A Hybrid Approach for Robust Multilingual Toponym Extraction and Disambiguation

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. Toponym extraction and disambiguation are key topics recently addressed by fields of Information Extraction and Geographical Information Retrieval. Toponym extraction and disambiguation are highly dependent processes. Not only toponym extraction effectiveness affects disambiguation, but also disambiguation results may help improving extraction accuracy. In this paper we propose a hybrid toponym extraction approach based on Hidden Markov Models (HMM) and Support Vector Machines (SVM). Hidden Markov Model is used for extraction with high recall and low precision. Then SVM is used to find false positives based on informativeness features and coherence features derived from the disambiguation results. Experimental results conducted with a set of descriptions of holiday homes with the aim to extract and disambiguate toponyms showed that the proposed approach outperform the state of the art methods of extraction and also proved to be robust. Robustness is proved on three aspects: language independence, high and low HMM threshold settings, and limited training data.

1 Introduction

Toponyms are names used to refer to locations without having to mention the actual geographic coordinates. The process of toponym extraction (recognition) is a subset of Named Entity Recognition (NER) that aims to identify location name boundaries in text. While toponym disambiguation (resolution) is the process of mapping between a toponym and an unambiguous spatial coordinates of the same place.

Toponyms extraction and disambiguation are highly challenging. 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. Around 46% of toponyms in GeoNames have more than one reference. Duplicate geographic names comes from the fact that emigrant settlers prefer to use their original land names to denote their new homes, leading to referential ambiguity of place names [1]. Another source of ambiguity is that some common English words have references in GeoNames and might be extracted as toponyms under some conditions. For example, words like {Shop, Park, Villa, Airport} represent location names in GeoNames.

A general principle in this work is our conviction that toponym extraction and disambiguation are highly dependent [2]. Mena et al. [3] studied not only

[1] www.geonames.org

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. They called this potential for mutual improvement, the *reinforcement effect*.

The extraction techniques fall into two categories: machine learning and rule-based approaches. The advantage of statistical techniques for extraction is that they provide alternatives for annotations along with confidence probabilities. Instead of discarding these, as is commonly done by selecting the top-most likely candidate, we use them to enrich the knowledge for disambiguation. It was proved that extraction probability can be used to enhance the disambiguation so that the contribution of each extracted item to the disambiguation of other extracted items is proportional to its extraction probability [3]. We believe that there is much potential in making the inherent uncertainty in information extraction explicit in this way. Certainty can also be improved using informativeness features and coherence features derived from the disambiguation results.

Most of existing extraction techniques are language-dependent as they need a POS tagger. And it is known that it takes some effort to tune the thresholds and that they are typically trained on large corpuses. In practice, one would like to have more robustness so that accuracy is not easily hampered. In this paper, we specifically address robustness against threshold settings, situations with other languages, and situations with limited training data.

In this paper we propose a hybrid extraction approach based on Hidden Markov Models (HMM) and Support Vector Machines (SVM). An initial HMM is trained and used for extraction. We used a low cutting threshold to achieve high recall resulting in low precision. A clustering based approach for disambiguation is then applied. A set of coherence features are extracted for the extracted toponyms based on the disambiguation results feedback and also on informativeness measures (like Inverse Document Frequency and Gain). A SVM is then trained with the extracted features to classify the HMM extracted toponyms into true positives and false positives resulting in improving the precision and hence the F1 measure. Our hybrid approach outperforms the Conditional Random Fields (CRF), the state of the art method of extraction and Stanford NER, the prominent Named Entity Recognition System. Furthermore, our hybrid approach is shown to be language independent as all the used methods are not based on language dependent techniques like Part Of Speech (POS) which is commonly used with the NER systems. Robustness of the proposed approach is experimentally proved by applying different HMM cutting thresholds, evaluating it across multiple languages and also with smaller training sets. More aspects of robustness like evaluating across multiple domains and using different types of named entities are left for future work.

To examine our hybrid approach, 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 2 for an example). As a representative example of

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

toponym extraction and disambiguation, we focused on the task of extracting toponyms from the description and using them to infer the country where the holiday property is located.

Contributions: We can summarize our contributions as follows: (1) We propose a hybrid toponym extraction approach based on HMM and SVM. (2) The proposed system is proved to be robust against three aspects: different languages, different cutting thresholds, and limited training data. (3) We introduce some features (informativeness and coherence-based) that can be used to enhance the process of toponym extraction.

The rest of the paper is organized as follows. Section 2 presents related work on toponym extraction and disambiguation. Our proposed approach for toponym extraction and disambiguation is 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.

2 Related work

Toponym extraction and disambiguation are special cases of a more general problem called Named Entity Recognition (NER) and Disambiguation (NED). In this section, we briefly survey a few major approaches for NER and toponym disambiguation.

