REPORT Full Automatic Archival Access Named Entity Retrieval on CABR

Martin Reynaert, Centre for Language and Speech Technology (CLST), Radboud University

26 September 2016

Management synopsis

Introduction

Creating a true Gold Standard (GS) is a costly and time-consuming business. Given the limitations of the available budget, time was in short supply. Aligning very noisy OCR text to a true Gold Standard is a very hard task. In the absence of a true Gold Standard, given the available time, an aligned OCR-GS version of these 100 archival items was far out of reach. An OCR-GS aligned version of the items is the only way to produce hard figures on the performance of the systems we have tested. For his very reason, the results we present here are tentative and only indicative of the systems' true performance. Nevertheless, we are confident that the measurements we have taken, tell a valid story that may perfectly serve as a solid basis for further managerial decision-making. We think the results show clearly that by using the tools tested, available free of charge as open source software , it is possible to gain true insight into the contents of the kilometers of archival boxes, given that their contents have been mass-digitized.

Test Corpus and its size

The corpus we have worked on consists of 100 pages drawn from just 2 of the CABR archival files. The archival boxes may hold up to 8,000 pages of the lightweight typewriting paper used predominantly at the period (roughly 1940-1960). Eight boxes take 1 meter of shelf space. The CABR archive takes about 4 kilometers of shelf space.

We see that this small sample amounts to about 35,000 word tokens, already roughly the size of a medium length novel.

The actual amount of text involved has had ramifications on our work. In the run-up to the current project, in the project proposal, we were shown a post card as an example. The card contains just two word tokens that one might hope that the OCR process might be able to properly recognize. The actual extent of the length of the texts encountered in the project, median length of each piece is x word tokens, came as bit of a surprise and also as a far larger work load than initially expected in terms of working hours required for ground truthing,

building the gold standard and Named Entity annotation and other preparatory work.

Gold Standard

In order to be able to measure the results of the tests, we needed a `perfect' version of the texts in the test corpus. This is generally referred to as a `Gold Standard' or GS.

The text for the GS was produced at Netwerk Oorlogsbronnen.

GS Annotators were Edwin Klijn (EKL) and Martin Reynaert (MRE). We evenly shared the work.

Tools used in this project

TICCL:
Text-Induced Corpus Clean-up or TICCL provides non-interactive spelling and OCR post correction facilities. TICCL aims at fully automatic postcorrection. TICCL is available in GitHub at https://github.com/martinreynaert/TICCL .

FROG:

One of the two major tools we used in this study is Frog. Originally developed at Tilburg University, development now continues mainly at Radboud University Nijmegen.

In the course of CLARIAH project PICCL (Philosophical integrator of Computational and Corpus Libraries, i.e. a web application/service workflow for automatically building text corpora from (possibly: images of) text incorporating Tesseract for OCRing, TICCL for automatic text correction, Frog for linguistically enriching the texts, BlackLab for indexing text ready for web interface WhiteLab) Frog is to be adapted for work on English and German, too, the latter being particularly relevant for possible future work on CABR. A new module geared at allowing researchers to more easily adapt Frog to their own domain, has also been developed within PICCL. This should e.g. allow to train Frog specifically to handle the CABR material. In fact it might already be trained specifically on the NER annotated Gold Standard we have developed in the framework of the current NIOD/NA project.

PICCL reference:
Reynaert, M., van Gompel, M., van der Sloot, K. & van den Bosch, A. (2015) PICCL: Philosophical Integrator of Computational and Corpus Libraries. 15 Oct 2015. Proceedings of
CLARIN Annual Conference 2015: Book of Abstracts. De Smedt, K. (ed.). Wrocław, Poland:
CLARIN ERIC, p. 75-79 5 p.

FOLIA

In the course of the current NIOD/NA project we have as a first step towards measuring Named Entity retrieval on the OCRed selection of CABR documents converted the documents to FoLiA XML format. FoLiA is also a product of CLARIN-NL projects at both Tilburg University and Radboud University Nijmegen. FoLiA: Format for Linguistic Annotation - FoLiA is a rich XML-based annotation format for the representation of language resources (including corpora) with linguistic annotations. A wide variety of linguistic annotations are supported, making FoLiA a useful format for NLP tasks and data interchange. URL: http://proycon.github.io/folia/ . A book chapter on FoLiA which focuses on the tools and further infrastructure which is currently available for the format is to appear in the fall of 2016 in a book presumably to be titled: "CLARIN in the Low Countries".

