Bootstrapping toponym classifiers

David A. Smith and **Gideon S. Mann**
Center for Language and Speech Processing
Computer Science Department, Johns Hopkins University
Baltimore, MD 21218, USA
{dasmith,gsm}@cs.jhu.edu

Abstract

We present minimally supervised methods for training and testing geographic name disambiguation (GND) systems. We train data-driven place name classifiers using toponyms already disambiguated in the training text — by such existing cues as "Nashville, Tenn." or "Springfield, MA" — and test the system on texts where these cues have been stripped out and on hand-tagged historical texts. We experiment on three English-language corpora of varying provenance and complexity: newsfeed from the 1990s, personal narratives from the 19th century American west, and memoirs and records of the U.S. Civil War. Disambiguation accuracy ranges from 87% for news to 69% for some historical collections.

1 Scope and Prior Work

We present minimally supervised methods for training and testing geographic name disambiguation (GND) systems. We train data-driven place name classifiers using toponyms already disambiguated in the training text — by such existing cues as "Nashville, Tenn." or "Springfield, MA" — and test the system on text where these cues have been stripped out and on hand-tagged historical texts.

As in early work with such named-entity recognition systems as Nominator (Wacholder et al., 1997), much previous work in GND has relied on heuristic rules (Olligschlaeger and Hauptmann, 1999; Kanada, 1999) and such culturally specific and knowledge intensive techniques as postal codes, addresses, and telephone numbers (McCurley, 2001). In previous work, we used the heuristic technique of calculating weighted centroids of geographic focus in documents (Smith and Crane, 2001). Sites closer to the centroid were weighted more heavily than sites far away unless they had some countervailing importance such as being a world capital.

News texts offer two principal advantages for bootstrapping geocoding applications. Just as journalistic style prefers identifying persons by full name and title on first mention, place names, when not of major cities, are often first mentioned followed by the name of their state, province, or country. Even if a toponym is strictly unambiguous, it may still be labelled to provide the reader with some "backoff" recognition. Although there is only one place in the world named "Wye Mills", an author would still usually append "Maryland" to it so that a reader who doesn't recognize the place name can still situate it within a rough area. In any case, the goal is to generalize from the kinds of contexts in which writers use a disambiguating label to one in which they do not.

Since news stories also tend to be relatively short and focused on a single topic, we can also exploit the heuristic of "one sense per discourse": unless otherwise indicated — e.g., by a different state label — subsequent mentions of the toponym in the story can be identified with the first, unambiguous reference. News stories often also have toponyms in their datelines that are disambiguated. Our news training corpus consists of two years (1989-90) of AP wire and two months (October, November, 1998) of Topic Detection and Tracking (TDT) data. The test set is the December, 1998, TDT data. See table 1 for the numbers of toponyms in the corpora.

In contrast to news texts, historical documents exhibit a higher density of geographical reference and level of ambiguity. To test the performance of our minimally-supervised classifiers in a particularly challenging domain, we test it on a corpus of historical documents where all place names have been marked and disambiguated. As with news texts, we initially train and test our classifiers on raw text. The range of geographic reference in these texts is somewhat similar to American news text: the corpus comprises the *Personal Memoirs* of Ulysses S. Grant and two nineteenth-century books of travel about California and Minnesota from the Library of Congress' American Memory project.[1] In all, we thus have about 600 pages of tagged historical text.

2 Experimental Setup

Dividing the corpora in training and test data, we train Naive Bayes classifiers on all examples of disambiguated toponyms in the training set. Although it is not uncommon for two places in the same state, for example, to share a name, we define disambiguation for purposes of these experiments as finding the correct U.S. state or foreign country. This asymmetry is reflected in U.S. news and historical text of the training data, where toponyms are specified by U.S. states or by foreign countries. We then run the classifiers on the test text with disambiguating labels, such as state or country names that immediately follow the city name, removed.

