Isolation and Mapping of Place-Name Forms in Toponymic Data

Tobias Roth
Schweizerdeutsches Wörterbuch
Auf der Mauer 5
8001 Zürich
Switzerland
tobias.roth@idiotikon.ch

Abstract

We apply a customised approximate matching method to toponymic text data in order to isolate single place-name forms. Current place-names are matched to current and historical variants in standard and non-standard spelling. Such one-to-one mappings are preferred to text snippets with context, e. g. in the case of geo-referencing historical documents. The presented method yields an error rate of about 2%, which can be reduced manually and with reasonable effort to approximately 1%.

1 Introduction

An important task in the digitisation of historical documents is the tagging of place-names and the geo-referencing of these place-names found. This geo-referencing task can be completed much more efficiently if the tagging tool has access to a mapping from historical place-name forms to geographical coordinates. A promising data source for such mappings are toponymic projects (books of place-names) where one can often find both geographical coordinates and place-name forms in the historical-evidence sections.

Many of the toponymic projects also have their data in digital form. Yet, the records used as evidence are normally given as a line of plain text with minimal context, but without explicitly indicating the place-name form itself. For the usage of such data in geo-referencing the actual place-name form has to be isolated first. The present paper will explore methods to detect place-name forms in toponymic records. The problem does not look very difficult at first sight. Typically, there is a reference form (normally the current name) and a line of text that contains a form of the same place-name (in standard or non-standard spelling). An example of this historical-evidence part is shown below. It is an abbreviated entry for the name *Waldrüti* from Reber (2014):[1]

Waldrüti
Sources:
[...]
1534: ein Stuck matten vf der wald Rütj am menweg (Zins und Zehnten F1, 90r)
1548: wider an die walld Rüttÿ, biß vff ... kalberweÿde (Gösg Urb 1548, unpag.)
[...]
1704: die waldrüthj sampt der Ziegermatt (Ber 159, 198r)
1826: Jn der Wald-Rüti, Matten & Holzland (Haue Gb 1826, 463)
1872: Waldrüti (HaIf ÜbPlan 1872, Übersichtsplan)
[...]

For toponymic data in German speaking Switzerland which is considered first here, digitally readily available data amounts to roughly 450 000 toponyms with an estimated number of about 1.2 million source records.[2] With this amount of data in mind the focus lies on automated matching, not manual annotation. The present matching problem is situated somewhere between the normalisation problem of historical spellings and named entity recognition.

The paper is organised as follows: it starts with a short description of the data in question. We then look at different matching methods and their results and try to optimise the matching method to our case. We conclude with an error analysis and a general summary.

2 Forms of place-names present in the data

The data we would like to cover is toponymic data in German speaking Switzerland. There are or have been several regionally organised toponymic projects. This alone accounts for a certain heterogeneity in the data although the projects on the

[1] Cf. also https://search.ortsnamen.ch/record/106016746 (22.07.2016) for the complete entry.

[2] Cf. https://www.ortsnamen.ch (22.07.2016).

whole are quite compatible. Many inter-project diffences can also be found within projects. Such differences that are particularly relevant with respect to the present form-isolation task are presence or absence of a dialectal reference form, length of the context given in evidence records, additional information coming with the reference form, etc.

2.1 Present-day forms

Present-day i. e. 20th and 21st century forms in the records are easiest to match: They often coincide with the reference form and comply with standard spelling. They frequently come from maps, so they have little or no context with them. One difficulty can be dialectal forms as they show non-standard spelling. There even are phonetically transcribed strings (following different transcription systems).

2.2 Historical forms

Historical forms tend to differ considerably from the present-day reference form. Different patterns of deviation can be observed.

2.2.1 Non-standard orthography

Older forms show more variety in spelling as there was no standard orthography established yet. Some characters were used differently and there were also characters that are not in use anymore. Tokenisation differs sometimes: a compound word written as one word today is often written in two or more words in historical documents.