FoLiA Reference:
Maarten van Gompel, Ko van der Sloot, Martin Reynaert and Antal van Den Bosch. (2016)
FoLiA in practice: The infrastructure of a linguistic annotation format. CLARIN-NL in the Low Countries, Chapter 6. To appear.

Overview of NE retrieval performance

We have performed four series of tests on our Gold Standard. In order to achieve this we first converted the raw text files to FoLiA XML, as this is the pivot format for our corpus building work flow that contains all the tools we used in these experiments.

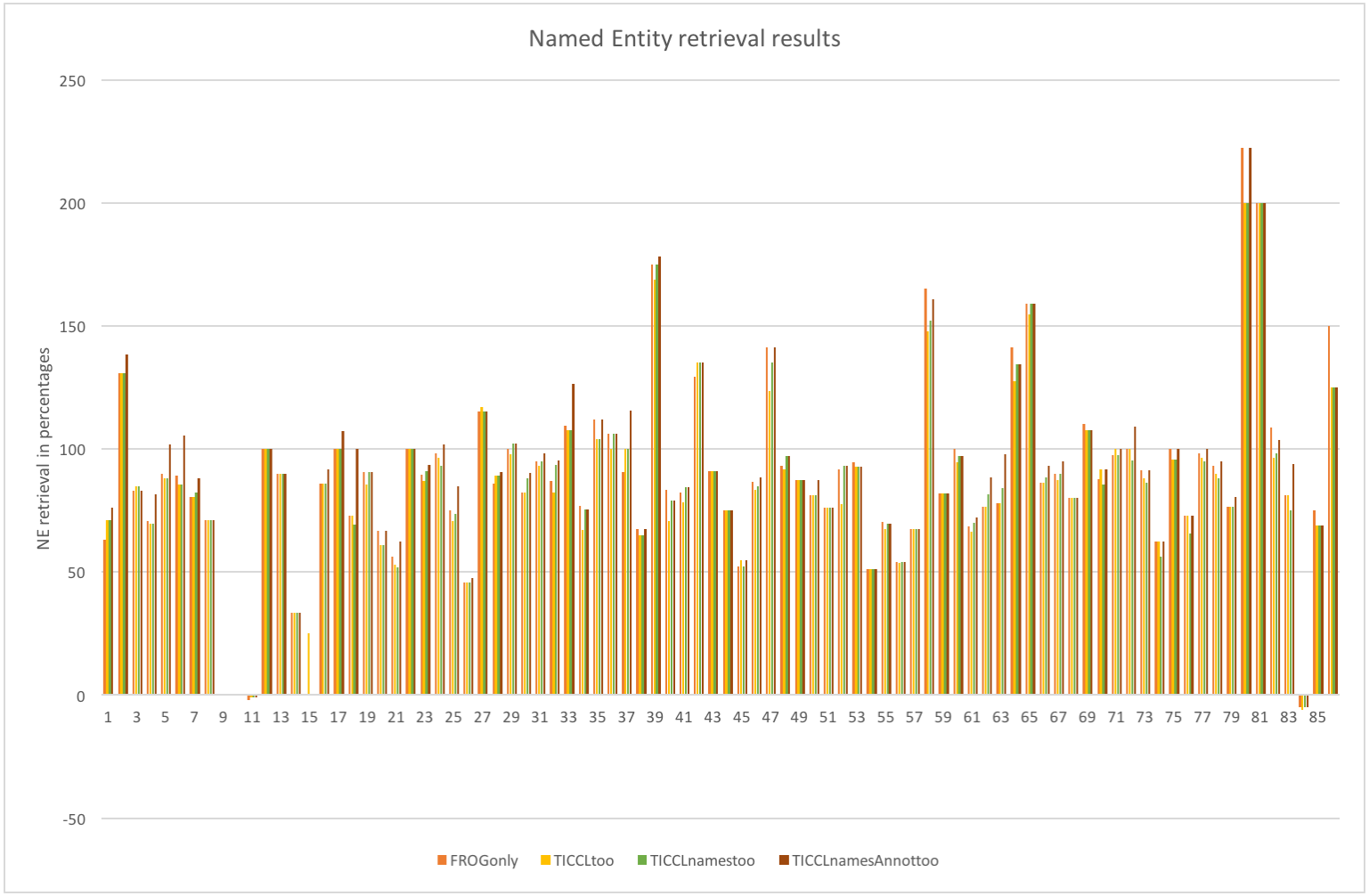
First, working name: FROGonly, we let Frog alone annotate the data for Named Entities.