2.1 Named entity extraction

NER is a subtask of Information Extraction (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) [4], however the field started much earlier. The vast majority of proposed approaches for NER fall in two categories: handmade rule-based systems and supervised learning-based systems.

One of the earliest rule-based system is FASTUS [5]. It is a nondeterministic finite state automaton text understanding system used for IE. The other category of NER systems is the machine learning based systems. Supervised learning techniques applied in NEE include Hidden Markov Models (HMM) [6], Decision Trees [7], Maximum Entropy Models [8], Support Vector Machines [9], and Conditional Random Fields (CRF) [10][11].

Multilingual NER is discussed by many researchers. Florian et al. [12] used classifier-combination experimental framework for multilingual NER in which four diverse classifiers are combined under different conditions. Szarvas et al. [13] introduced a multilingual NER system by applying AdaBoostM1 and the C4.5 decision tree learning algorithm. Richman and Schone utilized the multilingual characteristics of Wikipedia to annotate a large corpus of text with NER tags [14]. Similarly, Nothman et al. [15] automatically created multilingual training annotations for NER by exploiting the text and structure of parallel Wikipedia articles in different languages.

Using informativeness features in NER is introduced by Rennie et al. [16]. They conducted a study on identifying restaurant names from posts to a restaurant discussion board. They found the informativeness scores to be an effective restaurant word filter. Furche et al. [17] introduce a system called AMBER for extracting data from an entire domain. AMBER employs domain specific gazetteers to discern basic domain attributes on a web page, and leverages repeated occurrences of similar attributes to group related attributes into records.

Some researches focused only on toponym extraction. In [18], a method for toponym recognition is presented that is tuned for streaming news by leveraging a wide variety of recognition components, both rule-based and statistical. Another interesting toponym extraction work was done by Pouliquen et al. [19]. They present a multilingual method to recognize geographical references in free text that uses minimum of language-dependent resources, except a gazetteer. In this system, place names are identified exclusively through gazetteer lookup procedures and subsequent disambiguation or elimination.

2.2 Toponym disambiguation

Toponym reference disambiguation or resolution is a form of Word Sense Disambiguation (WSD). According to [20], 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 [21], 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, [22] 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.

Supervised learning approaches use machine learning techniques for disambiguation. [23] trained a naive Bayes classifier on toponyms with disambiguating clues and tested it on texts without these clues. Similarly, [24] used Support Vector Machines to rank possible disambiguations.

3 Proposed hybrid approach

The hybridness of our proposed approach can be viewed from two points of view. It can be viewed as a hybrid approach of toponym extraction and disambiguation processes. Clues derived from the disambiguation results are used to enhance extraction. Also our system can be viewed as a hybrid machine learning approach for extraction where HMM and SVM are combined to achieve better results. An initial HMM is trained and used for extraction with high recall. A SVM is then

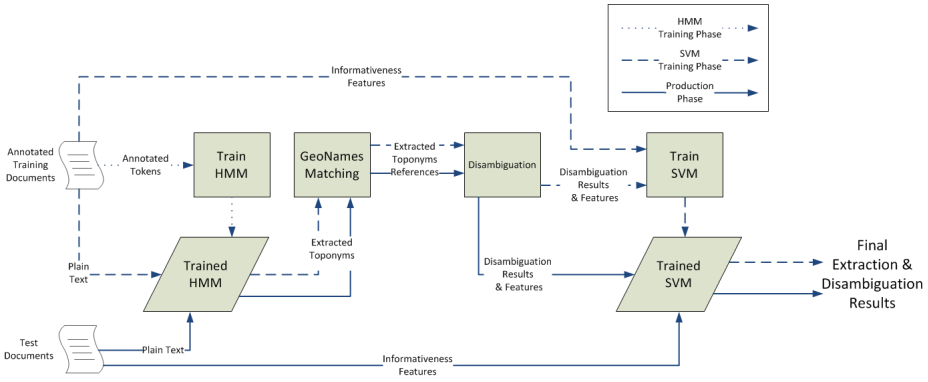


Fig. 1. Our proposed hybrid toponym extraction and disambiguation approach.

trained to classify the HMM extracted toponyms into true positives and false positives resulting in improving the precision and hence the F1 measure.

3.1 System phases

The system illustrated in Figure 1 has the following Phases:

Phase 1: HMM training

1. Training data is prepared by manually annotating all toponyms. Tokens are tagged, following the CoNLL[3] standards, by either a LOCATION or O tag which represents words that are not part of a location phrase.
2. Training data is used to train a HMM[4][5] [25] for toponym extraction. The advantage of statistical techniques for extraction is that they provide alternatives for annotations accompanied with confidence probabilities. Instead of discarding these, as is commonly done by selecting the top-most likely candidate, we use them to enrich the knowledge for disambiguation. The probabilities proved to be useful in enhancing the disambiguation process [3].

