Improving Toponym Extraction and Disambiguation Using Feedback Loop

Mena B. Habib and Maurice van Keulen

Faculty of EEMCS, University of Twente, Enschede, The Netherlands
{m.b.habib,m.vankeulen}@ewi.utwente.nl

Abstract. This paper addresses two problems with toponym extraction and disambiguation. First, almost no existing works examine the extraction and disambiguation interdependency. Second, existing disambiguation techniques mostly take as input extracted toponyms without considering the uncertainty and imperfection of the extraction process.

It is the aim of this paper to investigate both avenues and to show that explicit handling of the uncertainty of annotation has much potential for making both extraction and disambiguation more robust.

1 Introduction

Toponyms are names used to refer to locations without having to mention the actual geographic coordinates. The process of *toponym extraction* aims to identify location names in natural text.

Toponym disambiguation 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 GeoNames¹, the toponym “Paris” refers to more than sixty different geographic places around the world besides the capital of France. Another source of ambiguousness is that some toponyms are common English words.

A general principle in our work is our conviction that toponyms extraction and disambiguation are highly dependent. In previous work [1], we studied not only the positive and negative effect of the extraction process on the disambiguation process, but also the potential of using the result of disambiguation to improve extraction. We called this potential for mutual improvement, the *reinforcement effect* (see Figure 1).

In general, we concluded 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,

Fig. 1. The reinforcement effect between the toponym extraction and disambiguation processes

¹ www.geonames.org

all steps, including entity recognition, should produce *possible* alternatives with associated likelihoods and dependencies.

Our Contributions. In this paper, we focus on this principle. We turned to statistical approaches for toponym extraction. The advantage of statistical techniques for extraction is that they provide alternatives for annotations along with confidence probabilities. The probabilities proved to be useful in enhancing the disambiguation process. We believe that there is much potential in making the inherent uncertainty in information extraction explicit in this way.

Furthermore, extraction models are inherently imperfect and generate imprecise confidences. We were able to use the disambiguation result for increasing the confidence of true toponyms and reducing the confidence of false positives. This enhancement of extraction improves as a consequence the disambiguation (the aforementioned reinforcement effect). This process can be repeated iteratively, without any human interference, as long as there is improvement in the extraction and disambiguation.

2 Our Approach

The task we focus on is to extract toponyms from EuroCottage holiday home descriptions² and use them to infer the country where the holiday property is located. We use this country inference task as a representative example of disambiguating extracted toponyms.

We propose an entity extraction and disambiguation approach based on uncertain annotations. The general approach illustrated in Figure 2 has the following steps:

1. Prepare training data by manually annotating named entities.
2. Use the training data to build a statistical extraction model.
3. Apply the extraction model on test data and training data.
4. Match the extracted named entities against one or more gazetteers.
5. Use the toponym entity candidates for the disambiguation process.
6. Evaluate the extraction and disambiguation results for the training data.
Use a list of highly ambiguous named entities and false positives that affect the disambiguation results to re-train the extraction model.

Fig. 2. General approach

² www.eurocottage.com

7. The steps from 2 to 6 are repeated automatically until there is no improvement any more in either the extraction or the disambiguation.

Toponym Extraction. For toponym extraction, we developed two statistical named entity extraction modules³, one based on Hidden Markov Models (HMM) and one based on Conditional Random Fields (CRF).

The goal of HMM [2] is to find the optimal tag sequence (in our case, whether the word is assigned to toponym tag or not) $T = t_1, t_2, t_3, \dots, t_n$ for a given word sequence $W = w_1, w_2, w_3, \dots, w_n$ that maximizes $P(T | W)$.

Conditional Random Fields (CRF) can model overlapping, non-independent features [3]. Here we used a linear chain CRF, the simplest model of CRF.

Extraction Modes of Operation. We used the extraction models to retrieve sets of annotations in two ways:

- **First-Best:** In this method, we only consider the first most likely set of annotations that maximize the probability $P(T | W)$ for the whole text. This method does not assign a probability for each individual annotation, but only to the whole retrieved set of annotations.
- **N-Best:** This method returns a top-25 of possible alternative hypotheses for terms annotations in order of their estimated likelihoods $p(t_i | w_i)$. The confidence scores are assumed to be conditional probabilities of the annotation given an input token.

