DOWNLOAD PDF A SUPERVISED MACHINE LEARNING APPROACH TO TOPONYM DISAMBIGUATION

Chapter 1 : CiteSeerX â€” Approaches to Disambiguating Toponyms

11 â€¢ A Supervised Machine Learning Approach to Toponym Disambiguation Background This section reviews some background information related to this work, including.

Typ- ical fields addressing these topics are information retrieval, natural language processing, and semantic web. This paper addresses two problems with toponym extraction and disambiguation as a representative example of named entities. First, almost no existing works examine the extraction and disambiguation interdepen- dency. Second, existing disambiguation techniques mostly take as input extracted named entities 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. We conducted experiments with a set of holiday home descriptions with the aim to extract and disambiguate toponyms. We show that the extraction confidence probabilities are useful in enhancing the effectiveness of disambiguation. Reciprocally, retraining the extraction models with information automatically derived from the disambigua- tion results, improves the extraction models. This mutual reinforcement is shown to even have an effect after several automatic iterations. Figure 1 shows the top ten of the ing to predefined categories such as the names of per- most ambiguous geographic names. It also shows the sons, organizations, locations, expressions of times, long tail distribution of toponym ambiguity and the quantities, monetary values, percentages, etc. Named percentage of geographic names with multiple refer- entity extraction a. This process ponyms are common English words. One major type of named entities is the toponym. A Sample of English-words-like toponyms In natural language, toponyms are names used to re- fer to locations without having to mention the actual And 2 The 3 geographic coordinates. The process of toponym ex- General 3 All 3 traction a. The extraction tech- A 16 As 84 niques fall into two categories: A general principle in our work is our conviction Toponym disambiguation a. In previous work is referred to by a certain instance of a name. To- ponyms, as with named entities in general, are highly 1 www. Toponym ambiguity in GeoNames: Habib and van Keulen, , we studied not only tions along with confidence probabilities confidence the positive and negative effect of the extraction pro- for short. Instead of discarding these, as is com- cess on the disambiguation process, but also the po- monly done by selecting the top-most likely candi- tential of using the result of disambiguation to im- date, we use them to enrich the knowledge for disam- prove extraction. We called this potential for mutual biguation. The probabilities proved to be useful in en- improvement, the reinforcement effect see Figure 2. We believe that To examine there is much potential in making the inherent uncer- the reinforce- Direct effect tainty in information extraction explicit in this way. This Extraction Disambiguation restricts the negative effect of differences in naming experiments on d a collection of conventions of the gazetteer on the disambiguation Reinforcement effect holiday home process. The reinforcement ef- descriptions fect between the toponym extrac- from the Euro- tion and disambiguation processes. These descrip- Second, extraction models are inherently imper- tions contain fect and generate imprecise confidence. We were able general information about the holiday home including to use the disambiguation result to enhance the con- its location and its neighborhood See Figure 4 for an fidence of true toponyms and reduce the confidence example. As a representative example of toponym of false positives. This enhancement of extraction extraction and disambiguation, we focused on the improves as a consequence the disambiguation the task of extracting toponyms from the description and aforementioned reinforcement effect. This process using them to infer the country where the holiday can be repeated iteratively, without any human inter- property is located. Natural language has the innate property that it is multiply interpretable. Therefore, none of the processes in information ex- The rest of the paper is organized as follows. Sec- all steps, including entity recognition, should produce tion 3 presents a problem analysis and our general ap- possible alternatives with associated likelihoods and proach to iterative improvement of toponym extrac- dependencies. The adaptations we made

