Automatic Annotation of Bibliographical References with Target Language

Harald Hammarström
Dept. of Comp. Sci.
Chalmers University
S-412 96 Gothenburg
SWEDEN
harald2@chalmers.se

Abstract

In a large-scale project to list bibliographical references to all of the ca 7 000 languages of the world, the need arises to automatically annotated the bibliographical entries with ISO-639-3 language identifiers. The task can be seen as a special case of a more general Information Extraction problem: to classify short text snippets in various languages into a large number of classes. We will explore supervised and unsupervised approaches motivated by distributional characterists of the specific domain and availability of data sets. In all cases, we make use of a database with language names and identifiers. The suggested methods are rigorously evaluated on a fresh representative data set.

1 Introduction

There are about 7 000 languages in the world (Hammarström, 2008) and there is a quite accurate database of which they are (Gordon, 2005). Language description, i.e., producing a phonological description, grammatical description, wordlist, dictionary, text collection or the like, of these 7 000 languages has been on-going on a larger scale since about 200 years. This process is fully decentralized, and at present there is no database over which languages of the world have been described, which have not, and which have partial descriptions already produced (Hammarström, 2007b). We are conducting a large-scale project of listing all published descriptive work on the languages of the world, especially lesser-known languages. In this project, the following problem naturally arises:

Given: A database of the world's languages (consisting minimally of <unique-id, language-name>-pairs)

Input: A bibliographical reference to a work with descriptive language data of (at least one of) the language in the database

Desired output: The identification of which language(s) is described in the bibliographical reference

We would like to achieve this with as little human labour as possible. In particular, this means that thresholds that are to be set by humans are to be avoided. However, we will allow (and do make use of – see below) supervision in the form of databases of language references annotated with target language as long as they are *freely available*.

As an example, say that we are given a bibliographical reference to a descriptive work as follows:

> Dammann, Ernst 1957 *Studien zum Kwangali: Grammatik, Texte, Glossar*, Hamburg: Cram, de Gruyter & Co. [Abhandlungen aus dem Gebiet der Auslandskunde / Reihe B, Völkerkunde, Kulturgeschichte und Sprachen 35]

This reference happens to describe a Namibian-Angolan language called Kwangali [kwn]. The task is to automatically infer this, for an arbitrary bibliographical entry in an arbitrary language, using the database of the world's languages and/or databases of annotated entries, but without humanly tuned thresholds. (We will assume that

Coling 2008: Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, pages 57–64
Manchester, August 2008

the bibliographical comes segmented into fields, at least as to the title, though this does not matter much.)

Unfortunately, the problem is not simply that of a clean database lookup. As shall be seen, the distributional characteristics of the world language database and input data give rise to a special case of a more general Information Extraction (IE) problem. To be more precise, an abstract IE problem may be defined as follows:

- There is a set of natural language objects O
- There is a fixed set of categories C
- Each object in O belong to zero or more categories, i.e., there is a function $C : O \rightarrow Powerset(C)$
- The task is to find classification function f that mimics C.

The special case we are considering here is such that:

- Each object in O contains a small amount of text, on the order of 100 words
- The language of objects in O varies across objects, i.e., not all objects are written in the same language
- $|C|$ is large, i.e., there are many classes (about 7 000 in our case)
- $|C(o)|$ is small for most objects $o \in O$, i.e., most objects belong to very few categories (typically exactly one category)
- Most objects $o \in O$ contain a few tokens that near-uniquely identifies $C(o)$, i.e., there are some words that are very informative as to category, while the majority of tokens are very little informative. (This characteristic excludes the logical possibility that each token is fairly informative, and that the tokens *together*, on an equal footing, serve to pinpoint category.)

We will explore and compare ways to exploit these skewed distributional properties for more informed database lookups, applied and evaluated on the outlined reference-annotation problem.

2 Data and Specifics

The exact nature of the data at hand is felt to be quite important for design choices in our proposed algorithm, and is assumed to be unfamiliar to most readers, wherefore we go through it in some detail here.