2.2.2 Discontinuous forms

In some cases forms are discontinuous, with two patterns that can be found frequently. One is the coordination pattern with ellipsis as in *X or Y street* with the target form being *X street*.

The other rather frequent case is the swapping of elements as in *X street* vs. *street to X*.

2.2.3 Name change

Names can change over time. It is not the loss or gradual change of phonetic material that is addressed here. If places are completely renamed this is, of course, a nearly unsurmountable barrier for a linguistic matching algorithm. But sometimes it is just parts of names that change: attributes are omitted, added or replaced, etc.

2.2.4 Substitutions with synonyms

A special variant of name change is the substitution of name constituents with synonyms. This is frequent with classifying constituents (e. g. routes can be called *Strasse, Weg* or *Gasse* interchangeably), but can also happen with attributes (e. g. *unter-* vs. *nieder-* for English *lower;* or the historical form *leupriesters garten* for modern *Pfarrgarten,* English *parish garden).*

2.2.5 Translations

Some of the older sources are written in Latin. Place-names mentioned in these documents are often also translated to Latin, at least the readily translatable elements.

Examples are attributes like *inferior* for English *lower* as in the record *in Ernlisbach Inferiori*[3] for modern *Niedererlinsbach*. Another frequently translated element is *bonum* for German *Gut* (English *estate, manor),* e. g. in the record *Bonum Scherерin*[4] for a now extinct toponym *Schärersguet.*

2.2.6 Uncertain naming status

If you look at a record it is not always clear which elements belong to the name and which ones are merely additional attributes that describe the place. There are records where all the elements you can find in the modern name are already present, but the sentence structure suggests that it is not a name yet. Attributive relative clauses are instances of this pattern: for the modern form *Trimbacherstrasse* there are historical records like *an der strasß die gon Trimppach godt*[5] (English *at the road that goes to Trimbach).*

2.3 Inflected forms

Both historical and modern forms can occur with inflectional endings. Inflectional forms are more frequent in older sources as present-day sources are very often maps or geographical information systems.

3 Matching of place-names

The isolation of actual place-name formes in place-name data is a rather specialised approximate-matching task. Classical named entity recognition (NER, cf., e. g., Sekine and Ranchhod (2009)) is not likely to perform well in this case. Although context is very restricted there can occur many more place-names and other named entities in such a text snippet, not only the wanted form.

Algorithms for the normalisation or canonicalization of historical text could be more helpful here.

[3]Record from 1406 (Reber, 2014).
[4]Record from 1423 (Reber, 2014).
[5]Record from 1623 (Reber, 2014).

For a general overview see, e. g., Piotrowski (2012, 69ff.). Different methods of approximate matching have been proposed. Hauser and Schulz (2007) and Bollmann et al. (2011) both use training data to automatically deduce weights for use in the computation of an edit-distance based similarity score; very similar Pilz et al. (2008), but with manual rule derivation in addition. Jurish (2008) converts the text with an adapted letter-to-sound system before comparison.

3.1 Methods and results

3.1.1 Development and test data

For development and test purposes, in a random sample of 6 000 records from Reber (2014) the actual place-names have been tagged manually. About 800 of these records were not used because they concerned family names or because they were phonetic transcriptions in IPA. Half of the remaining records were used as a development set, the other half as the final test set.

Another set of around 30 000 records with their corresponding isolated name forms from Dittli (2007) was used in development only.

3.1.2 Similarity based on edit distance

The following example can help to show what the task in question exactly consists of. It is a record for the name *Waldrüti*:[6]
wider an die walld Rüttÿ, biß vff . . . kalberweÿde
Given the standard form *Waldrüti* the desired result of the matching task for this record is the string *walld Rüttÿ*.