Since not all toponyms in the test set will have been seen in training, we also train backoff classifiers to guess the states and countries related to a story. If, for example, we cannot find a classifier for "Oxford", but can tell that a story is about Mississippi, we will still be able to disambiguate. We use a gazetteer to restrict the set of candidate states and countries for a given place name. In trying to disambiguate "Portland", we would thus consider Oregon, Maine, and England, among other options, but not Maryland. As in the word sense disambiguation task as usually defined, we are classifying names and not clustering them. This approach is practical for geographic names, for which broad-coverage gazetteers exist, though less so for personal names (Mann and Yarowsky, 2003). System performance is measured with reference to the naive baseline where each ambiguous toponym is guessed to be the most commonly occurring place. London, England, would thus always be guessed rather than London, Ontario. Bootstrapping methods similar to ours have been shown to be competitive in word sense disambiguation (Yarowsky and Florian, 2003; Yarowsky, 1995).

3 Difficulty of the Task

Our ability to disambiguate place names should be weighed against the ease or difficulty of the task. In a world where most toponyms referred unambiguously to one place, we would not be impressed by near-perfect performance.

Before considering how toponyms are used in text, we can examine the inherent ambiguity of place names in isolation. The Getty *Thesaurus of Geographic Names*, with over a million toponyms, not only synthesizes many contemporary gazetteers but also contains a wealth of historical names. In table 2, we summarize for each continent the proportion of places that have multiple names and of names that can refer to more than one place. Although these proportions are dependent on the names and places selected for inclusion in this gazetteer, the relative rankings are suggestive. In areas with more copious historical records—such as Asia, Africa, and Europe—a place may be called by many names over time, but individual names are often distinct. With the increasing tempo of settlement in modern times, however, many places may be called by the same name, particularly by nostalgic colonists in the New World. Other ambiguities arise when people and places share names. Very few Greek and Latin place names are also personal names.[2] This is less true of Britain, where surnames (and surnames used as given names) are often taken from place names; in America, the confusion grows as numerous towns are named after prominent or obscure people. What may be called a lack of imagination in the many 41 Oxfords, 73 Springfields, 91 Washingtons, and 97 Georgetowns seems to plague the very area — North America — covered by our corpora.

If, however, one Washington or Portland predominates in actual usage, things are not as bad as they seem. At the

Corpus	Train	Test	Tagged
News	80,366	1464	0
Am. Mem.	11,877	3782	342
Civ. War	59,994	787	4153

Table 1: Experimental corpora with toponym counts in unsupervised training and test and hand-tagged test sections.

Continent	% places w/mult. names	% names w/mult. places
N. & Cent. America	11.5	57.1
Oceania	6.9	29.2
South America	11.6	25.0
Asia	32.7	20.3
Africa	27.0	18.2
Europe	18.2	16.6

Table 2: Places with multiple names and names applied to more than one place in the Getty *Thesaurus of Geographic Names*

[1] Our annotated data also includes disambiguated texts of Herodotus' *Histories* and Caesar's *Gallic War*, but toponyms in the ancient (especially Greek) world do not show enough ambiguity with personal names or with each other to be interesting.

[2] In Herodotus, for example, the only ambiguities between people and places are for foreign names such as 'Ninus', the name used of Nineveh and of its mythical king.

Corpus	$H(\text{class})$	$H(\text{class}\|\text{name})$	% ambig.
News	6.453	0.241	12.71
Am. Mem.	4.519	0.525	18.81
Civ. War	4.323	0.489	18.49

Table 3: Entropy (H) of the state/country classification task

very worst, for a baseline system, one can always guess the most predominant referent. We quantify the level of uncertainty in our corpora using entropy and average conditional entropy. As stated above, we have simplified the disambiguation problem to finding the state or country to which a place belongs. For our training corpora, we can thus measure the entropy of the classification and the average conditional entropy of the classification given the specific place name (table 3). These entropies were calculated using unsmoothed relative frequencies. The conditional entropy, not surprisingly, is fairly low, given that the percentage of toponyms that refer to more than one place *in the training data* is quite low. Since training data do not perfectly predict test data, however, we have to smooth these probabilities and entropy goes up.