Phase 2: SVM training

1. The trained HMM is then used to extract toponyms from the training set. A low cutting threshold is used to get high recall. The extracted toponyms are then matched against GeoNames gazeteer. For each toponym, a list of candidate references are fed to the disambiguation process.
2. The disambiguation process tries to find only one representative reference for each extracted toponym based on its coherency with other toponyms mentioned in the same document. Details of the disambiguation approach used is described in section 3.2.

[3] http://www.cnts.ua.ac.be/conll2002/ner/
[4] http://alias-i.com/lingpipe/
[5] We used an HmmCharLmEstimator which employs a maximum a posteriori transition estimator and a bounded character language model emission estimator.

3. Two sets of features (informativeness and coherence-based) are computed for each extracted toponym. Details of the selected features are described in section 3.3.
4. The extracted set of features are used to train the SVM classifier [6][7] to distinguish between true positives toponyms and false positives ones.

Phase 3: Production

1. The trained HMM is applied on the test set. The extracted toponyms are matched against GeoNames and their candidate references are disambiguated. Informativeness and coherence features are computed and fed to the trained SVM to find the final results of toponyms extraction process.
2. Disambiguation process can be repeated using the final set of extracted toponyms to get the improvement reflected on the disambiguation results.

The main intuition behind our approach is to make use of more clues than those often used by traditional extraction techniques (like POS, word shape, preceding and succeeding words). We deliberately use set of language-independent features to ensure robustness across multiple languages. To make use of those features we start with high recall and then filter the extracted toponyms based on those features. Even by using a higher cutting threshold, our approach is still able to enhance the precision at the expense of some recall resulting in enhancement of the overall F1 measure. Moreover, the features are found to be highly discriminative, so that only few training samples are required to train the SVM classifier good enough to make correct decisions.

3.2 Toponym disambiguation approach

For the toponym disambiguation task, we only select those toponyms annotated by the extraction models that match a reference in GeoNames. We use the clustering approach of [3] with the purpose to infer the country of the holiday home from the description. The clustering approach is an unsupervised disambiguation approach based on the assumption that toponyms appearing in same document are likely to refer to locations close to each other distance-wise. For our holiday home descriptions, it appears quite safe to assume this. For each toponym t_i, we have, in general, multiple entity candidates. Let $R(t_i) = \{r_{ix} \in \text{GeoNames gazetteer}\}$ be the set of reference candidates for toponym t_i. Additionally each reference r_{ix} in GeoNames belongs to a country $Country_j$. By taking one entity candidate for each toponym, we form a cluster. A cluster, hence, is a possible combination of entity candidates, or in other words, one possible entity candidate of the toponyms in the text. In this approach, we consider all possible clusters, compute the average distance between the candidate locations in the cluster, and choose the cluster $Cluster_{min}$ with the lowest average distance. We choose the most often occurring country $Country_{winner}$ in $Cluster_{min}$ for disambiguating the country of the document. In effect the abovementioned assumption states that the entities that belong to $Cluster_{min}$ are the

[6] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
[7] We used C-support vector classification (C-SVC) type of SVM with RBF kernel.

true representative entities for the corresponding toponyms as they appeared in the text.

3.3 Selected features

Coherence features derived from disambiguation results along with informativeness features are computed for all the extracted toponyms generated by the HMM. For each extracted toponym the following set of informativeness features are computed:

1. **Inverse Document Frequency (IDF)**: IDF is an informativeness score that embodies the principle that the more frequent a word is, the lower the chance it is a relevant toponym. The IDF score for an extracted toponym t is:
$$IDF = -log\frac{d_t}{D}$$
where d_t is the document frequency of the toponym t, and D is the total number of documents.
2. **Residual Inverse Document Frequency (RIDF)**: RIDF is an extension of IDF that has proven effective for NER [16]. RIDF is calculated as the difference between the IDF of a toponym and its expected IDF according to the poisson model. The RIDF score can be calculated by the formula:
$$expIDF = -log(1 - e^{-f_t/D})\ RIDF = IDF - expIDF$$
where f_t is the frequency of the toponym across all documents D.
3. **Gain**: Gain is a feature that can be used to identify "important" or informative terms. For a toponym t, Gain is derived as:
$$Gain(t) = \frac{d_t}{D}(\frac{d_t}{D} - 1 - log\frac{d_t}{D})$$
4. **Extraction Confidence (EC)**: Extraction confidence (probability) is the HMM conditional probability of the annotation given an input word. The goal of HMM is to find the optimal tag sequence $T = t_1, t_2, ..., t_n$ for a given word sequence $W = w_1, w_2, ..., w_n$ that maximizes:
$$P(T \mid W) = \frac{P(T)P(W|T)}{P(W)}$$
The prior probability $P(t_i|t_{i-2}, t_{i-1})$ and the likelihood probability $P(w_i|t_i)$ can be estimated from training data. The optimal sequence of tags can be efficiently found using the Viterbi dynamic programming algorithm [26]. The extraction confidence is the probability of being a part of toponym given a token $P(t|w)$.