Second, working name TICCLtoo, we first let TICCL automatically post-correct the OCR text using its standard diachronic lexicon which is in part based on the Historical Dutch lexicon built during the Impact project (aimed at improving OCR for historical texts) by then INL, now INT (Institute for the Dutch Language).

Third, working name TICCLnamestoo, we again let TICCL post-correct after augmenting its lexicon with all the first names and surnames from the CABR name list.

Fourth, working name TICCLnamesAnnottoo, we further augmented TICCL's lexicon with all the Named Entities manually annotated in the CABR files. This is not a real test, this serves as a reference. Its results in fact constitute an upper bound on what can possibly be achieved given 'perfect' resources such as name lists and our current tools.

The bar chart gives an overview of the results of these tests on 86 (?) annotated CABR files.


Named Entity retrieval results
NE retrieval in percentages
250
200
150
100
50
0
-50
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85
FROGonly
TICCLtoo
TICCLnamestoo
TICCLnamesAnnottoo

Results

The fact that FROGonly seems to perform better than the TICCLed versions is an artefact of the fact that Frog generalises over the data. It may thus very well identify a further unrecognizable character string on the basis of its context, i.e. the words preceding and following it, as an NE. TICCL replaces these strings by a single 'UNK' for 'unknown'. In our accounting, since this was originally an unrecognizable character string which Frog also assigns an NE label to (`Unk' is a place name somewhere), we disregard UNKs. The net effect of this is that while TICCLed versions may well score lower (even the Upper Bound ones) than FROGonly versions, the overall quality of the retrieved NEs is higher.

Analysis

There are a number of issues at play in the results that we need to clarify. In so far as they may skew the results, we shall try to explain the extent of this problem and thereby try to alleviate their effect.

We annotated dates and sometimes words that refer to some specific point in time. Frog currently does not deal with time references, of which we had 479 in our GS. Given sufficient and suitable training materials it should be possible to train Frog to deal with dates and time references. A specific tool for annotating dates and time references exists. It is called 'Heideltime' and specific rule sets for Dutch have even been developed by a colleague of ours, Matje van de Camp, first at Tilburg where she obtained her PhD and in the course of which she did this work, now at Radboud University Nijmegen. Heideltime is available in GitHub as open source software. The site https://github.com/HeidelTime/heideltime provides the system, but also the manually crafted Dutch rule sets.

On the other hand, Frog does annotate two classes of named entities we did not deal with: events (eve) and products (pro). In the data it annotated just 7 to 9 events, but between 173 and 184 products.

One further potential class of named entities neither we nor Frog annotated are monetary values. In so far as these reports often involve sums of money that were impounded and -- unexplained and unaccounted for -- went missing, we think this should be a class that is given due attention. Again, given sufficient and suitable training materials it should be possible to train Frog for monetary values.

Overall notes on performance

We observe there are no huge discrepancies between the results obtained by either of the TICCL enhanced versions and the established upper bound.

In the majority of the cases, results in terms of numbers of NEs automatically retrieved as compared to the numbers of manually annotated NEs as measured across the different classes are good to very good.

In so far as dates and other time references have been annotated in the Gold Standard and that Frog does not handle these, it is normal that on most files the systems do not reach 100%, i.e. retrieve the same number of NEs as were annotated in the Gold Standard. Nevertheless, some do. Others even surpass this number, even in the upper bound tests. We will further clarify this latter category in relation to particular files in the next Section.

Analyzing particular cases

In this section we look more closely at specific archival items.

We see that no NEs were retrieved for files
#10 in Excel or NL-HaNA_2.09.09_542_012.xml.txt and
#11 in Excel or NL-HaNA_2.09.09_542_013.xml.txt .

These contain hardly any text and indeed, not a single NE. So our systems scored perfectly, even though the bar chart might suggest they score 0% of the 100% they might have attained.

#12 in Excel or NL-HaNA_2.09.09_542_013.xml.txt
This one scores negatively (-200.00, -100.00, -100.00, -100.00). That is the result of a manipulation of ours to ensure that the result for files for which the annotators marked not a single NE visually obfuscate the other results. So in calculating the percentages we made the result negative and divided the actual numbers by 100. This we likewise did for
#85 in Excel or NL-HaNA_2.09.09_548_070.xml.folia.xml
The actual results were nevertheless not incorporated in the bar chart, they are high (-500.00, -600.00, -500.00, -500.00) and would otherwise dwarf the rest. These two are interesting. What we see happening in these is first that they are forms in list format. The second is a form to describe a person ('Signalement' in Dutch).