2.1 World Language Database

The Ethnologue (Gordon, 2005) is a database that aims to catalogue all the known living languages of the world.[1] As far as language inventory goes, the database is near perfect and language/dialect divisions are generally accurate, though this issue is thornier (Hammarström, 2005).

Each language is given a unique three-letter identifier, a canonical name and a set of variant and/or dialect names.[2] The three-letter codes are draft ISO-639-3 standard. This database is freely downloadable[3]. For example, the entry for Kwangali [kwn] contains the following information:

> Canonical name: Kwangali
> ISO 639-3: kwn
> Alternative names[4]: {Kwangali, Shisambyu, Cuangar, Sambio, Kwangari, Kwangare, Sambyu, Sikwangali, Sambiu, Kwangali, Rukwangali}.

The database contains 7 299 languages (thus 7 299 unique id:s) and a total of 42 768 name tokens. Below are some important characteristics of these collections:

- Neither the canonical names nor the alternative names are guaranteed to be unique (to one language). There are 39 419 unique name strings (but 42 768 name tokens in the database!). Thus the average number of different languages (= unique id:s) a name denotes is 1.08, the median is 1 and the maximum is 14 (for Miao).

[1] It also contains some sign languages and some extinct attested languages, but it does not aim or claim to be complete for extinct and signed languages.

[2] Further information is also given, such as number of speakers and existence of a bible translation is also given, but is of no concern for the present purposes.

[3] From `http://www.sil.org/iso639-3/download.asp` accessed 20 Oct 2007.

[4] The database actually makes a difference between dialect names and other variant names. In this case Sikwangali, Rukwangali, Kwangari, Kwangare are altername names denoting Kwangali, while Sambyu is the name of a specific dialect and Shisambyu, Sambiu, Sambio are variants of Sambyu. We will not make use of the distinction between a dialect name and some other alternative name.

- The average number of names (including the canonical name) of a language is 5.86, the median is 4, and the maximum is 77 (for Armenian [hye]).

- It is not yet well-understood how complete database of alternative names is. In the preparation of the test set (see Section 2.4) an attempt to estimate this was made, yielding the following results. 100 randomly chosen bibliographical entries contained 104 language names in the title. 43 of these names (41.3%) existed in the database as written. 66 (63.5%) existed in the database allowing for variation in spelling (cf. Section 1). A more interesting test, which could not be carried out for practical reasons, would be to look at a language and gather *all* publications relating to that language, and collect the names occurring in titles of these. (To collect the full range of names denoting languages used in the bodies of such publications is probably not a well-defined task.) The Ethnologue itself does not systematically contain bibliographical references, so it is not possible to deduce from where/how the database of alternative names was constructed.

- A rough indication of the ratio between spelling variants versus alternative roots among alternative names is as follows. For each of the 7299 sets of alternative names, we conflate the names which have an edit distance[5] of $\leq i$ for $i = 0, \ldots, 4$. The mean, median and max number of names after conflating is shown below. What this means is that languages in the database have about 3 names on average and another 3 spelling variants on average.

i	Mean	Median	Max
0	5.86	4	77 'hye'
1	4.80	3	65 'hye'
2	4.07	3	56 'eng'
3	3.41	2	54 'eng'
4	2.70	2	47 'eng'

2.2 Bibliographical Data

Descriptive data on the languages of the world are found in books, PhD/MA theses, journal articles, conference articles, articles in collections and manuscripts. If only a small number of languages is covered in one publication, the title usually carries sufficient information for an experienced human to deduce which language(s) is covered. On the other hand, if a larger number of languages is targeted, the title usually only contains approximate information as to the covered languages, e.g., *Talen en dialecten van Nederlands Nieuw-Guinea* or *West African Language Data Sheets*. The (meta-)language [as opposed to target language] of descriptive works varies (cf. Section 2.4).

