A confidence-based framework for disambiguating geographic terms

Erik Rauch, Michael Bukatin, and Kenneth Baker
MetaCarta, Inc.
875 Massachusetts Avenue, 6th Floor
Cambridge, MA 02139
{rauch,mishka,bakerkj}@metacarta.com

Abstract

We describe a purely confidence-based geographic term disambiguation system that crucially relies on the notion of "positive" and "negative" context and methods for combining confidence-based disambiguation with measures of relevance to a user's query.

1 Introduction

Many questions about geographic term disambiguation are standardly handled in a statistical framework: for example, we can ask that, in the absence of contextual information, with what probability does the word *Madison* refer to a person (e.g. *James Madison*), an organization (e.g. *Madison Guaranty Savings and Loan*), or a place (e.g. *Madison, Wisconsin*), and if no other disambiguation alternative exists, we can expect these three numbers to sum to 1 (i.e. behave like probabilities).

However, there are many other questions where a strictly probability-based framework is less appropriate. In particular, much of the information that could be used to disambiguate spatial references in natural language text is strongly non-local in character, and as we increase the amount of this background information, eventually we reach the point when the amount of training data per parameter is so low that there is no repeatable experiment to base probabilities on.

In such cases, "probabilities" are effectively used as a stand-in for what is really our confidence in one judgment or another. In this paper we describe some of the methods used in a purely confidence-based geographic term disambiguation system that crucially relies on the notion of "positive" and "negative" context.

Far more information is contained in unstructured text (such as the Web and message traffic) than in structured databases, so automatically processing ambiguous geographic references unlocks a large amount of information. Adding spatial dimensions to the document search systems requires new algorithms for determining the relevance of documents. We describe methods for combining confidence-based disambiguation with measures of relevance to a user's query.

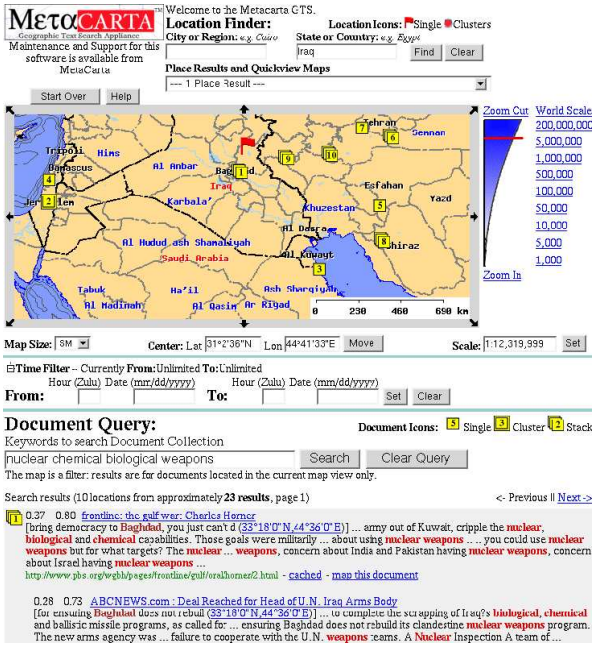

Figure 1: MetaCarta Geographic Text Search interface, showing query results ranked and plotted on a map.

It has become clear after several decades of artificial intelligence research that automated general natural language understanding is not feasible yet. However, we have been able to make progress by restricting our effort to the well-defined domain of geographic concepts, using statistical methods on extremely large corpora. To cope with billions of documents, we have built fast algorithms for extracting and disambiguating geographic

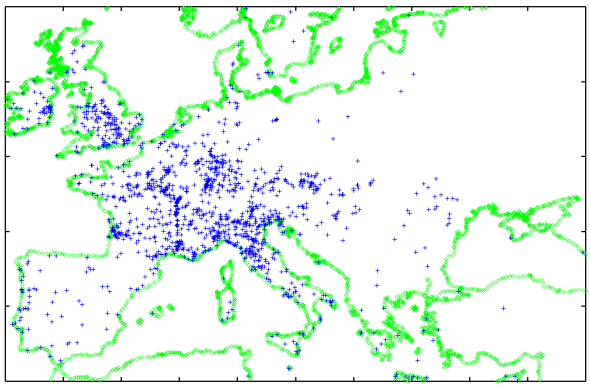

Figure 2: The distribution of occurrences of a term can identify geographic areas that it is relevant to. This example shows the distribution of the word *wine* in Europe.

information and fast database algorithms specifically for information which has a spatial component.