If we now turn our attention to the 16 or so files that obtain results above 100% (so for which Frog with/or without TICCL have marked more NEs than the humans annotators did), we see the highest outliers
81 in Excel or /opensonar/NIOD/NA/FR11SDKOutput_allSet/FROGonly/NL-HaNA_2.09.09_548_065.xml.folia.xml
82 in Excel or /opensonar/NIOD/NA/FR11SDKOutput_allSet/FROGonly/NL-HaNA_2.09.09_548_066.xml.folia.xml
What we mainly see happening is that while the human annotator saw only very few NEs in these files, these files are forms that were filled in (manually or typed). The field descriptions however were capitalized or denoted by a single capitalized letter: these were in turn erroneously picked up as names by Frog. What basically needs to be done is to train Frog not to report single or two character combinations as NEs.

Another factor which helps to explain the fact that Frog reports too many NEs is that Frog often reports multi-word expressions (MWEs) annotated as a single NE by the human annotators as two or more separate ones. Specific Frog-training on MWEs from this domain would definitely help to alleviate this problem.

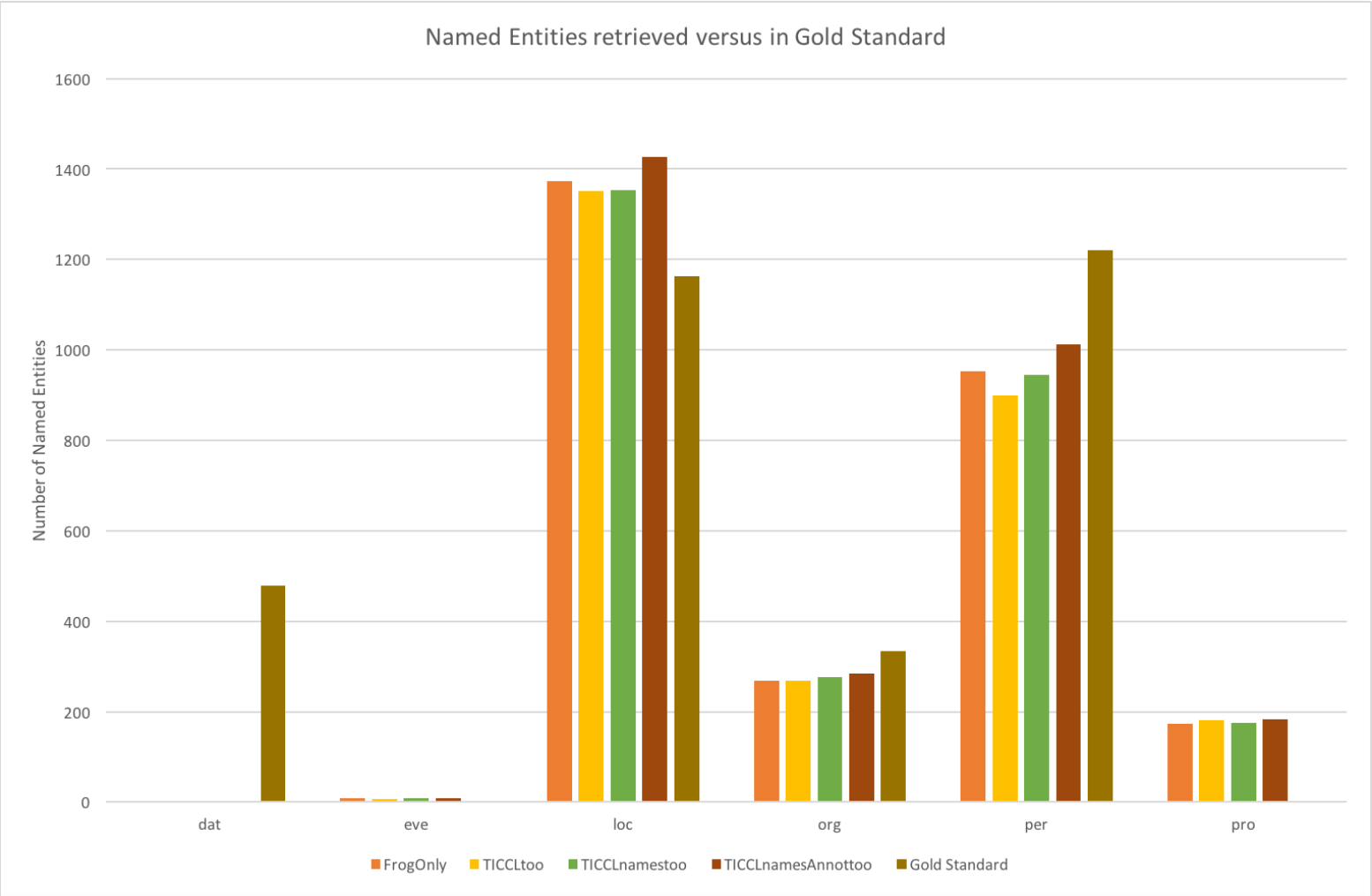


Figure: Numbers of Named Entities per class retrieved versus in Gold Standard

In the bar chart above we see, per class, the total numbers of NEs retrieved in each experiment contrasted to the numbers annotated in the Gold Standard (right-most column per class).

Dates:
As we explain further elsewhere, Frog is not currently trained to recognize dates and other time references and has consequently not retrieved any. They do form an important class of NEs, there are 479 of them in the Gold Standard so the account for over 13% of the NEs.

Events:
We actually did not annotate any events in de GS. Frog nevertheless reported a small number of them, from 7 to 9.

Locations:
Locations are by far the largest class of NEs. We see that all our experiments report too many, we explain elsewhere that in part this is due to Frog reporting several for multiword NEs where the human annotators saw only a single one.

Organisations:
These are often designated by acronyms, so show a lower number of MWEs. This in part accounts for the fact that Frog in this category does not report more than the human annotators did. Results on this class show clearly the contribution TICCL can make when it is given better resources such as names lists to help it.

Persons:
Persons form the second largest class. We see that we failed to retrieve about 1/6th of the annotated proper names. Please take into account that we here report results on word tokens, not types. Proper names often recur within the same document. If a name appears three times, but due to OCR misrecognition is retrieved only once, the system's results on word types would still be good: the retrieved metadata would include the name, as should be. That results on proper names are less good than on other classes of NEs can also in part be explained by the fact that names in these archival pages were often typed `spaced'. i.e. with a space between each distinct character, for the purpose of emphasis.