2.3 Free Annotated Databases

Training of a classifier ('language annotator') in a supervised framework, requires a set of annotated entries with a distribution similar to the set of entries to be annotated. We know of only two such databases which can be freely accessed[6]; WALS and the library catalogue of MPI/EVA in Leipzig.

WALS: The bibliography for the *World Atlas of Language Structures* book can now be accessed online (`http://www.wals.info/`). This database contains 5633 entries annotated to 2053 different languages.

MPI/EVA: The library catalogue for the library of the Max Planck Institute for Evolution Anthropology (`http://biblio.eva.mpg.de/`) is queryable online. In May 2006 it contained 7266 entries annotated to 2246 different languages.

Neither database is free from errors, imprecisions and inconsistencies (impressionistically 5% of the entries contain such errors). Nevertheless, for training and development, we used both databases put together. The two databases put together, duplicates removed, contains 8584 entries annotated to 2799 different languages.

2.4 Test Data

In a large-scale on-going project, we are trying to collect all references to descriptive work for lesser-known languages. This is done by tediously

[5] Penalty weights set to 1 for deletion, insertion and substitution alike.

[6] For example, the very wide coverage database worldcat (`http://www.worldcat.org/`) does not index individual articles and has insufficient language annotation; sometimes no annotation or useless categories such as 'other' or 'Papuan'. The SIL Bibliography (`http://www.ethnologue.com/bibliography.asp`) is well-annotated but contains only work produced by the SIL. (SIL has, however, worked on very many languages, but not all publications of the de-centralized SIL organization are listed in the so-called SIL Bibliography.)

going through handbooks, overviews and bibliographical for all parts of the world alike. In this bibliography, the (meta-)language of descriptive data is be English, German, French, Spanish, Portuguese, Russian, Dutch, Italian, Chinese, Indonesian, Thai, Turkish, Persian, Arabic, Urdu, Nepali, Hindi, Georgian, Japanese, Swedish, Norwegian, Danish, Finnish and Bulgarian (in decreasing order of incidence)[7]. Currently it contains 11788 entries. It is this database that needs to be annotated as to target language. The overlap with the joint WALS-MPI/EVA database is 3984 entries.[8] Thus $11788 - 3984 = 7804$ entries remain to be annotated. From these 7 804 entries, 100 were randomly selected and humanly annotated to form a test set. This test set was not used in the development at all, and was kept totally fresh for the final tests.

3 Experiments

We conducted experiments with three different methods, plus the enhancement of spelling variation on top of each one.

Naive Lookup: Each word in the title is looked up as a possible language name in the world language database and the output is the union of all answers to the look-ups.

Term Weight Lookup: Each word is given a weight according to the number of unique-id:s it is associated with in the training data. Based on these weights, the words of the title are split into two groups; informative and non-informative words. The output is the union of the look-up:s of the informative words in the world language database.

Term Weight Lookup with Group Disambiguation: As above, except that names of genealogical (sub-)groups and country names that occur in the title are used for narrowing down the result.

Following a subsection on terminology and definitions, these will be presented in increasing order of sophistication.

3.1 Terminology and Definitions

- C: The set of 7 299 unique three-letter language id:s

- N: The set of 39 419 language name strings in the Ethnologue (as above)

- $C(c)$: The set of names $\subseteq N$ associated with the code $c \in C$ in the Ethnologue database (as above)

- $LN(w) = \{id | w \in C(id), id \in C\}$: The set of id:s $\subseteq C$ that have w as one of its names

- $C_S(c) = \cup_{winC(c)} Spellings(w)$: The set of variant spellings of the set of names $\subseteq N$ associated with the code $c \in C$ in the Ethnologye database. For reference, the $Spelling(w)$-function is defined in detail in Table 1.