to toponym extrac- turned to statistical approaches for toponym extrac- tion and disambiguation techniques are described in tion. The advantage of statistical techniques for ex- Section 4. In Section 5, we describe the experimen- traction is that they provide alternatives for annota- tal setup, present its results, and discuss some obser- vations and their consequences. Finally, conclusions 2 http: NEE and NED are two areas of research that are well- Imprecision in information extraction is expected, covered in literature. Many approaches were devel- especially in unstructured text where a lot of noise ex- oped for each. NEE research focuses on improving ists. There is an increasing research interest in more the quality of recognizing entity names in unstruc- formally handling the uncertainty of the extraction tured natural text. NED research focuses on improv- process so that the answers of queries can be asso- ing the effectiveness of determining the actual entities ciated with correctness indicators. Only recently have these names refer to. As mentioned earlier, we focus information extraction and probabilistic database re- on toponyms as a subcategory of named entities. Is search been combined for this cause Gupta, Other methods extend this ap- 2. It is easy to extend probabilis- tic models like HMM and CRF to return the k high- NEE is a subtask of Information Extraction IE that est probability extractions instead of a single most aims to annotate phrases in text with its entity type likely one and store them in a probabilistic database such as names e. Managing uncertainty in tion name , or numeric expressions e. However, there is little work on main- started much earlier. The vast majority of proposed taining probabilities when the extraction is based on approaches for NEE fall in two categories: Within this context, Michelakis et al. It is a nondeterministic finite tems where the uncertainty arises due to the varying state automaton text understanding system used for precision associated with each rule by producing ac- IE. In the first stage of its processing, names and curate estimates of probabilities for the extracted an- other fixed form expressions are recognized by em- notations. They also capture the interaction between ploying specialized microgrammars for short, multi- the different rules, as well as the compositional nature word fixed phrases and proper names. Another ap- of the rules. It looks for single 2. One type is struc- etc. It supports hand-coded grammar rules that make tural ambiguity, where the structure of the tokens use of part of speech tags, semantic tags added in the forming the name are ambiguous e. The idea behind supervised learn- Another type of ambiguity is semantic ambiguity, ing is to discover discriminative features of named en- where the type of the entity being referred to is am- tities by applying machine learning on positive and biguous e. The aim is to automatically generate biguity, where it is unclear to which of several alter- rules that recognize instances of a certain category en- natives the toponym actually refers e. Sekine, , Maximum Entropy Models Borth- Toponym reference disambiguation or resolution wick et al. According to Buscaldi and Rosso, , existing Test methods for toponym disambiguation can be clas- data sified into three categories: An example of a map-based approach is Smith and including extracted alternatives Crane, , which aggregates all references for all toponyms with probabilities toponyms in the text onto a grid with weights repre- senting the number of times they appear. References Matching with a distance more than two times the standard de- highly ambiguous terms here: Following this hypothesis, Rauch et al. Another example of a knowledge-based approach is Overell and Ruger, which uses Wikipedia to Figure 3: General approach generate co-occurrence models for toponym disam- biguation. Supervised learning approaches use machine ambiguation is affected by the effectiveness of ex- learning techniques for disambiguation. Similarly, Mar- One major problem with the hand-coded gram- tins et al. We developed useful for the disambiguation process. We therefore a clustering-based approach for the toponym disam- propose an entity extraction and disambiguation ap- biguation task. This is described in Section 4. The general approach illustrated in Figure 3 has the following steps: Use the training data to build a statistical extrac- The task we focus on is to extract toponyms from Eu- tion model. Apply the extraction model on test data and train- infer the country where the holiday property is lo- ing data. Note that we explicitly allow uncertain cated. We use this country inference task as a rep- and alternative annotations with probabilities. Match the extracted named entities against one or ponyms. Our initial results from our previous work, where 5. Use the toponym entity candidates for the disam- we developed a set of hand-coded grammar rules to biguation process in our case we try to disam- extract

toponyms, showed that effectiveness of dis- biguate the country of the holiday home descrip- tion. As the relation between a word and its tag depends 6. Evaluate the extraction and disambiguation re- on the context of the word, the probability of the cur- sults for the training data and determine a list of rent word depends on the tag of the previous word and highly ambiguous named entities and false posi- the tag to be assigned to the current word. So P W T tives that affect the disambiguation results. Use can be calculated as: From test data one cannot deter- training data. The optimal sequence of tags can be mine a term to be a false positive, but only to be highly efficiently found using the Viterbi dynamic program- ambiguous. We also ization patterns. Conditional Random Fields CRF present our adaptations to enhance the disambigua- can model these overlapping, non-independent fea- tion by handling uncertainty and the imperfection in tures Wallach, This normalization factor is called the parti- tion function. The general form of a feature function is quences. It can be calculated by Markov assumption input sequence W , and the current location in the se- which states that the probability of a tag depends only quence i. An example for a feature function which pro- our holiday home descriptions, it appears quite safe duces a binary value for the current word shape is to assume this. For each toponym ti , we have, in gen- Capitalized: Additionally each reference rix 0 otherwise in GeoNames belongs to a country Country j. A cluster, hence, is a possible combination tional probability P T W. The standard parameter of entity candidates, or in other words, one possible learning approach is to compute the stochastic gradi- entity candidate of the toponyms in the text.