Products:
Please note that the bar for the GS annotations is not present: we did not annotate any product names in these pages. Frogs nevertheless reported between 173 and 184 product 'names'. We show the top most frequent 15 of these:

22 S.D.
21 1
14 Mei
14 2
12 Dossier
11 Zoo

11 Berends[1]
10 Berichten
9 4
8 3
6 5
5 N.S.B.
3 y
3 Strafrecht

Frog here too should be prevented from reporting single digits and characters. We see two acronyms for organisations and the name for a month. We will further explain how this could be remedied in future work. 'Zoo' here is not a zoological garden, but the older spelling for the word `zo', E. `so'. The four remaining words could be said to be products, but were never annotated as such in the GS.

A note on language recognition

A relatively small number of our test items are in fact in German (e.g. 548_028) and not Dutch. We should have applied language recognition to the documents handled. This we did not do.

The PICCL workflow we have used does contain a language recognizer. Once a document has been converted to FoLiA, the tool FoLiA-langcat can be invoked to annotate each paragraph with a language attribute. On the basis of this, TICCL or Frog can then be directed to not process those paragraphs marked as not being 'Dutch'.

Frog to date has not been trained for German text. This is, however, scheduled in the course of our CLARIAH project 'PICCL' to be done at Radboud University Nijmegen in early 2017. Frog will then also be trained to be able to process English text.

Further reflections on the NE annotation

* The most common denominators for places such as `house' and `wood' are not ever annotated as Named Entities. Nevertheless, a wood may be a place where specific, historically relevant events happened and the word casually mentioned in a number of archival pieces may well be fully defined in another, related document. Likewise, `houses' are places where perhaps other specific events may have happened. In our annotations, these were also never specifically marked, although often explicitly specified in conjunction with an address in e.g. testimonials or accounts of hearings. These mentions typically follow addresses, e.g. 'stationsstraat 100, huis te Amsterdam' (E: Station Road 100, house in Amsterdam). Perhaps in work like this, it should be recommended to specifically mark these possibly unspecified locations.

[1] Name was altered due to privacy reasons.