One form of information retrieval made possible by extracting geographic meaning in large corpora is geographic text search. Users are presented with an interface containing a traditional text search form combined with a map. They can zoom in on areas of the world that are of interest, and results of textual queries are plotted on the map (Figure 1). Other forms of data exploration are also made possible, such as exploring the spatial density pattern of documents satisfying a textual query (Figure 2).

In Section 2 we explore challenges of finding geographic meaning in natural language texts and give examples of typical ambiguities. In Section 3 we introduce some of our methods for determining geographic meaning in natural language. In Section 4 we describe some of the methods of determining geographic meaning during real-time processing. In Section 5 we describe some of our training methods. In Section 6 we describe methods for combining confidence-based disambiguation with measures of relevance to a user's query.

2 Challenges of finding geographic meaning in natural language text

Like other references in natural language text, geographic references are often highly under-specified and ambiguous. To take an extreme example, when encountering a reference to *Al Hamra*, the task is to determine which of the 65 places in the world with that name is being referred to, or even whether a place is being referred to at all, for the phrase also means *red* in Arabic. The same applies to the more than two dozen US towns named *Madison*. In fact, the majority of references to places are ambiguous in this way.

Human beings have a remarkable ability to derive useful information from ambiguous and under-specified references using real-world knowledge and experience, by deriving fuzzy rules from experience and knowing when to apply them. MetaCarta imitates this process using combinations of heuristics and data mining. For example, when encountering a mention of *Al Hamra*, a human analyst may notice that the rest of the document is focused on a region of *Oman*. Even if there is no mention of *Oman* itself, a mention of the nearby place *Safil* in the same document makes it likely that the *Al Hamra* in Oman is referred to. Even though there is another place named *Safil* in Iran, the towns of *Safil* and *Al Hamra* in Oman are close to each other, while there is no *Al Hamra* close to *Safil*, *Iran*.

People also apply real-world knowledge gained in other contexts: they know, for example, that a reference to a place called *Madison*, in the absence of a state, is more likely to refer to *Madison, Wisconsin* than the smaller *Madison, Iowa*; and they know that *James Madison* and *the Madison family* do not refer to places at all. Similarly they know that *Ishihara* does not refer to a place, even though there is a Japanese town of that name, if a government minister named *Ishihara* is being mentioned.

Moreover, much of the information people use to disambiguate references is not contained within the document itself, but is in the form of experience gained from reading many other documents. When encountering a name, people have various associations with the uses of this name they have seen before, and have a rough idea of how often it referred to places.

3 Methods for determining geographic meaning in natural language

MetaCarta has been able to imitate many aspects of this common-sense process because of the well-defined, low-dimensional space of geographic concepts. We begin with a gazetteer containing several million name-point and name-region pairs, and the enclosure relationship between regions and points. A given name n may refer to several points or regions, or refer to a non-geographic concept. To deal with ambiguity, for every potential reference of a name n to a point p, we estimate $c(p, n)$, the *confidence* that n really refers to p. The relevance of the document to each mentioned location must also be determined, in order to present the results that best satisfy the need for both correctness and relevance to a query, as described in Section 6.

There are two main phases of processing involved in the extraction of geographic information: training on large corpora, and real-time processing of a document.

In order to index large volumes of documents in a reasonable time, documents must be processed at a rate

of at least a hundred documents per second on a single workstation. This constraint affects the choice of heuristics used. Some of the methods of determining geographic meaning during real-time processing are described in Section 4.

The training phase requires some seed system capable of extracting the geographic information or, in the limiting case, some manually grounded documents. The quality of training depends on the quality of the seed, so as the system for the real-time processing of documents improves, we iterate the training process. Some details of the training process are described in Section 5.

4 Real-time processing of documents

4.1 Identifying candidate places

When processing a document, we begin by identifying potentially geographic references. For each, we identify all known candidates for the meaning of that reference. For example, a reference to 'Madison' can potentially mean any of 22 points with that name, or none of them.