- $LN_S(w) = \{id | w \in C_S(id), id \in C\}$: The set of id:s $\subseteq C$ that have w as a possible spelling of one of its names

- WE: The set of entries in the joint WALS-MPI/EVA database (as above). Each entry e has a title e_t and a set e_c of language id:s $\subseteq C$

- $Words(e_t)$: The set of words, everything lowercased and interpunctation removed, in the title e_t

- $LWEN(w) = \{id | e \in WE, w \in e_t, id \in e_c\}$: The set of codes associated with the entries whose titles contain the word w

- $TD(w) = LN(w) \cup LWEN(w)$: The set of codes tied to the word w either as a language name or as a word that occurs in a title of an code-tagged entry (in fact, an Ethnologue entry can be seen as a special kind of bibliographical entry, with a title consisting of alternative names annotated with exactly one category)

- $TD_S = LN_S(w) \cup LWEN(w)$: The set of codes tied to the word w either as a (variant spelling of a) language name or as a word that occurs in a title of an code-tagged entry

[7] Those entries which are natively written with a different alphabet always also have a transliteration or translation (or both) into ascii characters.

[8] This overlap at first appears surprisingly low. Part of the discrepancy is due to the fact that many references in the WALS database are in fact to secondary sources, which are not intended to be covered at all in the on-going project of listing. Another reason for the discrepancy is due to a de-prioritization of better-known languages as well as dictionaries (as opposed to grammars) in the on-going project. Eventually, all unique references will of course be merged.

- $WC(w) = |TD(w)|$: The number of different codes associated with the word w

- $WI(w) = |\{e_t|w \in Words(e_t), e_t \in WE\}|$: The number of different bibliographical entries for which the word w occurs in the title

- A: The set of entries in the test set (as above). Each entry e has a title e_t and a set e_c of language id:s $\subseteq C$

- $PA_A(X) = \frac{|\{e|X(e)==e_c, e \in A\}|}{|A|}$: The perfect accuracy of a classifier function X on test set A is the number of entries in A which are classified correctly (the sets of categories have to be fully equal)

- $SA_A(X) = \sum_{e \in A} \frac{|\{X(e) \cap e_c\}|}{|e_c \cup X(e)|}$: The sum accuracy of a classifier function X on a test set A is the sum of the (possibly imperfect) accuracy of the entries of A (individual entries match with score between 0 and 1)

3.2 Naive Union Lookup

As a baseline to beat, we define a naive lookup classifier. Given an entry e, we define naive union lookup (NUL) as:

$$NUL(e) = \cup_{w \in Words(e_t)} LN(w)$$

For example, consider the following entry e:

> Anne Gwenaïélle Fabre 2002 *Étude du Samba Leko, parler d'Allani (Cameroun du Nord, Famille Adamawa)*, PhD Thesis, Université de Paris III – Sorbonne Nouvelle

The steps in its NUL-classification is as follows are given in Table 2.

Finally, $NUL(e) = \{ndi, lse, smx, dux, lec, ccg\}$, but, simply enough, $e_c = \{ndi\}$.

The resulting accuracies are $PA_{NUL}(A) \approx 0.15$ and $SA_{NUL}(A) \approx 0.21$. NUL performs even worse with spelling variants enabled. Not surprisingly, NUL overclassifies a lot, i.e., it consistently guesses more languages than is the case. This is because guessing that a title word indicates a target language just because there is one language with such a name, is not a sound practice. In fact, common words like *du* [dux], *in* [irr], *the* [thx], *to* [toz], and *la* [wbm, lic, tdd] happen to be names of languages (!).

3.3 Term Weight Lookup

We learn from the Naive Union Lookup experiment that we cannot guess blindly which word(s) in the title indicate the target language. Something has to be done to individate the informativeness of each word. Domain knowledge tells us two relevant things. Firstly, a title of a publication in language description typically contains one or few words with very precise information on the target language(s), namely the name of the language(s), and in addition a number of words which recur throughout many titles, such as 'a', 'grammar', etc. Secondly, most of the language of the world are poorly described, there are only a few, if any, publications with original descriptive data. Inspired by the *tf-idf* measure in Information Retrieval (Baeza-Yates and Ribeiro-Neto, 1997), we claim that informativeness of a word w, given annotated training data, can be assessed as $WC(w)$, i.e., the number of distinct codes associated with w in the training data or Ethnologue database. The idea is that a uniquitous word like 'the' will be associated with many codes, while a fairly unique language name will be associated with only one or a few codes. For example, consider the following entry:

> W. M. Rule 1977 *A Comparative Study of the Foe, Huli and Pole Languages of Papua New Guinea*, University of Sydney, Australia [Oceania Linguistic Monographs 20]

Table 3 shows the title words and their associated number of codes associated (sorted in ascending order).