Furthermore, the following set of coherence features are computed based on the disambiguation results:

1. **Distance (D)**: The distance feature is the kilo-metric distance between the coordinates of the selected candidate reference r_{ij} for toponym t_i and the coordinates of the inferred country $Country_{winner}$.
$$Distance = Coordinates(r_{ij}) - Coordinates(Country_{winner})$$
2. **Standard Score (SS)**: It is calculated by dividing the distance between the coordinates of the r_{ix} and $Country_{winner}$ over the standard deviation of all selected references distances to $Country_{winner}$.
$$StandardScore = \frac{Coordinates(r_{ij}) - Coordinates(Country_{winner})}{\sigma}$$

1-room apartment 80 m2, on the ground floor, simple furnishings: living/dining room 70 m2 with 4 beds and satellite-TV. Open kitchen (4 hotplates, oven, micro wave) with dining table. Shower/WC. Floor heating. Facilities: hair dryer. Internet (Dial up/ISDN). The room is separated by 4 steps in bedroom and lounge. The bedroom has no direct light.
Olšova Vrata 5 km from **Karlovy Vary**: On the edge of the **Slavkovsky** les nature reserve. Small holiday hamlet next to the hotel which has been a popular destination for **Karlsbad** inhabitants for the past 30 years new, large house with 2 apartments, 2 storeys, built in 2004, surrounded by trees, above **Karlovy Vary**, in a secluded, sunny position, 10 m from the woods edge. Private, patio (20 m2), garden furniture. In the house: table-tennis. Central heating. Breakfast and half-board on request. Motor access to the house (in winter snow chains necessary). Parking by the house. Shop 4 km, grocers 1.5 km, restaurant 150 m, bus stop 550 m, swimming pool 6 km, indoor swimming pool 6 km, thermal baths 6 km, tennis 1 km, golf course 1.5 km, skisport facilities 25 km. Please note: car essential. Airport 1.5 km (2 planes/day). On request: Spa treatments, green fee. Ski resort **Klinovec**, 20 km.

Fig. 2. An example of a EuroCottage holiday home description (toponyms in bold).

3. **Number of GeoNames candidate references (#Geo)**: It is simply the number of candidate references for the toponym ti.
$$\#\text{GeoNames Refs} = |r_{ix}|$$
4. **Belongingness to the disambiguated country (Bel)**: Indicates whether or not r_{ij} belongs to $Country_{winner}$.
$$\text{Belongingness to } Country_{winner} = \begin{cases} 1 & \text{if } Country(r_{ij}) = Country_{winner} \\ 0 & \text{otherwise} \end{cases}$$

Informativeness features tend to find those false positives that appear multiple times across the collection. Those highly repeated words are more likely to be false positives toponyms. On the other hand, some false positives appear only rarely in the collection. Those toponyms can not be caught by informativeness features. Here where we make use of coherence-based features. Coherence features tend to find those false positives that are not coherent with other toponyms. The usage of a combination of both sets of features maximizes the extraction effectiveness (F1 measure).

Unlike traditional features commonly used with NER systems like (POS), all our selected features are language independent and thus our approach can be applied to any language as the GeoNames gazetteer has representations for toponyms in different languages. Furthermore we avoid using word shape features as languages like German require the capitalization of all nouns making capitalization a useless feature to extract NE.

4 Experimental results

In this section, we present the results of experiments with the proposed approach applied to a collection of holiday properties descriptions. The goals of the experiments are to compare our approach with the state of the art approaches and systems and to show its robustness in terms of language independence, high and low HMM threshold settings, and limited training data.

4.1 Data set

The data set we use for our experiments is a collection of traveling agent holiday property descriptions from the EuroCottage[8] 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. Descriptions are also available in German and Dutch. Some of these descriptions are direct translations and some others have independent descriptions of the same holiday cottage. The data set includes the country of each property which we use to validate our results. Figure 2 shows a representative example of a holiday property description. The manually annotated toponyms are written in bold. The data set consists of 1181 property descriptions for which we constructed a ground truth by manually annotating all toponyms for only the English version. The German and the Dutch versions of descriptions are annotated automatically by matching them against all toponyms that appear in the English version or their translations. For example "Cologne" in the English version is translated to "Köln" and matched in the German version and translated to "Keulen" and matched in the Dutch version. Although this method is not 100% reliable due to slight differences in translated versions, we believe that it is reliable enough as ground truth for showing the language independency of our approach.