DOWNLOAD PDF A SUPERVISED MACHINE LEARNING APPROACH TO TOPONYM DISAMBIGUATION

Chapter 2 : Supervised learning - Wikipedia

Abstract This chapter presents a toponym disambiguation approach based on supervised machine learning. The proposed approach uses a simple hierarchical geographic relationship model to describe geographic entities and geographic relationships among them.

I went fishing for some sea bass. The bass line of the song is too weak. To people who understand English, the first sentence is using the word " bass fish ", as in the former sense above and in the second sentence, the word " bass instrument " is being used as in the latter sense below. Developing algorithms to replicate this human ability can often be a difficult task, as is further exemplified by the implicit equivocation between " bass sound " and " bass instrument ". History[edit] WSD was first formulated into as a distinct computational task during the early days of machine translation in the s, making it one of the oldest problems in computational linguistics. Warren Weaver , in his famous memorandum on translation, [1] first introduced the problem in a computational context. Early researchers understood the significance and difficulty of WSD well. In fact, Bar-Hillel used the above example to argue [2] that WSD could not be solved by "electronic computer" because of the need in general to model all world knowledge. However, since WSD systems were at the time largely rule-based and hand-coded they were prone to a knowledge acquisition bottleneck. In the s, the statistical revolution swept through computational linguistics, and WSD became a paradigm problem on which to apply supervised machine learning techniques. The s saw supervised techniques reach a plateau in accuracy, and so attention has shifted to coarser-grained senses, domain adaptation, semi-supervised and unsupervised corpus-based systems, combinations of different methods, and the return of knowledge-based systems via graph-based methods. Still, supervised systems continue to perform best. Difficulties[edit] Differences between dictionaries[edit] One problem with word sense disambiguation is deciding what the senses are. In cases like the word bass above, at least some senses are obviously different. In other cases, however, the different senses can be closely related one meaning being a metaphorical or metonymic extension of another , and in such cases division of words into senses becomes much more difficult. Different dictionaries and thesauruses will provide different divisions of words into senses. One solution some researchers have used is to choose a particular dictionary, and just use its set of senses. Generally, however, research results using broad distinctions in senses have been much better than those using narrow ones. WordNet is a computational lexicon that encodes concepts as synonym sets e. And the question whether these tasks should be kept together or decoupled is still not unanimously resolved, but recently scientists incline to test these things separately e. It is instructive to compare the word sense disambiguation problem with the problem of part-of-speech tagging. Both involve disambiguating or tagging with words, be it with senses or parts of speech. However, algorithms used for one do not tend to work well for the other, mainly because the part of speech of a word is primarily determined by the immediately adjacent one to three words, whereas the sense of a word may be determined by words further away. These figures are typical for English, and may be very different from those for other languages. Inter-judge variance[edit] Another problem is inter-judge variance. WSD systems are normally tested by having their results on a task compared against those of a human. However, while it is relatively easy to assign parts of speech to text, training people to tag senses is far more difficult. Moreover, humans do not agree on the task at hand â€” give a list of senses and sentences, and humans will not always agree on which word belongs in which sense. Human performance, however, is much better on coarse-grained than fine-grained distinctions, so this again is why research on coarse-grained distinctions [10] [11] has been put to test in recent WSD evaluation exercises. This linguistic issue is called pragmatics. For example, comparing these two sentences: To properly identify senses of words one must know common sense facts. Also, completely different algorithms might be required by different applications. In machine translation, the problem takes the form of target word selection. In information retrieval, a sense inventory is not necessarily required, because it is enough to know that a word is used in the same sense in the query and a retrieved