Toponym Disambiguation. For the toponym disambiguation task, we only select those toponyms annotated by the extraction models that match a reference in GeoNames. We furthermore use an adapted version of the clustering approach of [1] to disambiguate to which entity an extracted toponym actually refers.

Handling Uncertainty of Annotations. Instead of giving equal contribution to all toponyms, we take the uncertainty in the extraction process into account to include the confidence of the extracted toponyms. In this way terms which are more likely to be toponyms have a higher contribution in determining the country of the document than less likely ones.

Improving Certainty of Extraction. In despite of the abovementioned improvement, the extraction probabilities are not accurate and reliable all the time. Some extraction models retrieve some false positive toponyms with high confidence probabilities. This is where we take advantage of the reinforcement effect. To be more precise. We introduce another class in the extraction model called ‘highly ambiguous’ and annotate those terms in the training set with this class that the disambiguation process finds more than τ countries for documents that contain this term. The extraction model is subsequently re-trained and the whole process is repeated without any human interference as long as there is improvement in extraction and disambiguation process for the training set. The intention is that the extraction model learns to avoid prediction of terms to be toponyms when they appear to confuse the disambiguation process.

³ We made use of the *lingpipe* toolkit for development: <http://alias-i.com/lingpipe>

3 Experimental Results

Here we present the results of experiments with the presented methods of extraction and disambiguation applied to a collection of holiday properties descriptions. The data set consists of 1579 property descriptions for which we constructed a ground truth by manually annotating all toponyms.

Experiment 1: Effect of Extraction with Confidence Probabilities. Table 1 shows the percentage of holiday home descriptions for which the correct country was successfully inferred. We can see that the **N-Best** method outperforms the **First-Best** method for both HMM and CRF models. This supports our claim that dealing with alternatives along with their confidences yields better results.

Table 1. Effectiveness of the disambiguation process for First-Best and N-Best methods in the extraction phase

	HMM	CRF
First-Best	62.59%	62.84%
N-Best	68.95%	68.19%

Table 2. Effectiveness of the disambiguation after iteration of refinement

	HMM	CRF
No Filtering	68.95%	68.19%
1st Iteration	73.28%	68.44%

Table 3. Effectiveness of the extraction process after iteration of refinement

	HMM		
	Pre.	Rec.	F1
No Filtering	0.3584	0.8517	0.5045
1st Iteration	0.7667	0.5987	0.6724

	CRF		
	Pre.	Rec.	F1
No Filtering	0.6969	0.7136	0.7051
1st Iteration	0.6989	0.7131	0.7059

Experiment 2: Effect of Extraction Certainty Enhancement. Tables 2 and 3 show the effectiveness of the disambiguation and the extraction processes respectively before and after one iteration of refinement. We can see an improvement in HMM extraction and disambiguation results. The initial HMM results showed a high recall rate with a low precision. In spite of this, our approach managed to improve precision through iteration of refinement. The refinement process is based on removing highly ambiguous toponyms resulting in a slight decrease in recall and an increase in precision. In contrast, CRF started with high precision which could not be improved by the refinement process.

4 Conclusion and Future Work

Named entity extraction and disambiguation are inherently imperfect processes that moreover depend on each other. The aim of this paper is to examine and make use of this dependency for the purpose of improving the disambiguation by iteratively enhancing the effectiveness of extraction, and vice versa.

We examined how handling the uncertainty of extraction influences the effectiveness of disambiguation, and reciprocally, how the result of disambiguation can be used to improve the effectiveness of extraction. The extraction models are automatically retrained after discovering highly ambiguous false positives among the extracted toponyms. This process improves the precision of the extraction.

References

1. Habib, M.B., van Keulen, M.: Named entity extraction and disambiguation: The reinforcement effect. In: Proc. of MUD 2011, Seattle, USA, pp. 9–16 (2011)
2. Ekbal, A., Bandyopadhyay, S.: A Hidden Markov Model Based Named Entity Recognition System: Bengali and Hindi as Case Studies. In: Ghosh, A., De, R.K., Pal, S.K. (eds.) PReMI 2007. LNCS, vol. 4815, pp. 545–552. Springer, Heidelberg (2007)
3. Wallach, H.: Conditional random fields: An introduction. Technical Report MS-CIS-04-21, University of Pennsylvania (2004)