We split the data set into a training set and a validation test set with ratio 2 : 1. We used the training set for training the HMM extraction model and the SVM classifier, and the test set for evaluating the extraction and disambiguation effectiveness for "new and unseen" data.

4.2 Experiment 1: Data set analysis

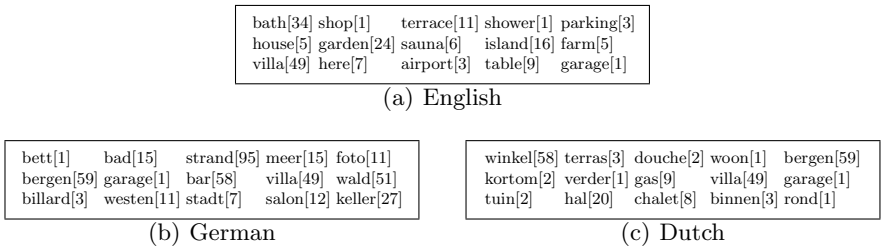
The aim of this experiment is to show some statistics about the test set in all versions through different phases of our system pipeline. Table 1 shows the number of toponyms per property description [*#Top./Doc.*], the number of toponyms per property that have references in GeoNames [*#Top./Doc. ∈GeoNames*], and the average degree of ambiguity per toponyms [*Degree of ambiguity*] (i.e the average number of references in GeoNames for a given toponym). *Ground Truth* represents manual annotations statistics. *HMM(0.1)* represents statistics of the extracted toponyms resulting from applying HMM on the test set with cutting probability threshold 0.1, while *HMM(0.1)+SVM* represents statistics of the extracted toponyms resulting from applying SVM after HMM on the test set.

As can be observed from table 1 that HMM extracts many false positives. Examples of those false positives that have references in GeoNames are shown in figure 3[9].

It can also be noticed that the English version contains more toponyms per property description. Our method of automatically annotating the German and the Dutch texts misses a few annotations. This doesn't harm the evaluation process of the proposed method as our approach works on improving the precision

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

[9] We match the extracted toponyms against names of places, their ascii representation and their alternative representations in GeoNames gazeteer.



(a) English

(b) German

(c) Dutch

Fig. 3. Examples of false positives (toponyms erroneously extracted by HMM(0.1)) and their number of references in GeoNames.

with some loss in recall. Hence, we can claim that precision/recall/F1 measures of our proposed approach applied on German and Dutch versions shown on the section 4.4 can be regarded as a lower bound.

Table 1. Test set statistics through different phases of our system pipeline.

	#Top./Doc.			**#Top./Doc. ∈GeoNames**			**Degree of ambiguity**		
	EN	DE	NL	EN	DE	NL	EN	DE	NL
Ground Truth	5.04	4.62	3.51	3.47	3.10	2.46	7.24	6.15	6.78
HMM(0.1)	12.02	11.31	11.38	6.51	5.72	5.85	8.69	9.27	10.33
HMM(0.1)+SVM	5.24	5.04	3.91	3.59	3.18	2.58	8.43	7.38	7.78

4.3 Experiment 2: SVM features analysis

In this experiment we evaluate the selected set of features used for SVM training on the English collection. We want to show the effect of these features on the effectivness of the SVM classifier. The aim of the SVM is to find the false positives toponyms among those extracted by the HMM. Two groups of features are used. Informativness features and coherence features (features derived from disambiguation results). Table 2 shows:

- Extraction and disambiguation results using each of the features individually to train the SVM classifier.
- Information Gain [*IG*] for each feature. *IG* measures the amount of information in bits about the class prediction (in our case true positive toponym or false positive).
- The extraction and disambiguation results using each group of features (Informativeness (*Inf*) and coherence (*Coh*)) and using both combined (*All*).
- Extraction and disambiguation results for only *HMM* with threshold 0.1 (prior to the usage of the SVM).
- Disambiguation results using manually annotated toponyms (*Ground Truth*).

Extraction results are evaluated in terms of precision [*Pre.*], recall [*Rec.*] and [*F1*] measures, while disambiguation results [*Dis.*] are evaluated in terms of the percentage of holiday home descriptions for which the correct country was inferred.