So far so good, we now have an informativeness value for each word, but at which point (above which value?) do the scores mean that word is a near-unique language name rather than a relatively ubiquitous non-informative word? Luckily, we are assuming that there are only those two kinds of words, and that at least one near-unique language will appear. This means that if we cluster the values into two clusters, the two categories are likely to emerge nicely. The simplest kind of clustering of scalar values into two clusters is to sort the values and put the border where the relative increase is the highest. Typically, in titles where there is exactly one near-unique language name, the border will almost always isolate that name. In the example above, where we actually have three near-

<table>
<tr><th>#</th><th>Substition Reg. Exp.</th><th>Replacement</th><th>Comment</th></tr>
<tr><td>1.</td><td>\′\‘\^\~\"</td><td>′′</td><td>diacritics truncated</td></tr>
<tr><td>2.</td><td>[qk](?=[ei])</td><td>qu</td><td>k-sound before soft vowel to qu</td></tr>
<tr><td>3.</td><td>k(?=[aou]|$)|q(?=[ao])</td><td>c</td><td>k-sound before hard vowel to c</td></tr>
<tr><td>4.</td><td>oo|ou|oe</td><td>u</td><td>oo, ou, oe to u</td></tr>
<tr><td>5.</td><td>[hgo]?u(?=[aouei]|$)</td><td>w</td><td>hu-sound before hard vowel to w</td></tr>
<tr><td>6.</td><td>((?:[^aouei]*[aouei]
[^aouei]*)+?)
(?:an$|ana$|ano$|o$)</td><td>\1a</td><td>an? to a</td></tr>
<tr><td>7.</td><td>eca$</td><td>ec</td><td>eca to ec</td></tr>
<tr><td>8.</td><td>tsch|tx|tj</td><td>ch</td><td>tsch, tx to ch</td></tr>
<tr><td>9.</td><td>dsch|dj</td><td>j</td><td>dsch, dj to j</td></tr>
<tr><td>10.</td><td>x(?=i)</td><td>sh</td><td>x before i to sh</td></tr>
<tr><td>11.</td><td>i(?=[aouei])</td><td>y</td><td>i before a vowel to y</td></tr>
<tr><td>12.</td><td>ern$|i?sche?$</td><td>′′</td><td>final sche, ern removed</td></tr>
<tr><td>13.</td><td>([a-z])\1</td><td>\1</td><td>remove doublets</td></tr>
<tr><td>14.</td><td>[bdgv]</td><td>b/p,d/t,g/k,v/f</td><td>devoice b, d, g, v</td></tr>
<tr><td>15.</td><td>[oe]</td><td>o/u,e/i</td><td>lower vowels</td></tr>
</table>

Table 1: Given a language name w, its normalized spelling variants are enumerate according to the following (ordered) list of substitution rules. The set of spelling variants $Spelling(w)$ should be understood as the strings $\{w/action_{1-i} | i \leq 15\}$, where $w/action_{1-i}$ is the string with substitutions 1 thru i carried out. This normalization scheme is based on extensive experience with language name searching by the present author.

$Words(e_t)$	$LN(Words(e_t))$	$Words(e_t)$	$LN(Words(e_t))$
etude	$\{\}$	cameroun	$\{\}$
du	$\{dux\}$	du	$\{dux\}$
samba	$\{ndi, ccg, smx\}$	nord	$\{\}$
leko	$\{ndi, lse, lec\}$	famille	$\{\}$
parler	$\{\}$	adamawa	$\{\}$
d'allani	$\{\}$		

Table 2: The calculation of NUL for an example entry

unique identifiers, this procedure correctly puts the border so that Foe, Pole and Huli are near-unique and the rest are non-informative.

