Spatial characteristics of a large web n-gram corpus

Sautier, Jerome ; Derungs, Curdin

Abstract: N-gram corpora, though prominently used to structure and index large natural language corpora, are rarely in the focus of GIR. In this study we describe a step in this direction by characterizing spatial information in a large Web n-gram corpus provided by Microsoft. We explore how continent and country toponyms are represented in this corpus and if basic topological relations can be correctly retrieved. Results suggest that toponym ambiguity has major impact and that although retrieved topological relations are often correct, recall is considerably low. We conclude that further research is required if more fine grained spatial information is to be retrieved from n-grams.

DOI: https://doi.org/10.1145/2837689.2837691

Spatial Characteristics of a large Web N-gram Corpus

Jerome Sautier
University of Zurich
Department of Geography
jerome.sautier@outlook.com

Curdin Derungs
University of Zurich
Department of Geography
curdin.derungs@geo.uzh.ch

ABSTRACT

N-gram corpora, though prominently used to structure and index large natural language corpora, are rarely in the focus of GIR. In this study we describe a step in this direction by characterizing spatial information in a large Web n-gram corpus provided by Microsoft. We explore how continent and country toponyms are represented in this corpus and if basic topological relations can be correctly retrieved. Results suggest that toponym ambiguity has major impact and that although retrieved topological relations are often correct, recall is considerably low. We conclude that further research is required if more fine grained spatial information is to be retrieved from n-grams.

CCS Concepts

•**Information systems** → *Digital libraries and archives; Web mining;*

Keywords

Web N-gram; GIR; Spatial Information; Ambiguity

1. INTRODUCTION AND BACKGROUND

N-grams (NG) and associated frequency or probability measures are widely used for structuring and indexing large natural language corpora. NG are for instance used as an input in applications of machine translation or query completion [4]. However, retrieval of spatial information from NG remains a largely unexplored field (an exception is [1]). It is often not explicitly stated, but an important initial step in GIR is to know a corpus well. It is thus our goal to explore basic characteristics of spatial information in a large Web NG corpus [5] and to further tackle the question what spatial information can be retrieved, given the very short text samples in NG. In the first experiment we explore probabilities of occurrence of continent and country toponyms. In the second experiment we set out to test if basic topological relations, i.e. *IN* and *BORDERING*, can be successfully retrieved from NG.

GIR '15, November 26-27, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2837689.2837691

2. METHODOLOGY AND RESULTS

The Microsoft Web N-gram corpus can be accessed through a REST interface. The corpus consists of the Web documents indexed by BING in the EN-US market (i.e. some hundred billion Web pages) [5]. The interface allows to submit n-token sequences and retrieve respective joint probabilities (jP) (i.e. frequency of occurrence) and ranked lists of suggestions for the follow up token. Token sequences with $n > 4$ can not be processed.

2.1 Mapping Toponym Probabilities

Figure 1 shows jP values as retrieved for continent and the top 10 most frequent country toponyms, as well as a map of some 180 country jP values, where jP is visualized using 20% quantile.

The generally unsurprising distribution of continent jP values shows two peculiarities. Firstly, *Australia* appears as the most frequent continent toponym in NG, which might be due to the fact that it is both, a referent to a continent and a country (i.e. referent ambiguity, [3]). Secondly, *North America* gains a low jP value. Presumably for the reason that its official name consists of two tokens but is usually referred to as *America* in text. The top 10 most frequent country toponyms are predominated by large or prominent instances, with maybe the exception of *Australia* and *Mexico*, with the latter again being prone to referent ambiguity (city and country). The map of country jP values shows that for instance most European countries gain high country jP values, while countries in Africa are associated with low values. On the one hand, this indicates the uneven spatial coverage of the Internet, as for instance shown by [2]. On the other hand, it is interesting that the high jP value of Africa as a continent is not represented on country level within Africa, whereas in Europe an average continental jP is contrasted by relatively high values for the individual European countries. Among the top 20 countries with highest jP we also find *Georgia* and *Jordan*. Both toponyms are popular surnames and thus prone to semantic toponym ambiguity[3], plus Georgia being a state in the US (referent ambiguity).