The coherence features can be only calculated for toponyms that belong to GeoNames. This implies that its effect only appears on false positives that belong to GeoNames. To make their effect more clear, we presented two sets of results:

- *All extracted toponyms*: where all toponyms are used to train HMM and SVM regardless of whether they exist in GeoNames or not. Evaluation is done for all extracted toponyms.
- *Only toponyms* $\in$ *GeoNames*: where only toponyms existing in GeoNames are used to train and evaluate HMM and SVM.

By looking at $[IG]$ of each feature we can observe that the $[Bel]$, $[IDF]$ and $[EC]$ are highly discriminative features, while $[\#Geo]$ seems to be a bad feature as it has no effect at all on the SVM output.

Using manually annotated toponyms for disambiguation, the best possible input one would think, may not produce the best possible disambiguation result. For example, the disambiguation result of HMM(0.1)+SVM(Gain) is higher than that of the ground truth. This is because some holiday cottages are located on the border with other country, so that description mentions cities from other country rather than the country of the cottage. This does not mean that the correct representative candidates for toponyms are missed. Moreover, since our disambiguation result is based on voting, we attribute this effect to chance: the NER may produce a false positive toponym which happens to sway the vote to the correct country, in other words, there are cases of correct results for the wrong reasons.

It can be also observed that low recall leads to poor disambiguation results. That is because low recall may result in extracting no toponyms from the property description and hence the country of that property is misclassified.

Table 2 shows how using the SVM classifier enhances the extraction and the disambiguation results. The effect of combining both set of features is more clear in the results of $[$*Only toponyms* $\in$ *GeoNames*$]$. Precision is improved significantly, and hence the F1 measure, by using the coherence features beside the informativeness ones.

Table 3 shows the extracted toponyms for the property shown in figure 2 using different methods. Informativeness features tend to find those false positives that appear multiple times across the collection like {In, Shop}. On the other hand, disambiguation features tend to find those false positives that are not coherent with other toponyms like {Airport}. The usage of a combination of both sets of features maximizes the extraction effectiveness (F1 measure).

4.4 Experiment 3: Multilinguality, different thresolding robustness and competitors

In this experiment we want to show the multiligualitiy and system robustness across different languages and against different threshold settings. Multilinguality is guaranteed by our approach as we only use language independent methods of extraction and filtering. We effectively avoided using Part-Of-Speech (POS) as feature since it is highly language-dependent and for many languages there are no good automatic POS-tagger available. Table 4 shows the effectiveness of

Table 2. Extraction and disambiguation results using different features for English version.

	All extracted toponyms				
	IG	**Pre.**	**Rec.**	**F1**	**Dis.**
Ground Truth		**1**	**1**	**1**	79.1349
HMM(0.1)		0.3631	**0.8659**	0.5116	75.0636
HMM(0.1)+SVM(IDF)	0.1459	0.5514	0.8336	0.6637	80.4071
HMM(0.1)+SVM(RIDF)	0.1426	0.5430	0.8472	0.6618	80.4071
HMM(0.1)+SVM(Gain)	0.1013	0.5449	0.8205	0.6549	**80.9160**
HMM(0.1)+SVM(EC)	0.2223	0.7341	0.7489	0.7414	78.3715
HMM(0.1)+SVM(D)	0.0706	0.6499	0.5726	0.6088	74.5547
HMM(0.1)+SVM(SS)	0.0828	0.6815	0.5166	0.5877	68.4478
HMM(0.1)+SVM(#Geo)	0.1008	0.4800	0.6099	0.5372	71.7557
HMM(0.1)+SVM(Bel)	0.3049	**0.8106**	0.4942	0.6140	73.0280
HMM(0.1)+SVM(Inf)		0.7764	0.7756	0.7760	79.8982
HMM(0.1)+SVM(Coh)		0.8106	0.4940	0.6138	73.0280
HMM(0.1)+SVM(All)		0.7726	0.8014	**0.7867**	79.8982

	Only extracted toponyms ∈ GeoNames				
	IG	**Pre.**	**Rec.**	**F1**	**Dis.**
Ground Truth		**1**	**1**	**1**	79.1349
HMM(0.1)		0.4874	**0.9121**	0.6353	75.0636
HMM(0.1)+SVM(IDF)	0.2652	0.7612	0.8983	0.8241	**81.1705**
HMM(0.1)+SVM(RIDF)	0.2356	0.7536	0.9107	0.8247	80.9160
HMM(0.1)+SVM(Gain)	0.1754	0.6419	0.8656	0.7372	76.3359
HMM(0.1)+SVM(EC)	0.2676	0.8148	0.8243	0.8195	78.3715
HMM(0.1)+SVM(D)	0.1375	0.6563	0.8584	0.7439	77.6081
HMM(0.1)+SVM(SS)	0.1077	0.6802	0.7444	0.7108	68.4478
HMM(0.1)+SVM(#Geo)	0.0791	0.4878	0.9121	0.6356	75.0636
HMM(0.1)+SVM(Bel)	0.3813	0.8106	0.7117	0.7579	73.0280
HMM(0.1)+SVM(Inf)		0.8181	0.8823	0.8490	80.6616
HMM(0.1)+SVM(Coh)		0.8117	0.7451	0.7770	76.3359
HMM(0.1)+SVM(All)		**0.8865**	0.8453	**0.8654**	79.8982

