Modelling Swedish toponyms using a GRU

Linnea Strand

November 2019

1 Introduction

2 Background and preliminaries

2.1 How Swedish place names get their names

According to the Swedish Institute for Language and Folklore, *"[the] place names often describe a place, for example what it looks like, which animals and plants that live there or which person first cultivated the land."*[1]

This is typical of Swedish toponyms; they are often formed by combining two or more morphemes describing the geographical properties of their referent. Some examples are *sjö* (lake), *berg* (mountain or hill) and *skog* (forest), which can be combined in any which way to form e.g. *Skogsjö*, *Sjöskog* or *Bergsjön*.

As a matter of fact, place names can encode much more than just geographical information. For example, the suffix *-arv* means "land which has been inherited", and *-vi/-ve* means "holy place". Other frequent suffixes are ones pertaining to the size of a settlement: *-sta/-stad* (city/town), *-borg* (fortress) and *-by* (village).

2.2 Focus question

Swedish speakers seem to have an intuition when it comes to toponyms; when presented with a Swedish-sounding place name, be it familiar to them or not, they can usually guess whether it belongs to a rural area or an urban one. If we assume that such an intuition exists, we should be able to model it using machine learning. However, this 'intuition' could also be the effect of subconscious knowledge of Swedish geography.

As for this particular experiment, no groundbreaking results are to be expected. Firstly, the training data is sparse (see section 3.1). Secondly, place names are generally given when a new city or town is founded. With time, the conditions under which it was first founded are very likely to change; the population size might grow or shrink, industries come and go etc. As a result, they

[1] https://www.sprakochfolkminnen.se/sprak/namn/ortnamn.html

may not necessarily convey any up-to-date information about the referenced area.

2.3 Småort vs. tätort

The Swedish terms *småort* and *tätort*[2] were chosen to represent the "small town" and "large town" classes. According to Haldorson and Ben Daher (2016), *"The purpose of the [tätort] statistics is to show where in [Sweden] the population is concentrated or sparse"*, meaning these classes are quite ideal for the present study.

A *tätort* is defined as a populated area with more than 200 inhabitants, a concentration of holiday cottages less than 50% and a distance of 200 meters between houses[3]. A *småort* has a population of 50-199, with a distance of 150 m between houses. As of 2015, there are 2011 *tätorter* and 3135 *småorter* in Sweden, which also corresponds to the amount of each class in our data.

3 Method

3.1 The data

All the training and test data used in this experiment is available for free at Statistiska centralbyråns website and can be downloaded in `.xlxs` format. The data in question is from 2015.

A brief look at the data suggests that there may be no obvious correlation between place name and population size. As an example, there are 102 names ending in *-by* (village) in the småort group and 146 of the same in the tätort group. Conversely, there are 58 småort names ending in *stad* (town), while only 46 in the tätort group. Disregarding the contribution of any alternative spellings of these suffixes (*-sta, -städe, -byn*), this may well be a cue that these suffixes are no longer indicative of the population size.

3.2 Network

The following three methods were used to encode the place names:

1. Basic character encoding, i.e. each letter of the place name was represented by an index,

2. Splitting each place name into (semi) meaningful subparts, and then encoding these subparts in the same manner as above,

[2] As defined by Statistics Sweden, a Swedish government agency that provides official data to both the public as well as other Swedish government agencies

[3] This leaves some room for interpretation, as stated in the report *Översyn av metod och definition för SCBs avgränsningar av koncentrerad bebyggelse* (Haldorson and Ben Daher, 2016)

3. Splitting each place name into (semi) meaningful subparts, and then representing each subpart with a pretrained fastText vector.

The training and test data is selected by a random split using sklearns train_test_split method. Training will be done done on a batch size of 32 with a learning rate of 0.01 for 20 epochs, using Binary Cross Entropy as the loss function.

3.3 Preprocessing

3.3.1 SALDO sammansättningsanalys

To be able to use pretrained fastText embeddings, compositional names had to be split into their subparts:

(1) Halm-stad
straw-town

(2) Ljus-dal
light-valley

To achieve this, I used the compound analysis tool developed by the SALDO initiative (*Swedish Associative Thesaurus version 2*, Borin et al., 2008), which can be retrieved from their web page.

While it is impossible to say how many of these compounds were successfully analysed, this tool is to my knowledge the only compound analysis tool available for Swedish. A glance at the data reveals some mistakes:

(3) Tors-ås
Thor.GEN-ridge

(4) *Tor-sås
Thor-sauce

(5) Fot-ö
foot-island

(6) *Fo-tö
?-till

This approach poses several other limitations, one being that SALDO already contains some place names and therefore analyses these as common nouns instead of compound words. Some of these examples may be overridden by the next step in preprocessing, as explained in the next segment.

3.3.2 Swedish place name suffixes

In order to get around some of the limitations of the SALDO compound analysis mentioned above, common Swedish suffix names were extracted from the Swedish Wikipedia entry *Svenska ortnamnsefterled* (Swedish place name suffixes, 2020), which contains a total of 184 suffixes. These entries were scraped using the BeautifulSoup library (Richardson, 2007). Any place name ending in one of the suffix strings identified is then split at that point, overruling the compound analysis made by the SALDO tool, so that `Tor-sås` → `Tors-ås`.

Moreover, the following common place name suffixes were arbitrarily chosen and hard coded into the script:

-sta	town
-hult	forest?
-borg	fort
-vik	bay
-strand	beach
-berg	mountain
-norra	north(ern)
-södra	south(ern)
-västra	west(ern)
-östra	east(ern)
-bro	bridge
-bron	bridge.DEF
-sund	strait

3.4 fastText embeddings

To encode the subparts, pretrained fastText vectors, trained on Common Crawl and Wikipedia, from *Word vectors for 157 languages* (Grave et al., 2018) were used. More specifically, *"[t]hese models were trained using [Continuous Bag of Words] with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives".*

As only roughly 40 % of the vocabulary could be found among the fasttext vectors, the remaining word representations were generated using the print-word-vectors function.

4 Results

The character-based model achieved an accuracy of around 52%, meaning it is marginally better than a lucky guess. The second approach using the compound analysis (without pretrained embeddings) performs better, with an accuracy of about 60%.

5 Conclusion

I have presented a binary place name classifier trained on strings and compounds respectively.

It is fully possible that a fastText model covering more of the vocabulary words could improve the overall performance.

The experiment could be extended further by incorporating Swedish place names from Finland into the data (as of 1960, the Nordic countries share the same definition of *tätort* (Haldorson and Ben Daher, 2016)).

References

Borin, L., Forsberg, M., and Lönngren, L. (2008). Saldo 1.0 (svenskt associationslexikon version 2). *Språkbanken, University of Gothenburg*.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Haldorson, M. and Ben Daher, K. (2016). Översyn av metod och definition för scbs avgränsningar av koncentrerad bebyggelse. $\mathrm{https} www.scb.se Statistik_P ublikationer MI0810\_2015A01\_BR_M IFT1601 pdf, retrievedonDecember10$

Richardson, L. (2007). Beautiful soup documentation. *April*.

Wikipedia contributors (2020). Svenska ortnamnsefterled — Wikipedia, the free encyclopedia. [Online; accessed 20-Nov-2019].