document; what sense that is, is unimportant. Discreteness of senses[edit] Finally, the very notion of " word sense " is slippery and controversial. Most people can agree in distinctions at the coarse-grained homograph level e. It does not divide up easily into distinct or discrete sub-meanings. The art of lexicography is to generalize from the corpus to definitions that evoke and explain the full range of meaning of a word, making it seem like words are well-behaved semantically. However, it is not at all clear if these same meaning distinctions are applicable in computational applications , as the decisions of lexicographers are usually driven by other considerations. Recently, a task â€" named lexical substitution â€" has been proposed as a possible solution to the sense discreteness problem. Approaches and methods[edit] As in all natural language processing , there are two main approaches to WSD â€" deep approaches and shallow approaches. Deep approaches presume access to a comprehensive body of world knowledge. Knowledge, such as "you can go fishing for a type of fish, but not for low frequency sounds" and "songs have low frequency sounds as parts, but not types of fish", is then used to determine in which sense the word bass is used. These approaches are not very successful in practice, mainly because such a body of knowledge does not exist in a computer-readable format, outside very limited domains. They just consider the surrounding words, using information such as "if bass has words sea or fishing nearby, it probably is in the fish sense; if bass has the words music or song nearby, it is probably in the music sense. However, it can be confused by sentences like The dogs bark at the tree which contains the word bark near both tree and dogs. There are four conventional approaches to WSD: Dictionary - and knowledge-based methods: These rely primarily on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence. Semi-supervised or minimally supervised methods: These make use of a secondary source of knowledge such as a small annotated corpus as seed data in a bootstrapping process, or a word-aligned bilingual corpus. These make use of sense-annotated corpora to train from. These eschew almost completely external information and work directly from raw unannotated corpora. These methods are also known under the name of word sense discrimination. Almost all these approaches normally work by defining a window of n content words around each word to be disambiguated in the corpus, and statistically analyzing those n surrounding words. In recent research, kernel-based methods such as support vector machines have shown superior performance in supervised learning. Graph-based approaches have also gained much attention from the research community, and currently achieve performance close to the state of the art. Dictionary- and knowledge-based methods[edit] The Lesk algorithm [19] is the seminal dictionary-based method. It is based on the hypothesis that words used together in text are related to each other and that the relation can be observed in the definitions of the words and their senses. Two or more words are disambiguated by finding the pair of dictionary senses with the greatest word overlap in their dictionary definitions. For example, when disambiguating the words in "pine cone", the definitions of the appropriate senses both include the words evergreen and tree at least in one dictionary. A similar approach [20] searches for the shortest path between two words: Finally, the first word is disambiguated by selecting the semantic variant which minimizes the distance from the first to the second word. An alternative to the use of the definitions is to consider general word-sense relatedness and to compute the semantic similarity of each pair of word senses based on a given lexical knowledge base such as WordNet. Graph-based methods reminiscent of spreading activation research of the early days of AI research have been applied with some success. More complex graph-based approaches have been shown to perform almost as well as supervised methods [21] or even outperforming them on specific domains. Supervised methods[edit] Supervised methods are based on the assumption that the context can provide enough evidence on its own to disambiguate words hence, common sense and reasoning are deemed unnecessary. Probably every machine learning algorithm going has been applied to WSD, including associated techniques such as feature selection , parameter optimization , and ensemble learning. Support Vector Machines and memory-based learning have been shown to be the most successful approaches, to date, probably because they can cope with the high-dimensionality of the feature space. However, these supervised methods are subject to a new knowledge acquisition bottleneck since they rely on substantial amounts of manually sense-tagged corpora for training,

