intuition that, as each toponym in a document may refer to many alternatives, the more of those appear in a certain country, the more probable it is that the document belongs to that country. For example, it is common to find lakes, rivers or mountains with the same name as a neighboring city. We also take into consideration the GeoNames Feature Class (GFC) of the reference. As shown in Table 2, we assign a weight to each of the 9 GFCs representing its contribution to the country of the toponym, basically choosing a higher weight for cities, populated places, regions, etc. We define the *popularity* of a country c for a certain document d to be the average over all toponyms of d of the sum of the weights of the references of those toponyms in c:

$$Pop_d(c) = \frac{1}{|d|} \sum_{t_i \in d} \sum_{r_{ix} \in R(t_i) \rceil c} wgfc(r_{ix}) \qquad (6)$$

where $R(t_i)\rceil c = \{r_{ix} \in R(t_i) \mid c_{ix} = c\}$ is the restriction of the set of references $R(t_i)$ to those in country c, and *wgfc* is the weight of the GeoName Feature Class as specified in Table 2. For disambiguating the country of a document, we choose the country with the highest popularity.

GeoName Feature Classes (GFC)	**Weight** *wgfc*
Administrative Boundary Features	3
Hydrographic Features	1
Area Features	1
Populated Place Features	3
Road / Railroad Features	1
Spot Features	1
Hypsographic Features	1
Undersea Features	1
Vegetation Features	1

Table 2: The feature classes of GeoNames along with the weights we use for each class

3.2.3 *Clustering Approach*

This is an unsupervised approach based on the assumption that toponyms appearing in same document are likely to refer to locations close to each other *distance-wise*. For each toponym, we have, in general, multiple alternatives. By taking one alternative for each toponym, we form a cluster. A cluster, hence, is a possible combination of alternatives, or in other words, one possible interpretation of the toponyms in the text. In this approach, we consider all possible clusters, compute the average distance between the alternative locations in the cluster, and choose the cluster $Cluster_{min}$ with the lowest average distance.

$$Clusters = \{\{r_{1x}, r_{2x}, \ldots, r_{mx}\} \mid \forall t_i \in d \bullet r_{ix} \in R(t_i)\} \quad (7)$$

$$Cluster_{min} = \underset{Cluster_k \in Clusters}{\arg\min} \text{ average distance of } Cluster_k \quad (8)$$

For disambiguating the country of the document, we choose the most often occurring country in $Cluster_{min}$.

4. EXPERIMENTAL RESULTS

In this section, we present the results of experiments with the presented methods of extraction and disambiguation applied on a collection of holiday properties descriptions. The goal of the experiments is to investigate the influence of extraction effectiveness on disambiguation effectiveness and vice versa, and ultimately to show that they can *reinforce* each other.

4.1 Data Set

The data set we use for our experiments is a collection of traveling agent holiday properties descriptions from the Eurocottage[3] portal. The descriptions not only contain information about the property itself and its facilities, but also a description of its location, neighboring cities and opportunities for sightseeing. The data set includes the country of each property which we use to validate our results. Figure 5 shows an example for a holiday property description.

> Bargecchia 9 km from Massarosa: nice, rustic house ”I Cipressi”, renovated in 2000, in the center of Bargecchia 11 km from the center of Viareggio, 29 km from the center of Lucca, in a central, quiet, sunny position on a slope. Private, terrace (60 m2), garden furniture, barbecue. Steep motor access to the house. Parking in the grounds. Grocers, restaurant, bar 100 m, sandy beach 11 km. Please note: car essential.
> 3-room house 90 m2 on 2 levels, comfortable and modern furnishings: living/dining room with 1 double sofa bed, open fireplace, dining table and TV, exit to the terrace. Kitchenette (oven, dishwasher, freezer). Shower/bidet/WC. Upper floor: 1 double bedroom. 1 room with 1 x 2 bunk beds, exit to the balcony. Bath/bidet/WC. Gas heating (extra). Small balcony. Terrace 60 m2. Terrace furniture, barbecue. Lovely panoramic view of the sea, the lake and the valley. Facilities: washing machine. Reserved parking space n 2 fenced by the house. Please note: only 1 dog accepted.