Table 3. Extracted toponyms for the property shown in figure 2

	HMM(0.1)	**HMM(0.1) + SVM(Inf)**	**HMM(0.1) + SVM(Dis)**	**HMM(0.1) + SVM(All)**
[+]Olšova Vrata	+	+	+	+
[+]Karlovy Vary	+	+	+	+
[+]Slavkovsky	+	+	+	+
[+]Karlsbad	+	+	+	+
[+]Karlovy Vary	+	+	+	+
[+]Klinovec	+	+	+	+
[-]In	+	-	+	-
[-]Shop	+	-	+	-
[-]Airport	+	+	-	-

our proposed approach applied on English, German, and Dutch versions in terms of the F1 and the disambiguation results over the state of the art: the CRF, and the Stanford NER models [10]. CRF is considered one of the famous techniques in NER. We trained a CRF on set of features described in [3]. One of the used features is POS which we were only able to extract for the English version. Stanford is a NER system based on CRF model trained on CoNLL data collection. It

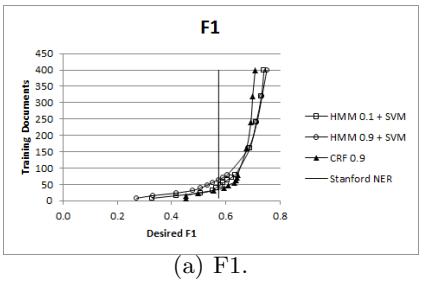
[10] http://nlp.stanford.edu/software/CRF-NER.shtml

Table 4. Extraction and disambiguation results for all versions.

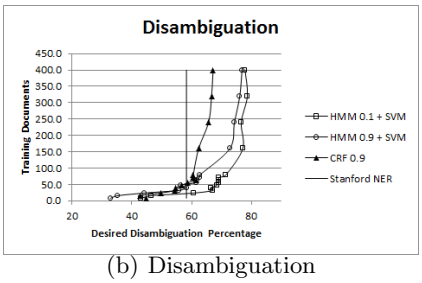
	English			
	Pre.	**Rec.**	**F1**	**Dis.**
Ground Truth	**1**	**1**	**1**	79.1349
HMM(0.1)	0.3631	**0.8659**	0.5116	75.0636
HMM(0.1)+SVM(All)	0.7726	0.8014	0.7867	**79.8982**
HMM(0.9)	0.6638	0.7806	0.7175	78.3715
HMM(0.9)+SVM(All)	0.8275	0.7591	**0.7918**	79.3893
Stanford NER	0.8375	0.4365	0.5739	58.2697
CRF(0.9)	**0.9383**	0.6205	0.7470	69.4656

	German			
	Pre.	**Rec.**	**F1**	**Dis.**
Ground Truth	**1**	**1**	**1**	**81.4249**
HMM(0.1)	0.3399	**0.8306**	0.4824	79.3893
HMM(0.1)+SVM(All)	0.6722	0.7321	0.7009	79.6438
HMM(0.9)	0.6169	0.7085	0.6595	77.8626
HMM(0.9)+SVM(All)	**0.7414**	0.6876	**0.7135**	77.3537
Stanford NER	0.5351	0.2723	0.3609	40.4580

	Dutch			
	Pre.	**Rec.**	**F1**	**Dis.**
Ground Truth	**1**	**1**	**1**	**73.0280**
HMM(0.1)	0.2505	**0.8128**	0.3830	68.4478
HMM(0.1)+SVM(All)	0.6157	0.6872	**0.6495**	70.4835
HMM(0.9)	0.4923	0.6713	0.5680	67.1756
HMM(0.9)+SVM(All)	**0.6762**	0.6197	0.6467	67.6845



(a) F1.



(b) Disambiguation

Fig. 4. The required training data required to achieve desired extraction and disambiguation results

incorporates long-distance information [11]. Stanford provides NER models for English and German. Unfortunately, we didn't find a suitable NER system for Dutch to compare with.

It can be observed that the CRF models achieve better precision at the expense of recall. Low recall sometimes leads to extracting no toponyms from the property description and hence the country of that property is misclassified. This results in a poor disambiguation results.