The main source of geographic references are names from the high-quality MetaCarta gazetteer. See (Axelrod, 2003) for the process of building and updating this gazetteer. The procedure used to obtain realistic initial confidences associated with the gazetteer names is described in Section 5.1.

We mention some of the alternative sources of potentially geographic references here. We have capabilities allowing to match US postal addresses and pass them to third-party geolocation software producing a coordinate for the address.

Coordinates such as `38°01'10.5"N 121°44'48.8"W` or `56.51°N 25.86°E` are matched. We match some of JINTACCS (Department of the Army, 1990) message traffic formats such as `163940N 1062920E` (means 16°39'40"N 106°29'20"E).

The matches are then assigned initial confidences, and disambiguated using local and non-local information within the document.

4.2 Geographic disambiguation by local linguistic context

Similarly to other statistical NLP efforts, we use the local document context that a potentially geographic name occurs in. For example, the words *city of* or *mayor of* preceding or the words *community college* following a name like *Madison* are strong positive indicators of the geographic nature of this name. At the same time, the words *Mr.*, *Dr.*, or a common first name preceding or the words *will arrive* following a potential city name are strong negative indicators that the name in question is geographic.

We use the mixture of data mining procedures described in Section 5.2 and domain knowledge repositories containing context strings such as first names to form the sets of contexts we are using and to determine their strength as positive and negative indicators.

Heuristics then adjust the confidence $c_{geo}(n)$ that n refers to any geographic location (though not whether it refers to one of several synonymous locations) according to the nature and strength of these indicators.

Other local clues, such as absence of upper-case letters in the name itself or the resemblance of the name to an acronym have also proven useful to further adjust the values of c_{geo}.

The values of c_{geo} are then modified by non-local information as described below.

4.3 Geographic disambiguation by spatial patterns of geographic references in documents

We have found that there is a high degree of spatial correlation in geographic references that are in textual proximity. This applies not only to points that are nearby, such as Madison and Milwaukee, but also to the situation when points are enclosed by regions, e.g. Madison and Wisconsin. This correlation between geographic and textual distance is considered in estimating the confidence that a name refers to a point.

Some of our heuristics increase $c(p, n)$ based on how many and which points (and enclosing regions) are mentioned in the same document as n and their proximity. We make use of the characteristics of the nearby locations, and weight their influence as a decreasing function of geographic relationships to p and textual relationships to n. $c(p, n)$ is then increased by a saturating function of these influences.

4.4 Domain knowledge: population heuristics

In addition, population data in the gazetteer is also used. A place with a high population is more likely to be mentioned than a place with a lower one. Thus when disambiguating multiple referents with the same name, the population of each is considered. The confidence of a place p is decreased by an amount proportional to the logarithm of the ratio of the population of p to the population of all places with the name n.

4.5 Relative references

Until now we discussed the processing of stand-alone geographic references. We also process relative geographic references such as *15 miles northeast of Portland*. This relative reference is resolved in correspondence with the disambiguation of its anchor reference, *Portland*. If we decided that *Portland* refers to *Portland, Oregon* with confidence c, then we assume that *15 miles northeast of*

Portland refers to the point 15 miles northeast of *Portland, Oregon* with confidence $f(c)$, where $f(c)$ is greater than c, since the presence of a well-defined relative reference serves as an additional linguistic clue.

4.6 Temporal information

While not strictly a geographic issue, we mention here that the system also extracts temporal information from natural language documents. Currently we recognize *military date/time group* Zulu formats (Combined Communications-Electronics Board, 1983) such as `301535Z AUG 01` (means August 30, 2001 15:35:00 Zulu).

5 Training

5.1 Determining the geographic significance of gazetteer names

The methods for disambiguating geographic terms described above can also be exploited at the level of the corpus, despite the fact that the data used for training are untagged and therefore noisy. Since the real-time document processing system is high throughput, it can be applied to a training corpus consisting of a few hundred million documents.

If a name n is often given a high confidence of referring to a point p, then n is likely to refer to p even in the absence of other evidence in the document. Thus, each name-point pair n, p is given an initial confidence which is the average confidence assigned to an instance in the training corpus.