which are laborious and expensive to create. Semi-supervised methods[edit] Because of the lack of training data, many word sense disambiguation algorithms use semi-supervised learning , which allows both labeled and unlabeled data. The Yarowsky algorithm was an early example of such an algorithm. From observation, words tend to exhibit only one sense in most given discourse and in a given collocation. The bootstrapping approach starts from a small amount of seed data for each word: The seeds are used to train an initial classifier , using any supervised method. This classifier is then used on the untagged portion of the corpus to extract a larger training set, in which only the most confident classifications are included. The process repeats, each new classifier being trained on a successively larger training corpus, until the whole corpus is consumed, or until a given maximum number of iterations is reached. Other semi-supervised techniques use large quantities of untagged corpora to provide co-occurrence information that supplements the tagged corpora. These techniques have the potential to help in the adaptation of supervised models to different domains. Also, an ambiguous word in one language is often translated into different words in a second language depending on the sense of the word. Word-aligned bilingual corpora have been used to infer cross-lingual sense distinctions, a kind of semi-supervised system. Word sense induction Unsupervised learning is the greatest challenge for WSD researchers. The underlying assumption is that similar senses occur in similar contexts, and thus senses can be induced from text by clustering word occurrences using some measure of similarity of context, [26] a task referred to as word sense induction or discrimination. Performance has been lower than for the other methods described above, but comparisons are difficult since senses induced must be mapped to a known dictionary of word senses. If a mapping to a set of dictionary senses is not desired, cluster-based evaluations including measures of entropy and purity can be performed. Alternatively, word sense induction methods can be tested and compared within an application. For instance, it has been shown that word sense induction improves Web search result clustering by increasing the quality of result clusters and the degree diversification of result lists.

Chapter 3 : Word-sense disambiguation - Wikipedia

Supervised Machine Learning. The majority of practical machine learning uses supervised learning. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Steps[edit] In order to solve a given problem of supervised learning, one has to perform the following steps: Determine the type of training examples. Before doing anything else, the user should decide what kind of data is to be used as a training set. In case of handwriting analysis, for example, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting. Gather a training set. The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements. Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector , which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality ; but should contain enough information to accurately predict the output. Determine the structure of the learned function and corresponding learning algorithm. For example, the engineer may choose to use support vector machines or decision trees. Run the learning algorithm on the gathered training set. Some supervised learning algorithms require the user to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset called a validation set of the training set, or via cross-validation. Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set. Algorithm choice[edit] A wide range of supervised learning algorithms are available, each with its strengths and weaknesses. There is no single learning algorithm that works best on all supervised learning problems see the No free lunch theorem. There are four major issues to consider in supervised learning: Bias-variance dilemma A first issue is the tradeoff between bias and variance. The prediction error of a learned classifier is related to the sum of the bias and the variance of the learning algorithm. A learning algorithm with low bias must be "flexible" so that it can fit the data well. But if the learning algorithm is too flexible, it will fit each training data set differently, and hence have high variance. Function complexity and amount of training data[edit] The second issue is the amount of training data available relative to the complexity of the "true" function classifier or regression function. If the true function is simple, then an "inflexible" learning algorithm with high bias and low variance will be able to learn it from a small amount of data. But if the true function is highly complex e. Dimensionality of the input space[edit] A third issue is the dimensionality of the input space. If the input feature vectors have very high dimension, the learning problem can be difficult even if the true function only depends on a small number of those features. This is because the many "extra" dimensions can confuse the learning algorithm and cause it to have high variance. Hence, high input dimensionality typically requires tuning the classifier to have low variance and high bias. In practice, if the engineer can manually remove irrelevant features from the input data, this is likely to improve the accuracy of the learned function. In addition, there are many algorithms for feature selection that seek to identify the relevant features and discard the irrelevant ones. This is an instance of the more general strategy of dimensionality reduction , which seeks to map the input data into a lower-dimensional space prior to running the supervised learning algorithm. Noise in the output values[edit] A fourth issue is the degree of noise in the desired output values the supervisory target variables. If the desired output values are often incorrect because of human error or sensor errors , then the learning algorithm should not attempt to find a function that exactly matches the training examples. Attempting to fit the data too carefully leads to overfitting. You can overfit even when there are no measurement errors stochastic noise if the function you are trying to learn is too complex for your learning model. In such a situation, the part of the