Table 4 also shows the robustness of our approach against different HMM thresholding settings. We used two different cutting thresholds (0.1, 0.9) for HMM. It is clear that our approach improves the precision and F1 measure on both cases.

4.5 Experiment 4: Low training data robustness

Robustness across different languages and using different cutting probability threshold is shown in the previous sections. In this section we want to prove the third aspect of robustness of our system which is its capability to work even with limited training samples. Figures 4(a) and 4(b) shows the required size of training data to achieve a desired result for F1 and disambiguation respectively (applied on the English collection). It can be observed that our approach requires low number of training data to outperform our competitors the CRF and Stanford

NER. Only 160 annotated documents are required to achieve 0.7 F1 and 75% correct disambiguation and to outperform the the CRF. Much less documents are required to outperform the CRF disambiguation results as we mentioned before that the high precision of CRF systems is accompanied by low recall leading to poor disambiguation results.

Conclusion and future work

In this paper we introduced a hybrid approach for toponym extraction and disambiguation. We used a HMM for extraction and a SVM classifier to classify the HMM output into false positive and true positive toponyms. Informativeness features beside coherence features derived from disambiguation results were used to train the SVM. Experiments were conducted with a set of holiday home descriptions with the aim to extract and disambiguate toponyms. Our system is proved to be robust on three aspects: language differences, high and low HMM threshold settings, and limited training data. It also outperforms the state of the art methods of NER.

For future research, we plan to apply and enhance our approach for other types of named entities and other domains. We claim that this approach is also robust against domain differences and can be adapted to suit any kind of named entities. To achieve this it is required to develop a mechanism to find false positives among the extracted named entities. Coherency measures can be used to find highly ambiguous named entities. We also want to estimate locations of toponyms not existing in gazetteers using other toponyms found in the textual context of the unknown toponym.

References

1. Jochen L Leidner. *Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* Universal Press, Boca Raton, FL, USA, 2008.
2. Mena B. Habib and Maurice van Keulen. Named entity extraction and disambiguation: The reinforcement effect. In *Proc. of MUD 2011, Seatle, USA*, pages 9–16, 2011.
3. Mena B. Habib and Maurice van Keulen. Improving toponym disambiguation by iteratively enhancing certainty of extraction. In *Proc. of KDIR 2012*, pages 399–410, 2012.
4. 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.
5. 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.
6. G. Zhou and J. Su. Named entity recognition using an hmm-based chunk tagger. In *Proc. ACL2002*, pages 473–480, 2002.
7. S. Sekine. NYU: Description of the Japanese NE system used for MET-2. In *Proc. of 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 MUC-7*, 1998.

9. H. Isozaki and H. Kazawa. Efficient support vector classifiers for named entity recognition. In *Proc. of 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 CoNLL 2003*, pages 188–191, 2003.
11. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of the 43nd ACL 2005*.
12. Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named entity recognition through classifier combination. In Walter Daelemans and Miles Osborne, editors, *Proc. of CoNLL-2003*, pages 168–171. Edmonton, Canada, 2003.
13. György Szarvas, Richárd Farkas, and András Kocsor. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In *Proc. of the 9th international conference on Discovery Science*, DS'06, pages 267–278, 2006.
14. Alexander E. Richman and Patrick Schone. Mining wiki resources for multilingual named entity recognition," acl'08, 2008.
15. Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 2012.
16. Jason D. M. Rennie. Using term informativeness for named entity detection. In *In Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 353–360, 2005.
17. Tim Furche, Giovanni Grasso, Giorgio Orsi, Christian Schallhart, and Cheng Wang. Automatically learning gazetteers from the deep web. In *Proc. of the 21st international conference companion on World Wide Web*, pages 341–344, 2012.
18. Michael D. Lieberman and Hanan Samet. Multifaceted toponym recognition for streaming news. In *Proc. of SIGIR'11*, pages 843–852, 2011.
19. Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Flavio Fluart, Wajdi Zaghouani, Anna Widiger, Ann charlotte Forslund, and Clive Best. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proc. of LREC-2006*, pages 53–58, 2006.
20. D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *Journal of Geographical Information Science*, 22(3):301–313, 2008.
21. 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.
22. E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Workshop Proc. of the HLT-NAACL 2003*, pages 50–54, 2003.
23. D.A. Smith and G.S. Mann. Bootstrapping toponym classifiers. In *Workshop Proc. of HLT-NAACL 2003*, pages 45–49, 2003.
24. B. Martins, I. Anastácio, and P. Calado. A machine learning approach for resolving place references in text. In *Proc. of AGILE 2010*, 2010.
25. Bob Carpenter. Character language models for chinese word segmentation and named entity recognition. In *Association for Computational Linguistics*, pages 169–172, 2006.
26. A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260 – 269, 1967.