Figure 5: An example of a EuroCottage holiday home description.

The data set consists of 29707 property descriptions. This set has been partitioned into a *training set* of 26610 descriptions for the Bayes supervised approach, and a *test set* containing the remaining 3097 descriptions. The *annotation test set* is a subset of the test set containing 1579 descriptions for which we constructed a ground truth by manually annotating all toponyms.

It turned out, however, that not all manually annotated toponyms had a match in the GeoNames database. For example, we annotated phrases like “Columbus Park” as a toponym, but no entry for this toponym in GeoNames exists. Therefore, we constructed, besides this *full ground truth*, also a *matching ground truth* where all non-matching annotations have been removed.

4.2 Experiment 1: Initial effectiveness of extraction

The objective of the first set of experiments is to evaluate the initial effectiveness of the extraction rules in terms of precision and recall.

Table 3 contains the precision and recall of the extraction rules on the annotation test set evaluated against both ground truths. As expected, recall is higher with the matching ground truth, because there are less toponyms to find, and precision is lower, because more of the extracted toponyms are not in the matching ground truth.

Ground truth	**Precision**	**Recall**
Full ground truth	72%	78%
Matching ground truth	51%	80%

Table 3: Effectiveness of the extraction rules

4.3 Experiment 2: Initial effectiveness of disambiguation

The second set of experiments aims to evaluate the initial effectiveness of the proposed disambiguation approaches and its sensitivity to the effectiveness of the extraction process.

The top part of Table 4 contains the precision of country disambiguation, i.e., the percentage of correctly inferred countries using the automatically annotated toponyms. As expected, the supervised approach performs better than both unsupervised approaches.

The bottom part of Table 4 aims at showing the influence of the imprecision of the extraction process on the disambiguation process. We compare the results of using the automatically extracted toponyms with using the (better quality) manually annotated toponyms. Since we only have manual annotations for the annotation test set and not for the training set, we have no results for the Bayes approach. Even though the annotation test set is smaller, we can observe that the results for the automatically extracted toponyms are very similar to those of the full test set, hence we assume that our conclusions are also valid for the test set. We can conclude that both unsupervised approaches signicantly benefit from better quality toponyms.

	Bayes approach	**Popularity approach**	**Clustering approach**
On full test set			
Automatically extracted toponyms	94.2%	65.45%	78.19%
On annotation test set			
Automatically extracted toponyms	-	65.4%	78.95%
Manually annotated toponyms	-	75.6%	86%

Table 4: Precision of country disambiguation

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

4.4 Experiment 3: The reinforcement effect

Examining the results of disambiguation, we discovered that there were many false positives among the automatically extracted toponyms, i.e. words extracted as a toponym and having a reference in GeoNames, that are in fact no toponyms. A sample of such words is shown in Figure 6.

access attention beach breakfast chalet
cottage double during floor garden
golf holiday haus kitchen market
olympic panorama resort satellite shops
spring thermal villa village wireless
world you

Figure 6: A sample of false positives among extracted toponyms.

These words affect the disambiguation result, because the matching entries in GeoNames belong to many different countries.

A possible improvement for the extraction process, hence, is filtering out extracted toponyms that match GeoNames entries belonging to too many countries. The intuition is that these toponyms, whether they are actual toponyms in reality or not, confuse the disambiguation process. We set the threshold to five, i.e., words referring to more than five countries in GeoNames are filtered out from the extracted toponyms. In this way, 197 toponyms were filtered out.

Note that we used the *result of disambiguation* for an improvement of *extraction*. Therefore, this is an example of the 'Reinforcement effect' in Figure 2.

To evaluate the effect of this improvement, we repeated the experiments but now while using the filtered set of automatically extracted toponyms. Tables 5 and 6 present the repetition of the first and second experiment, respectively.