2.2 Retrieving Topological Relations

In Figure 2 retrieval results of basic topological relations for countries and continents are summarized. The relations are retrieved by submitting token sequences of the form

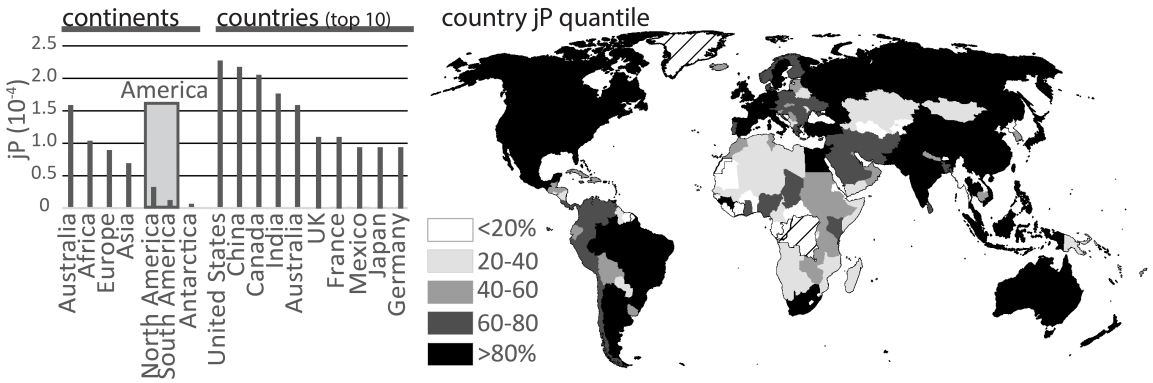

Figure 1: JP values for continent and country toponyms.

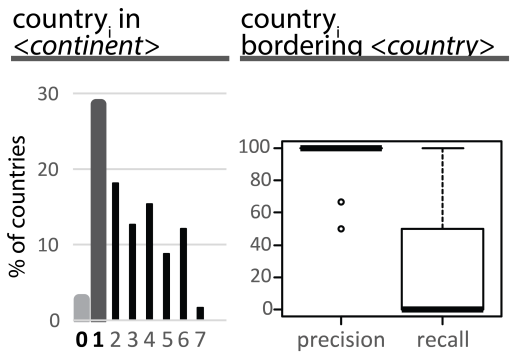

Figure 2: Evaluation of the retrieval of the topological relations *IN* and *BORDERING*.

country$_i$ IN (left) and *country$_i$ BORDERING*, with [1...*i*] representing all 192 countries to the NG interface. We then analyze the (up to) 1000 suggestions for follow up tokens of each sequence for the occurrence of continent and country toponyms respectively. The goal is to evaluate if correct *IN* and *BORDERING* relations can be retrieved. 5% of countries are *IN* no continent (left, **O**). Almost 30% of countries are *IN* the (one) correct continent (**1**) and the other two thirds of countries are *IN* several continents, with for instance 20% being *IN* five or more. This is another indication of the presents of ambiguity. To examine the correctness of retrieved *BORDERING* relations we use a GIS dataset with country borders as ground truth. Precision of *BORDERING* relations is close to 100% - i.e. almost all suggested *BORDERING* countries are also contained in the ground truth (left boxplot). Recall, in contrast, is in average 28%, with a median at 0% (right boxplot) - i.e. less than one third of all *BORDERING* relations can be retrieved from NG.

3. CONCLUDING DISCUSSION

Our results suggest that the occurrence of continent and country toponyms in NG is biased by toponym ambiguity (e.g. Australia or Georgia), length of token sequences, productive language use (e.g. North America instead of America) and uneven Internet coverage (e.g. African countries [2]) [2]. The retrieval of basic spatial relations shows mixed results. 30% or all countries are IN the (one) correct continent and BORDERING relations between countries can be retrieved with high precision but critically low recall. We could also show that the frequency of occurrence of continent toponyms does not reflect the frequency distribution on country level. It seems that for different regions, different spatial aggregations are used as a primary reference in the Internet. In Africa this is the continent level, whereas in Europe country toponyms are used more prominently.
We see two major implications of these results. Firstly, the retrieval of correct spatial information from NG requires profound toponym disambiguation, which is very challenging for short text samples. Additionally, we made the experience that spatial information, despite the vast amount of web pages incorporated in the NG corpus, is available to only a limited degree. Considering for instance the small amount of BORDERING relations between countries retrieved from NG. The impact of both implications can be expected to increase as we move beyond the relatively coarse continent or country resolution represented in this study. Thus, further research is required before NG might become a novel data source in GIR.

4. REFERENCES

[1] C. Derungs and R. Purves. Where is near? In *Proceedings of the 8th International Conference on Geographic Information Science*, 2014.

[2] M. Graham, B. Hogan, R. K. Straumann, and A. Medhat. Uneven geographies of user-generated information: patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):746–764, 2014.

[3] J. L. Leidner and M. D. Liberman. Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. In R. Purves and C. Jones, editors, *Letters on Geographic Information Retrieval*, pages 5–12. ACM Sigspatial Special, 2011.

[4] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[5] K. Wang, C. Thrasher, E. Viegas, X. Li, and B.-j. P. Hsu. An overview of Microsoft Web N-gram corpus and applications. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 45–48. Association for Computational Linguistics, 2010.