This initial confidence is then used as a starting point and modified by the other heuristics described above to obtain confidence for a name instance in a specific document during real-time document processing. Thus the training process is iterative.

5.2 Data mining of geographically significant local linguistic contexts

We currently use data mining on tagged corpora to learn the contexts in which geographic and non-geographic references occur, the words and phrases leading up to and trailing the name n. The tagged corpora were obtained using the Alembic tagger (Day et al., 1997). The accumulated statistics allow us to determine whether a specific context is a positive or negative indicator of a term being geographic, and the strength of this particular indicator. For any context C, an adjustment is applied to the confidence which is a nonlinear function of the probability of a geographic reference occurring in C in the tagged corpus.

6 Relevance

The addition of geographic dimensions to information retrieval means that in addition to the relevance of documents to a textual query, the relevance to the places mentioned in those documents must also be considered in order to rank the documents. The two kinds of relevance, traditional textual query relevance R_w and *georelevance* R_g, must be properly balanced to return documents relevant to a user's query. The traditional textual query relevance is obtained using standard techniques (Robertson and Jones, 1997).

Georelevance is based on both the geographic confidence of the place names used to place the document on the map, and the emphasis of the place name in the document. Emphasis is affected by the position P_n of the name in the document, and the prominence B_n. The latter is a function of whether it is in the title or header, whether it is emphasized or rendered in a large font, and other clues related to the nature and formatting of a document. This is similar to term relevance heuristics in information retrieval (Robertson and Jones, 1997), but the pattern of emphasis of geographic references is somewhat different. The function that assigns the emphasis component that is a function of in-document position is somewhat different than those usually used. It decreases from a maximum at the beginning of the document to a low number near the end of a long document, but increases near the bottom of the document to account for the increased relevance of information in footers. The frequency of the name F_n in the document is considered in a similar way to standard information retrieval techniques (Robertson and Jones, 1997).

Emphasis is also a function of the number of other geographic references S in the document. This is based on the assumption that a document does not have an unlimited amount of relevance to "spend" on places. Thus, a place mentioned in a document with many others is likely to be less relevant. Once emphasis $E(P_n, B_n, F_n, S)$ is calculated, it is multiplied by geoconfidence C_g to obtain the georelevance R_g.

We also compute a georelevance-like function for each location that could be referenced by a document. It varies as a function of character position in the document and is independent of geoconfidence.

Finally, the textual query relevance and georelevance are balanced as follows. The more terms m are in the user's query, the higher the weight W_w we assign to the term component of the query; however we use a function W_w that saturates at a maximal weight M ($.5 < M < 1$). The term relevance weight is defined as

$$W_w(m) = .5 + \frac{m-1}{m}(M - .5)$$

Georelevance and term relevance R_w are then combined as $(1 - W_w(m))R_g + W_w(m)R_w$.

7 Conclusion

The successful deployment of an industrial high-volume system partially based on the methods described here, even in the absence of large amounts of tagged data, has shown that many elements of common sense relating to geographic disambiguation can be encoded as heuristics in a confidence-based framework.

Acknowledgements

We would like to thank András Kornai and John Frank for fruitful suggestions and discussions regarding extraction of geographic information and to acknowledge MetaCarta team involved in other aspects of Geographic Text Search.

References

Amittai Axelrod. 2003. *On building a high performance gazetteer database*, this volume.

Combined Communications-Electronics Board. 1983. *Communication Instructions General (Unclassified) ACP 121(F)*, London. Available via URL `http://www.dtic.mil/jcs/j6/cceb/acps/acp121f.pdf`

David Day, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson and Marc Vilain. 1997. *Mixed-Initiative Development of Language Processing Systems*. In Proceedings of Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Washington, DC.

Department of the Army. 1990. *FM 24-33. Communications techniques: electronic counter-countermeasures*, Appendix A. Washington, DC. Available via URL `http://www.fas.org/irp/doddir/army/fm24-33/fm2433.htm`

S.E. Robertson and K. Sparck Jones. 1997. *Simple, proven approaches to text retrieval.* University of Cambridge Computer Laboratory Technical Report, May 1997.