Comparing Tables 5 and 3, we can observe, albeit relatively small, some improvement in the effectiveness of extraction by filtering out the 'confusing' words. Nevertheless, if we compare Tables 6 and 4, we observe a significant improvement for the subsequent disambiguation. Note that the precision is now very close to the precision of using manually annotated toponyms.

This shows that the idea of multiple iterations of extraction and disambiguation may reinforce each other. In the next section, we explore this idea somewhat further by presenting observations from deeper analysis and discussing possible ways of exploiting the reinforcement effect.

Ground truth	**Precision**	**Recall**
Full ground truth	74%	77%
Matching ground truth	52%	79%

Table 5: Effectiveness of the extraction rules with filtering.

	Popularity approach	**Clustering approach**
On annotation test set		
Filtered automatically extracted toponyms	73.5%	84.1%

Table 6: Precision of country disambiguation with filtering.

4.5 Further analysis and discussion

From further analysis of results and causes, we like to mention the following observations and thoughts.

4.5.1 Ambiguous toponyms

The improvement described above was based on filtering out toponyms that have alternatives in five or more countries. The intuition was that these terms ordinarily do not constitute toponyms but general terms that happen to be common topological names as well, such as those of Figure 6. In total, 197 extracted toponyms were filtered out in this way. We have observed, however, that some of these were in fact true toponyms, for example, "Amsterdam", "France", and "Sweden". Apparently, these toponyms appear in more than five countries. We believe, however, that filtering them out, had a positive effect anyway as they were harming the disambiguation process.

4.5.2 Multi-token toponyms

Sometimes the structure of the terms constituting a toponym in the text is ambiguous. For example, for "Lake Como" it is dubious whether or not "Lake" is part of the toponym or not. In fact, it depends on the conventions of the gazetteer which choice produces the best results. Furthermore, some toponyms have a rare structure, such as "Lido degli Estensi". The extraction rules of Figure 3 failed to extract this as one toponym and instead produced two toponyms: "Lido" and "Estensi" with harmful consequences for the holiday home country disambiguation.

4.5.3 All-or-nothing

Related to this, we can observe that entity extraction is ordinarily an all-or-nothing activity: one can only annotate either "Lake Como" or "Como", but not both.

4.5.4 Near-border ambiguity

We also observed problems with near-border holiday homes, because their descriptions often mention places across the border. For example, the description in Figure 7 has 4 toponyms in The Netherlands, 5 in Germany and 1 in the UK, whereas the holiday home itself is in The Netherlands and not in Germany. Even if an approach like the clustering approach is succesful in correctly interpreting the toponyms themselves, it may still assign the wrong country.

4.5.5 Non-expressive toponyms

Finally, we observed many properties with no or non-expressive toponyms, such as "North Sea". In such cases, it remains hard and error prone to correctly disambiguate the country of the holiday home.

4.5.6 Proposed new approach based on uncertain annotations

We believe that many of the observed problems are caused by an improper treatment of the inherent ambiguities. Natural language has the innate property that it is multiply interpretable. Therefore, none of the processes in information extraction should be 'all-or-nothing'. In other words, all steps, including entity recognition, should produce *possible* alternatives with associated likelihoods and depedencies (see Figure 8). Multiple iterations of recognition, matching, and disambiguation are then aimed at adjusting likelihoods and expanding or reducing alternatives (see Figure 9). Scalable

This charming holiday home is in a small holiday park in the village of **Nutter**^{NL}. The village is in the province of **Overijssel**^{NL}. The holiday home is comfortably furnished and equipped with every modern convenience.
The home is furnished in an **English**^{UK} style and has a romantic atmosphere. You can relax on the veranda in the evenings and enjoy delightful views of the orchard. The surrounding area has much to offer.
There are plenty of excellent walking and cycling routes. Interesting towns such as **Ootmarsum**^{NL} and **Almelo**^{NL} are well worth a visit. Children will enjoy the **German**^{GER} Animal Park in **Nordhorn**^{GER}. If you're prepared to travel a little further afield, you can reach the **Apfelkorn Distillery**^{GER} in **Haselüne**^{GER} in **Germany**^{GER}, in around one hour. It's not to be missed.