As a kind of baseline, matching was first performed using a similarity ratio based on simple edit distance (Levenshtein, 1966) computed with all strings transformed to lower case (column *ED* in table 1; see column *BL* in table 1 for baseline rates with random selection of words[7]). The ratio was computed as follows – with a cost of 1 for delete and insert, cost 2 for replace operations:

$$sim(x,y) = \frac{length(x) + length(y) - dist(x,y)}{length(x) + length(y)}$$

The error rate in the test set was 3.3% with this method. It got slightly better (3.1%) if all diacritics in the text were removed (column *ASC* in table 1); i. e. the text *wider an die walld rutty, biß vff . . . kalberweyde* was compared to the converted version of the reference word (*waldruti*).

Century	Count	Error rates in %			
		BL	ED	ASC	CST
<15th	68	80.9	13.2	13.2	10.3
15th	144	93.1	9.0	9.0	6.3
16th	524	94.5	6.1	5.7	3.6
17th	195	81.5	1.5	1.5	1.5
18th	226	77.4	4.9	4.4	1.8
19th	781	52.5	1.8	1.7	0.9
20th	312	11.5	1.0	1.0	0.6
21st	391	5.9	0.3	0.3	0.3
total	2641	56.3	3.3	3.1	2.0

BL = baseline; random selection of items
ED = edit distance, lower case
ASC = edit distance, lower case, diacritics removed
CST = edit distance after customised transformation

Table 1: Error rates with different matching methods (by century).

3.1.3 Weighted similarity

As we could use a set of 30 000 records with manually pre-annotated place-name forms (Dittli, 2007) we tried to improve the simple edit-distance based method above by taking these records as training data. We deduced replacement rules from this data set, comparable to methods described in Bollmann et al. (2011) and Hauser and Schulz (2007). The rules operated on one character with one character of context to the left and to the right. The cost of a given replacement depended on the ratio of its application in the training data, with a maximum cost assigned to unseen replacements. The similarity ratio was then computed as above, but with this weighted cost function.

The resulting error rate in the test set was at 3.4% and thus even lower than with the simple edit-distance based approach. A closer look at the training data suggests that there are too many errors in it,[8] so we decided not to further pursue the weighted-similarity branch for lack of adequate data.

3.1.4 Customised transformation and matching

Another possible approach are all the methods that try to simplify the strings before matching (after

[6] Record from 1548 (Reber, 2014).

[7] The low error rates for the 20th and 21st century even with random selection are again a sign of the many one-word records that come from maps or geographical databases.

[8] This data was not used in the printed version, so at some point it presumably was not maintained properly anymore.

the model of phonetic simplification as in methods like *Soundex* or *Kölner Phonetik* (Postel, 1969), see also Piotrowski (2012) and Jurish (2008)).

We adopted such a method with a very small set of manually selected replacement rules (see also Pilz et al. (2008) and Jurish (2008)). We replaced different writing variants of umlaut to e (e. g. *ö* to *oe*), merged *i, y, j* and *ie* to *i*, *th* to *t*, removed certain diacritics such as accents, etc. The rule set comprised less than 30 rules, for the simple reason that, at this point, rules we added (collected from our experiments in 3.1.3) did not further improve the error rate.

Sequences of identical characters were then reduced to one occurrence. Computation of the similarity ratio was done like in 3.1.2. Certain character alternations were not replaced before computation but were assigned reduced cost. An example is *v* that frequently alternates with *u* but also with *f* and *w*. It is easier to assign a reduced cost afterwards than to decide beforehands whether it is used as a vowel or as a consonant.

The inspection of the remaining errors in the development set led us to allow for discontinuous forms and discontinuous forms with swapped order (cf. 2.2.2). We also introduced penalties for forms that started or ended in certain words such as articles or prepositions, and we favoured forms that occurred just after an article or the like.

As a result we could lower the error rate in the test set to 2.0% (see column *CST* in table 1 for detailed results by century).

4 Error analysis and error management

There are some error types though that cannot be handled well with this procedure. Notably the types mentioned in 2.2.3–2.2.5 (name change, synonym substitution, translation) where the difference is not just a matter of spelling or sound change. A much more sophisticated apparatus than the one set up would be needed to account for these error types.