Now, that we have a method to isolate the group of most informative words in a title e_t (denoted $SIG_{WC}(e_t)$), we can restrict lookup only to them. TWL is thus defined as follows:

$$TWL(e) = \cup_{w \in SIG_{WC}(e_t)} LN(w)$$

In the example above, $TWL(e_t)$ is $\{fli, kjy, foi, hui\}$ which is almost correct, containing only a spurious [fli] because Huli is also an alternative name for Fali in Cameroon, nowhere near Papua New Guinea. This is a complication that we will return to in the next section.

The resulting accuracies jump up to $PA_{TWL}(A) \approx 0.57$ and $SA_{TWL}(A) \approx 0.73$.

Given that we "know" which words in the title are the supposed near-unique language names, we can afford, i.e., not risk too much overgeneration, to allow for spelling variants. Define TWL_S ("with spelling variants") as:

$$TWL_S(e) = \cup_{w \in SIG_{WC}(e_t)} LN_S(w)$$

We get slight improvements in accuracy $PA_{TWL_S}(A) \approx 0.61$ and $SA_{TWL_S}(A) \approx 0.74$.

The $WC(w)$-counts make use of the annotated entries in the training data. An intriguing modification is to estimate $WC(w)$ without this annotation. It turns out that $WC(w)$ can be sharply estimated with $WI(w)$, i.e., the raw number of entries in the training set in which w occurs in the

foe	pole	huli	papua	guinea	comparative	new	study	languages	and	a	the	of
1	2	3	57	106	110	145	176	418	1001	1101	1169	1482
1.0	2.0	1.5	19.0	1.86	1.04	1.32	1.21	2.38	2.39	1.10	1.06	1.27

Table 3: The values of $WC(w)$ for w taken from an example entry (mid row). The bottom row shows the *relative increase* of the sequence of values in the mid-row, i.e., each value divided by the previous value (with the first set to 1.0).

title. This identity breaks down to the extent that a word w occurs in many entries, all of them pointing to one and the same language id. From domain knowledge, we know that this is unlikely if w is a near-unique language name, because most languages do not have many descriptive works about them. The TWL-classifier is now unsupervised in the sense that it does not have to have annotated training entries, but it still needs raw entries which have a realistic distribution. (The test set, or the set of entries to be annotated, can of course itself serve as such a set.)

Modeling Term Weight Lookup with WI in place of WC, call it TWI, yields slight accuracy drops $PA_{TWI}(A) \approx 0.55$ and $SA_{TWI}(A) \approx 0.70$, and with spelling variants $PA_{TWI_S}(A) \approx 0.59$ and $SA_{TWI_S}(A) \approx 0.71$. Since, we do in fact have access to annotated data, we will use the supervised classifier in the future, but it is important to know that the unsupervised variant is nearly as strong.

4 Term Weight Lookup with Group Disambiguation

Again, from our domain knowledge, we know that a large number of entries contain a "group name", i.e., the name of a country, region of genealogical (sub-)group in addition to a near-unique language name. Since group names will naturally tend to be associated with many codes, they will sorted into the non-informative camp with the TWL-method, and thus ignored. This is unfortunate, because such group names can serve to disambiguate inherent small ambivalences among near-unique language names, as in the case of Huli above. Group names are not like language names. They are much fewer, they are typically longer (often multi-word), and they exhibit less spelling variation.

Fortunately, the Ethnologue database also contains information on language classification and the country (or countries) where each language is spoken. Therefore, it was a simple task to build a database of group names with genealogical groups and sub-groups as well as countries. All group names are unique[9] as group names (but some group names of small genetic groups are the same as that of a prominent language in that group). In total, this database contained 3 202 groups. This database is relatively complete for English names of (sub-)families and countries, but should be enlarged with the corresponding names in other languages.