Figure 7: Example holiday home description illustrating the vulnerability of the clustering approach for near-border holiday homes. '**T**^{C}' depicts a toponym T in country C.

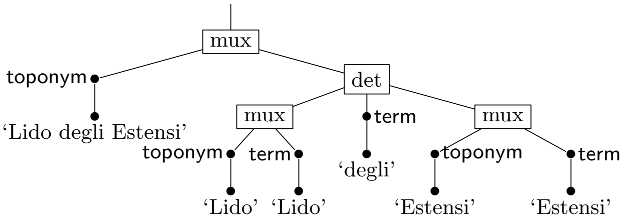


Figure 8: Probabilistic XML fragment representing all possible interpretations of the token sequence "Lido degli Estensi" (Notation from [23])

solutions for managing huge volumes of 'uncertain' annotations can be found in probabilistic relational (See, e.g., [24, 25]) and probabilistic XML (See, e.g., [23, 26]) databases.

As we have shown in this paper, steps in the information extraction process can reinforce each other. With 'uncertain alternatives', reinforcement techniques such as refining extraction rules, establishing lists of exceptional cases, or even learning rules, can be more gradual and refined. One can imagine, for example, that it can be *automatically and gradually learned* that "Lake Como" is more likely to be the best naming convention rather than "Como", or that "degli" may connect two terms into one toponym, or that for country disambiguation, what threshold to use for the number of alternative countries above which such toponyms start to harm the disambiguation process.

In this way, the entire process becomes more robust against ambiguous situations and can gradually learn. In other words, we believe there is much potential in making the inherent uncertainty in information extraction explicit.

5. CONCLUSION AND FUTURE WORK

Named entity extraction and disambiguation are highly dependent processes. The aim of this paper is to examine this dependency and show how one affects the other, and vice versa. Experiments were conducted with a set of descriptions of holiday homes with the aim to extract and disambiguate toponyms as a representative example of named entities. Three approaches for disambiguation were applied with the purpose to infer the country of the holiday home from the description. We examined how the effectiveness of extraction influences the effectiveness of disambiguation, and reciprocally, how the result of disambiguation can be used to improve extraction. As an example of the latter we filtered out toponyms that were discovered to be highly ambiguous. Results showed that the effectiveness of extraction and, in turn, disambiguation improved, thereby showing that both can reinforce each other. We also analyzed the results more closely and formulated a general approach based on *uncertain annotation* for which we argue that it has much potential for making information extraction more robust against ambiguous situations and allowing it to gradually learn.

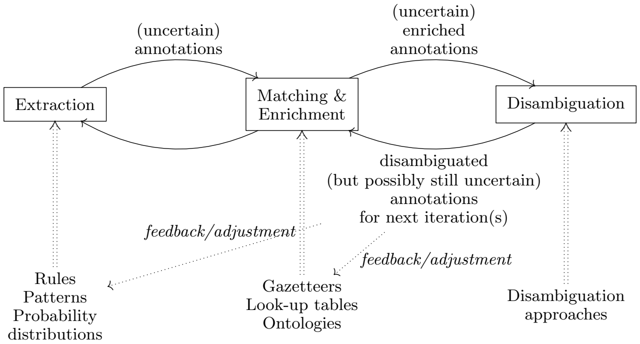


Figure 9: Activities and propagation of uncertainty

For future work, we plan to investigate the abovementioned potential. We also plan to examine statistical techniques for extraction, matching, and disambiguation as they seem to fit well in such an approach based on uncertain annotations.

6. REFERENCES

[1] M.B. Habib. Neogeography: The challenge of channelling large and ill-behaved data streams. In *Workshops Proc. of the 27th IEEE Int'l Conf. on Data Engineering (ICDE 2011)*, pages 284–287, 2011.

[2] R. Grishman and B. Sundheim. Message understanding conference - 6: A brief history. In *Proc. of Int'l Conf. on Computational Linguistics*, pages 466–471, 1996.