An error analysis with error rates by similarity ratio can show whether a threshold for the similarity ratio might be useful or how efficient manual post-processing might be. Table 2 presents these figures for our test set. The second and third columns give record counts and error rates for every similarity range. The two last columns show cumulated percentages of record counts as well as the proportion of all errors within these records. There are, for example, 75 records with a similarity ratio of 0.6–0.7, the error rate within these 75 records is at 21.3%; the records with a similarity ratio of up to 0.7 constitute 4.2% of all records, and they contain 50% of all errors.

Sim. ratio	**Count**	**Err. %**	**Cumulated**	
			% records	% err.
0.0–0.4	4	50.0	0.2	3.8
0.4–0.5	7	42.9	0.4	9.6
0.5–0.6	26	19.2	1.4	19.2
0.6–0.7	75	21.3	4.2	50.0
0.7–0.8	175	6.3	10.9	71.2
0.8–0.9	537	1.3	31.2	84.6
0.9–1.0	1817	0.4	100.0	100.0

Table 2: Error analysis by similarity ratio.

As 50% of all errors are in a well-defined set of around 4% of the records it could be considered to correct these errors manually. You could thus – with a reasonable effort – reduce the overall error rate to about 1%.[9]

Depending on the application one could also introduce a threshold for the similarity ratio of e. g. 0.7, but then you would lose the correctly classified forms of this range and these are likely to be the most interesting ones.

5 Conclusion and outlook

This paper has shown that place-name forms can reliably be detected in toponymic data using an automated matching method with a few manually set replacement rules and an edit-distance based similarity score (2% errors). The algorithm performs rather well in discerning doubtful cases: half of all the errors are in those 4% of the records with the lowest similarity ratio. Manual correction of these 4% can further reduce the error rate to about 1%.

For even further improvement such manual corrections could be taken as additional reference forms. Or, if set up as a web service for georeferencing historical documents, freshly annotated forms could and should be fed back into the original system.

References

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Applying rule-based normalization to different types of historical texts – an evaluation. In *Pro-*

[9] For the situation in Switzerland depicted in the introduction, it would mean that about 50 000 records would have to be checked manually.

ceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2011). Poznan, Poland.

Beat Dittli. 2007. *Zuger Ortsnamen. Lexikon der Siedlungs-, Flur- und Gewässernamen im Kanton Zug. Lokalisierung, Deutung, Geschichte. 5 Bände und Kartenset.* Balmer Verlag, Zug.

Andreas W. Hauser and Klaus Schulz. 2007. Unsupervised learning of edit distance weights for retrieving historical spelling variations. In Stoyan Mihov and Klaus U. Schulz, editors, *Finite State Techniques and Approximate Search, Proceedings of the First Workshop on Finite-State Techniques and Approximate Search*, pages 1–6, Borovets, Bulgaria.

Bryan Jurish. 2008. Finding canonical forms for historical German text. In Angelika Storrer, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner, editors, *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing* (KONVENS 2008), pages 27–37. Mouton de Gruyter, Berlin.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and Control Theory*, 10:707–710.

Thomas Pilz, Andrea Ernst-Gerlach, Sebastian Kempken, Paul Rayson, and Dawn Archer. 2008. The identification of spelling variants in english and german historical texts: Manual or automatic? *Literary and Linguistic Computing*, 23(1):65–72. `http://llc.oxfordjournals.org/cgi/content/full/fqm044?ijkey=0RtdsFnq2rH7gwL&keytype=ref`.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Number 17 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers LLC.

Hans Joachim Postel. 1969. Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. *IBM-Nachrichten*, 19. Jahrgang:925–931.

Jacqueline Reber, editor. 2014. *Die Flur- und Siedlungsnamen der Amtei Olten-Gösgen*, volume 3 of *Solothurner Namenbuch*. Schwabe, Basel.

Satoshi Sekine and Elisabete Ranchhod, editors. 2009. *Named Entities: Recognition, Classification, and Use*. John Benjamins Publishing, Amsterdam and Philadelphia.