	PA	SA
NUL	0.15	0.21
TWL	0.57	0.73
TWL_S	0.61	0.74
TWI	0.55	0.70
TWI_S	0.59	0.71
TWG	0.59	0.74
TWG_S	0.64	0.77

Table 4: Summary of methods and corresponding accuracy scores.

We can add group-based disambiguation to TWL as follows. The non-significant words of a title is searched for matching group names. The set of languages denoted by a group name is denoted $L(g)$ with $L(g) = C$ if g is not a group name found in the database.

$$\begin{aligned} TWG(e) \quad = \quad & (\cup_{w \in SIG_{WC}(e_t)} LN(w)) \\ & \cap_{g \in (Words(e_t) \setminus SIG_{WC}(e_t))} L(g) \end{aligned}$$

We get slight improvements in accuracy $PA_{TWG}(A) \approx 0.59$ and $SA_{TWG}(A) \approx 0.74$. The corresponding accuracies with spelling variation enabled are $PA_{TWG}(A) \approx 0.64$ and $SA_{TWG}(A) \approx 0.77$.

5 Discussion

A summary of accuracy scores are given in Table 4.

All scores conform to expected intuitions and motivations. The key step beyond naive lookup

[9] In a few cases they were forced unique, e.g., when two families X, Y were listed as having subgroups called Eastern (or the like), the corresponding group names were forced to Eastern-X and Eastern-Y respectively.

is the usage of term weighting (and the fact the we were able to do this without a threshold or the like).

In the future, it appears fruitful to look more closely at automatic extraction of groups from annotated data. Initial experiments along this line were unsucessful, because data with evidence for groups is sparse. It also seems worthwhile to take multiword language names seriously (which is more implementational than conceptual work). Given that near-unique language names and group names can be reliably identified, it is easy to generate frames for typical titles of publications with language description data, in many languages. Such frames can be combed over large amounts of raw data to speed up the collection of further relevant references, in the typical manner of contemporary Information Extraction.

6 Related Work

As far as we are aware, the same problem or an isomorphic problem has not previously been discussed in the literature. It seems likely that isomorphic problems exist, perhaps in Information Extraction in the bioinformatics and/or medical domains, but so far we have not found such work.

The problem of language identification, i.e., identify the language of a (written) document given a set of candidate languages and training data for them, is a very different problem – requiring very different techniques (see Hammarström (2007a) for a survey and references).

We have made important use of ideas from Information Retrieval and Data Clustering.

7 Conclusion

We have presented (what is believed to be) the first algorithms for the specific problem of annotating language references with their target language(s). The methods used are tailored closely to the domain and our knowledge of it, but it is likely that there are isomorphic domains with the same problem(s). We have made a proper evaluation and the accuracy achieved is definetely useful.

8 Acknowledgements

We wish to thank the responsible entities for posting the Ethnologue, WALS, and the MPI/EVA library catalogue online. Without these resources, this study would have been impossible.

References

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. 1997. *Modern Information Retrieval*. Addison-Wesley.

Gordon, Jr., Raymond G., editor. 2005. *Ethnologue: Languages of the World*. SIL International, Dallas, 15 edition.

Hammarström, Harald. 2005. Review of the Ethnologue, 15th ed., Raymond G. Gordon, Jr. (ed.), SIL international, Dallas, 2005. *LINGUIST LIST*, 16(2637), September.

Hammarström, Harald. 2007a. A fine-grained model for language identification. In *Proceedings of iNEWS-07 Workshop at SIGIR 2007, 23-27 July 2007, Amsterdam*, pages 14–20. ACM.

Hammarström, Harald. 2007b. *Handbook of Descriptive Language Knowledge: A Full-Scale Reference Guide for Typologists*, volume 22 of *LINCOM Handbooks in Linguistics*. Lincom GmbH.

Hammarström, Harald. 2008. On the ethnologue and the number of languages in the world. Submitted Manuscript.