[3] J.R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson. Fastus: A system for extracting information from text. In *Proc. of Human Language Technology*, pages 133–137, 1993.

[4] R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks. University of Sheffield: Description of the LaSIE system as used for MUC-6. In *Proc. of the 6th Conf. on Message understanding (MUC-6)*, pages 207–220, 1995.

[5] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of Sheffield: Description of the Lasie-II system as used for MUC-7. In *Proc. of the 7th Conf. on Message Understanding (MUC-7)*, 1998.

[6] G. Zhou and J. Su. Named entity recognition using an hmm-based chunk tagger. In *Proc. of the 40th Ann.*

Meeting of the Association for Computational Linguistics, pages 473–480, 2002.

[7] S. Sekine. NYU: Description of the Japanese NE system used for MET-2. In *Proc. of the 7th Conf. on Message Understanding MUC-7*, 1998.

[8] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. NYU: Description of the MENE named entity system as used in MUC-7. In *Proc. of the 7th Conf. on Message Understanding (MUC-7)*, 1998.

[9] H. Isozaki and H. Kazawa. Efficient support vector classifiers for named entity recognition. In *Proc. of the 19th Int'l Conf. on Computational Linguistics (COLING 2002)*, pages 1–7, 2002.

[10] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proc. of 7th Conf. on Natural Language Learning (CoNLL 2003)*, pages 188–191, 2003.

[11] N. Wacholder, Y. Ravin, and M. Choi. Disambiguation of proper names in text. In *Proc. of the 5th Conf. on Applied Natural Language Processing (ANLC 1997)*, pages 202–208, 1997.

[12] D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *Int'l Journal of Geographical Information Science*, 22(3):301–313, 2008.

[13] D. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries*, volume 2163 of *LNCS*, pages 127–136, 2001.

[14] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Proc. of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 50–54, 2003.

[15] J.M.S. Overell and S. Ruger. Place disambiguation with co-occurrence models. In *Proc. of the Working Notes of the Cross Language Evaluation Forum Workshop (CLEF 2006)*, 2006.

[16] D.A. Smith and G.S. Mann. Bootstrapping toponym classifiers. In *Proc. of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 45–49, 2003.

[17] B. Martins, I. Anastácio, and P. Calado. A machine learning approach for resolving place references in text. In *Proc. of the 13th AGILE Int'l Conf. on Geographic Information Science*. Springer, 2010.

[18] S. Sekine and C. Nobata. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proc. of Conf. on Language Resources and Evaluation (LREC 2004)*, pages 1977–1980, 2004.

[19] H. Cunningham. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[20] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[21] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *Proc. of the AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48, 1998.

[22] S.-B. Kim, K.-S. Han, H.-C. Rim, and S.H. Myaeng. Some effective techniques for naive bayes text classification. *IEEE Trans. on Knowledge and Data Engineering*, 18:1457–1466, 2006. ISSN 1041-4347.

[23] S. Abiteboul, B. Kimelfeld, Y. Sagiv, and P. Senellart. On the expressiveness of probabilistic xml models. *The VLDB Journal*, 18(5):1041–1064, 2009.

[24] M. Mutsuzaki, M. Theobald, A. de Keijzer, J. Widom, P. Agrawal, O. Benjelloun, A. Das Sarma, R. Murthy, and T. Sugihara. Trio-One: Layering uncertainty and lineage on a conventional DBMS (demo). In *3rd Biennial Conf. on Innovative Data Systems Research (CIDR 2007)*, pages 269–274, 2007.

[25] L. Antova, T. Jansen, C. Koch, and D. Olteanu. Fast and simple relational processing of uncertain data. In *Proc. of the 24th Int'l Conf. on Data Engineering (ICDE 2008)*, pages 983–992. IEEE, April 2008.

[26] M. van Keulen and A. de Keijzer. Qualitative effects of knowledge rules and user feedback in probabilistic data integration. *The VLDB Journal*, 18(5):1191–1217, 2009.