target function that cannot be modeled "corrupts" your training data - this phenomenon has been called deterministic noise. When either type of noise is present, it is better to go with a higher bias, lower variance estimator. In practice, there are several approaches to alleviate noise in the output values such as early stopping to prevent overfitting as well as detecting and removing the noisy training examples prior to training the supervised learning algorithm. There are several algorithms that identify noisy training examples and removing the suspected noisy training examples prior to training has decreased generalization error with statistical significance. Heterogeneity of the data. If the feature vectors include features of many different kinds discrete, discrete ordered, counts, continuous values , some algorithms are easier to apply than others. Many algorithms, including Support Vector Machines , linear regression , logistic regression , neural networks , and nearest neighbor methods , require that the input features be numerical and scaled to similar ranges e. Methods that employ a distance function, such as nearest neighbor methods and support vector machines with Gaussian kernels , are particularly sensitive to this. An advantage of decision trees is that they easily handle heterogeneous data. Redundancy in the data. If the input features contain redundant information e. These problems can often be solved by imposing some form of regularization. Presence of interactions and non-linearities. If each of the features makes an independent contribution to the output, then algorithms based on linear functions e. However, if there are complex interactions among features, then algorithms such as decision trees and neural networks work better, because they are specifically designed to discover these interactions. Linear methods can also be applied, but the engineer must manually specify the interactions when using them. When considering a new application, the engineer can compare multiple learning algorithms and experimentally determine which one works best on the problem at hand see cross validation. Tuning the performance of a learning algorithm can be very time-consuming. Given fixed resources, it is often better to spend more time collecting additional training data and more informative features than it is to spend extra time tuning the learning algorithms.

Chapter 4 : CiteSeerX â€" A Supervised Machine Learning Approach to Conjunction Disambiguation in Na

A Supervised Machine Learning Approach to Conjunction Disambiguation in Named Entities PaweÅ, Mazurâˆ—â€”âˆ—Institute of Applied Informatics WrocÅ,aw University of Technology Wyb. Wyspianskiego 27,Â´.

Most of these approaches share a common scheme: A score for each referent hal, version 1 - 27 Sep is calculated according to the context toponyms, and the referent with the highest score is selected. According to the method used to calculate the score, Toponym Disambiguation TD methods may be grouped into three main categories, as proposed by [7]: Map-based methods usually do not need any information other than the coordinates of the places appearing in context. The size of a location, for example measured through the number of inhabitants living in that place, is an important clue in knowledge-based methods: Population represents a good rule-of-thumb if other context information is not available. This heuristic was included in the methods of [12, 1, 2]. Structural information derived from the containment, or part-of, relationship was also used to develop methods that are based on the idea that the places in the context are usually contained in the same region or geographical area. Some examples of hierarchy-based algorithms are [3, 7]. Wikipedia was also successfully used by [11], who took advantage of article templates, categories and referents links to other articles in Wikipedia. Data-driven methods are not commonly used in TD, mostly because of the lack of geo- graphically tagged data and the problem in the classification of unseen toponyms. Nev- ertheless, supervised classifiers has been implemented with mixed results. Toponym Ambiguity The ambiguity of a given toponym depends strictly on the specific resource used to rep- resent the world usually a gazetteer or an ontology: GeoPlanet and 40 according to Geonames. Gazetteers and geographical ontologies may be characterised according to their scope, coverage and detail or granularity. The scope of a geographic resource indicates whether a resource is limited to a region or a country for instance, the scope of Ordnance Survey data is limited to Great Britain , or it is a resource covering all the parts of the world. Coverage is determined by the number of toponyms listed in the resource. Detail is related to how fine-grained is the resource with respect to the area covered. With an highly detailed resource, the chances of finding an ambiguity within a single state or re- gion increase; therefore, the methods that relies on the containment relationship to resolve ambiguities may be negatively affected by the use of a highly detailed resource. While the resource determines the disambiguation method, the task determines the resource to be used. In [6], the objective was to disambiguate toponyms in a local Italian newspa- per, where the granularity of toponyms was intended to be at the level of street names. These requirements made it necessary to create a custom resource by merging Geonames data with data extracted from the Google Maps API, because there was no freely avail- able resource listing street names together with city, region and country names. Street name ambiguities may frequently occur at the province level, with the result that it was impossible to apply disambiguation methods based on hierarchies. Some classes of toponyms are more ambiguous, on average, than others. It can be assumed that the average degree of ambiguity is proportional to the depth of a place in a geographical ontology. Evaluation The lack of a rigid evaluation procedure and a standardised test-set has represented a major issue in the evaluation of toponym disambiguation methods. Most of the initial works published calculated their results under particular conditions and using small test- sets that makes reproduction of their results difficult. In some cases, the same method applied to different test-sets obtained very different results: Therefore, in recent years there has been an effort to produce test collections specifically aimed to the evaluation of to- ponym disambiguation methods. In these test collections, toponyms have been manually labelled with the correct referent. A set of documents was manually annotated with co- ordinates from a custom gazetteer derived from Geonames. The resulting resource contains 6, toponym instances. SemCor is a collection of texts extracted from the Brown Corpus of American English, where each word has been labelled with a WordNet sense synset. It contains 1, toponym instances in its final version. The SpatialML corpus is a manually tagged collection of documents drawn mainly from broadcast news, newsgroups, and weblogs, which contains 4, toponyms instances. This corpus is