4 Evaluation

We evaluate our system's performance on geographic name disambiguation using two tasks. For the first task, we use the same sort of untagged raw text used in training. We simply find the toponyms with disambiguating labels — e.g., "Portland, Maine" —, remove the labels, and see if the system can restore them from context. For the second task, we use texts all of whose toponyms have been marked and disambiguated. The earlier heuristic system described in (Smith and Crane, 2001) was run on the texts and all disambiguation choices were reviewed by a human editor.

Table 4 shows the results of these experiments. The baseline accuracy was briefly mentioned above: if a toponym has been seen in training, select the state or country with which it was most frequently associated. If a site was not seen, select the most frequent state or country from among the candidates in the gazetteer. The columns for "seen" and "new" provide separate accuracy rates for toponyms that were seen in training and for those that were not. Finally, the overall accuracy of the trained system is reported. For the American Memory and Civil War corpora, we report results on the hand-tagged as well as the raw text.

Not surprisingly, in light of its lower conditional entropy, disambiguation in news text was the most accurate, at 87.38%. Not only was the system accurate on news text overall, but it degraded the least for unseen toponyms. The relative accuracy on the American Memory and Civil War texts is also consistent with the entropies presented above. The classifier shows a more marked degradation when disambiguating toponyms not seen in training.

Corpus	Baseline	Seen	New	Overall
News	86.36	87.10	69.72	87.38
Am. Mem.	68.48	74.60	46.34	69.57
(tagged)	80.12	91.74	10.61	77.19
Civ. War	78.27	77.23	33.33	78.65
(tagged)	21.94	71.07	9.38	21.82

Table 4: Disambiguation accuracy (%) on test corpora. Hand-tagged data were available for the American Memory and Civil War corpora.

The accuracy of the classifier on restoring states and countries in raw text is significantly, but not considerably, higher than the baseline. It seems that many of toponyms mentioned in text might be only loosely connected to the surrounding discourse. An obituary, for example, might mention that the deceased left a brother, John Doe, of Arlington, Texas. Without tagging our test sets to mark such tangential statements, it would be hard to weigh errors in such cases appropriately.

Although accuracy on the hand-tagged data from the American memory corpus was better than for the raw text, performance on the Civil War tagged data (Grant's *Memoirs*) was abysmal. Most of this error seems came from toponyms unseen in training, for with the accuracy was 9.38%. In both sets of tagged text, moreover, the full classifier performed below baseline accuracy due to problems with unseen toponyms. The back-off state models are clearly inadequate for the minute topographical references Grant makes in his descriptions of campaigns. Including proximity to other places mentioned is probably the best way to overcome this difficulty. These problems suggest that we need to more robustly generalize from the kinds of environments with labelled toponyms to those without.

5 Conclusions

Lack of labelled training or test data is the bane of many word sense disambiguation efforts. For geographic name disambiguation, we can extract training and test instances from contexts where the toponyms are disambiguated by the document's author. Tagging accuracy is quite good, especially for news texts, which have a lower entropy in the disambiguation task. In real applications, however, we do not usually need to disambiguate toponyms that already have state or country labels; we need to disambiguate unmarked place names. We investigated the ability of our classifier to generalize by evaluating on hand-corrected texts with all toponyms marked and disambiguated. The mixed results show that more generalization power is needed in our models, particularly the back-off models that handle toponyms unseen in training.

In future work, we hope to try further methods from WSD such as decision lists and transformation-based learning on the GND task. In any event, we hope that this should improve the accuracy on toponyms seen in training. As for disambiguating unseen toponyms, incorporating our prior work on heuristic proximity-base disambiguation into the probabilistic framework would be a natural extension. A fully hand-corrected test corpus of news text would also provide us with more robust evidence for classifier generalization.

Evidence learned by classifiers to disambiguate toponyms includes the names of prominent people and industries in a particular place, as well as the topics and dates of current and historical events, and the titles of newspapers (see figures 1 and 2). In our news training corpus, for example, Hawaii was most strongly collocated with "lava" and Poland with "solidarity" (case was ignored). In addition to their use for GND, such associations should be useful in their own right for event detection (Smith, 2002), personal name disambiguation, and augmenting the information in gazetteers.

References

[Kanada1999] Yasusi Kanada. 1999. A method of geographical name extraction from Japanese text for thematic geographical search. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pages 46–54, Kansas City, Missouri, November.

[Mann and Yarowsky2003] Gideon S. Mann and David Yarowsky. 2003. Unsupervised personal name disambiguation. In *CoNLL*, Edmonton, Alberta. (to appear).

[McCurley2001] Kevin S. McCurley. 2001. Geospatial mapping and navigation of the web. In *Proceedings of the Tenth International WWW Conference*, pages 221–229, Hong Kong, 1–5 May.

[Olligschlaeger and Hauptmann1999] Andreas M. Olligschlaeger and Alexander G. Hauptmann. 1999. Multimodal information systems and GIS: The Informedia digital video library. In *Proceedings of the ESRI User Conference*, San Diego, California, July.

[Smith and Crane2001] David A. Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *Proceedings of ECDL*, pages 127–136, Darmstadt, 4-9 September.

[Smith2002] David A. Smith. 2002. Detecting and browsing events in unstructured text. In *Proceedings of the 25th Annual ACM SIGIR Conference*, pages 73–80, Tampere, Finland, August.

[Wacholder et al.1997] Nina Wacholder, Yael Ravin, and Misook Choi. 1997. Disambiguation of proper names in text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 202–208, Washington, DC, April. Association for Computational Linguistics.

[Yarowsky and Florian2003] David Yarowsky and Radu Florian. 2003. Evaluating sense disambiguation performance across diverse parameter spaces. *Journal of Natural Language Engineering*, 9(1).

[Yarowsky1995] David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised mehtods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.

NASHVILLE , Tenn - Singer Marie Osmond will receive the 1988 Roy Acuff Community Service Award from the **Country Music** Foundation. She will be honored for her work as national chairwoman of the Osmond Foundation ... The honor is named for a **Grand Ole Opry star** known as "the king of **country music**".

NASHVILLE , Tenn - The home of **country music** is singing the blues after the sale of its last locally owned **music publising company** to CBS Records. Tree International Publishing, ranked as Billboard magazine 's No. 1 **country music** publisher for the last 16 years, is being sold to New York-based CBS for a reported $45 million to $50 million, The Tennessean reported today.

NASHVILLE , Tenn - **Country music** entertainer Johnny Cash was scheduled to be released from **Baptist Hospital** Tuesday, two weeks after undergoing heart bypass surgery, a hospital spokeswoman said Monday ...

Figure 1: Documents with Dateline of Nashville, having strong collocation **country music**

PORTLAND, Ore - Federal court hearing on whether to permit logging on timber tracts where northern **spotted owl** nests.

GRANTS PASS, Ore - ... "As more and more federal lands are set aside for **spotted owls** and other types of wildlife and recreation areas, the land available for perpetual commercial timber management decreases"...

SEATTLE - Interior Secretary Manuel Lujan says federal law should allow economic considerations to be taken into account in deciding whether to protect species like the northern **spotted owl**...

SAN FRANCISCO - Environmental groups can sue the government to try to stop logging of old-growth fir near **spotted owl** nests in western Oregon, a federal appeals court ruled Tuesday...

PORTLAND, Ore - Environmentalists trying to protect the northern **spotted owl** cheered a federal judge's decision halting logging on five timber tracts...

WASHINGTON - Are the **spotted owls** that live in the ancient forest of the Northwest really endangered or are they being victimized by the miniature radio transmitters that scientists use to track their movements?

SEATTLE - A federal court extended a ban Thursday on U.S Forest Servi ce plans to sell nearly 1 billion board feet of ancient timber from nine nationa l forests in two states where the northern **spotted owl** lives.

Figure 2: A sample of new stories with the keyword **spotted owl**, most are Oregon/Washington