focused on smaller newspapers with a localised audience, which have a large presence on the Internet, in contrast with the other resources which have been derived from news wires, which is usually written for a broadly distributed geographical audience. The primary measures used for the evaluation of toponym disambiguation are precision and recall, where precision is calculated as the number of correctly disambiguated to- ponyms divided by the number of disambiguated toponyms, and recall is calculated as the number of correctly disambiguated toponyms divided by the number of toponyms in the test collection. Correctness is determined by comparing the location ID assigned by the disambiguation process to the location ID label contained in the resource. A fuzzy hal, version 1 - 27 Sep criterion for correctness could use the inverse of the distance between the assigned location and the correct one. This criterion reduces the importance of errors such as mismatch- ing the province with the city of the same name [10] though problems still exist since point-based geometry assigns a single location for example a centroid irrespective of the spatial footprint related to a toponym. Areas with Success and Failure In the last years, the relationship between Toponym Disambiguation and Information Retrieval has been object of many studies. This is also a result of the introduction in of the GeoCLEF evaluation campaign see also the notes from Mandl and Cardoso in this issue. In [5] an experiment using the whole collection showed that disambiguation improved results only in the case of short queries and with a relatively detailed geographical resource GeoNames was used, a configuration more similar to Web retrieval than to the IR ad-hoc task. Recent work by [16] shows an improvement in the retrieval performance when applying TD to IR over geological texts. Some tasks where toponym disambiguation has as of yet failed to prove useful are the scope resolution task and Question Answering QA. In the first case, [2] observed that content- based features, such as person names, are more useful for the resolution of geographical scopes than the disambiguation of ambiguous toponyms in text. In [5], the experiments in QA showed that the errors derived by the wrong labelling of ambiguous toponyms do not alter significantly the process of answer extraction. Conclusions In the last decade, toponym disambiguation has progressively gained importance within the field of Natural Language Processing. The number of works dealing with this task has been increasing: However, there is still room for further work: Geographically Constrained Information Retrieval. PhD thesis, University of Groningen. Toponym disambiguation by arborescent relationships. Journal of Computer Science, 6 6: Location and the web LocWeb Toponym Disambiguation in Information Retrieval. Grounding toponyms in an italian local news corpus. A conceptual density-based approach for the disambiguation of toponyms. International Journal of Geographical Information Systems, 22 3: An evaluation dataset for the toponym resolution task. Computers, Environment and Urban Systems, 30 4: Annotation Scheme, Corpora, and Tools. Lieberman, Hanan Samet, Geotagging with local lexicons to build indexes for textually-specified spatial data. Classification, Disambiguation and Modelling. PhD thesis, Imperial College London. A confidence-based framework for disambiguating geographic terms. Toponym disambiguation using events. Disambiguating geographic names in a historical digital library. Association for Computational Linguistics. Co-occurrence based place name disambiguation and its application to retrieval of geological text. An empirical study of the effects of nlp components on geographic ir performance. International Journal of Geographical Information Science, 22 3:

Chapter 5 : Approaches to disambiguating toponyms | Davide Buscaldi - blog.quintoapp.com

Approaches to Disambiguating Toponyms. â€¢ data-driven or supervised: methods based on machine learning techniques. Toponym Disambiguation, i.e. the task of assigning to place name their.