* Another type of possible Named Entity, i.e. stated values for amounts of money, were also not part of our annotations. Nevertheless, in the course of annotating these specific accounts, it became apparent that these have specific bearing on the events described and likely even on the motives and motivations of some of the people mentioned. Martin Reynaert specifically notes here that the value of 1,250, and multiples thereof, is a highly recurrent value in these archival pieces and that he thinks it is likely that the fact that part of the population at the time seems to have had at hand this particular amount in cash at all times may well be part of the larger story.

Reflections on TICCL

In silent understanding with our Director of Research, prof. dr. Antal van Den Bosch, I have managed over this past summer to gradually enlist the help of our senior programmer at CLST, Ko van der Sloot. Officially in the framework of our shared CLARIAH project PICCL, we have addressed a number of slated issues regarding further TICCL development. One of these is ngram correction, in casu simultaneous correction of combinations up to three words, geared at enabling TICCL to correct split words (space and/or punctuation inserted) and run-ons (missing space).

Even when given the help of the SoNaR-500 Language Model, TICCL on these data clearly struggles with data sparseness. The SoNaR corpus, being composed of Dutch written text produced mainly between 1990 and about 2010, clearly does not cover either the period nor the domain of the CABR archive. We firmly think that if TICCL were fed with the full contents of just one dossier in order to correct this dossier in this archive, it would outperform our current version with the SoNaR Language Model. The point is that the same names appear to recur sometimes elevated amounts of times providing better statistical evidence towards the correct spelling of the names present.

Workable alternative approach using TICCL

We see a possible alternative or complementary route we could take using TICCL. In the current study we have used TICCL with a large diachronical lexicon, later augmented with the CABR-names and then still with all the annotated NERs taken from our CABR Gold Standard.

TICCL performs in this study has performed isolated non-word error correction. We do not, we cannot yet, report on ngram correction. Each word is evaluated in isolation. When the word string is not in the lexicon, it is taken to be an Out-Of-Vocabulary (OOV) word and correction candidates are sought. This may and does result in incorrect names not being corrected. For the recurrent name 'van der Krijn' we have seen variants due to OCR misrecognition that TICCL missed, i.e. let go, because part of the misrecognized word happened to have resulted in another in-vocabulary word. The CABR corpus contains `van der erijn' and `van der urijn'. These were consequently not corrected.

This situation can be avoided if we perform ngramcorrection, e.g. correction on combinations of three words and give TICCL only a validated ngram name list instead of a `full' lexicon. If we do this, on the basis of the higher frequency name `van der Krijn' the lower frequency 'real word' errors `van der erijn' and `van der urijn' would effectively be corrected and retrieved. [2]

We think in follow-up work this complementary approach should also be evaluated.

Reflections on FROG and Named Entity Recognition

In these experiments we have only tried Frog to perform Named Entity Recognition on our data. Developing the NER module for Frog has in fact between a haste job performed by its programmer at TiCCl in between other jobs on the basis of the then recently available SoNaR1 gold standard for NEs developed in the STEVIN project SoNaR by Bart Desmet at LT3, University of Ghent. Desmet developed his own NER system called NERD on the basis of the same training data. No doubt, NERD is the better performing NE Recognizer, having been fully optimized and tuned on the Gold Standard SoNaR-1 data, whereas Frog was given only the very basic training data features. We, however, in the course of this small pilot project did not manage to get the NERD source code we still had around from the SoNaR project in late 2011 to compile on our current servers.

One more thing that has become abundantly clear to us in analyzing Frog's performance on these files, is that Frog has not been properly trained on recognizing fully capitalized names as names. Since it should be rather straightforward to do this, we will recommend to its developers that this be done. Fully capitalized names in these files, for extra emphasis and clarity, are in essence the norm.

What we also have not been able to do in the framework of this limited pilot project is to train the Named Entity Recognizer on this specific data. This would be highly recommended in a follow-up. Due to the specificity of the data, the NER would benefit highly from textual markers announcing a name such as e.g. 'de familie X' (E: the family X), 'de woning van Y' (E: the house of Y).

This leads us to discuss how Frog handles the context around / within words: (e.g. -straat, -gebouw, etc.]. Frog has been trained on the NEs manually annotated in the STEVIN project SoNaR-1 in 1 million words of running text. Basically the training materials consist of phrases of a particular number of words, some of which are NEs. Given words in a list, e.g. on a preprinted form, it lacks this context. It is possible to train Frog on compounds, i.e. have it take into account intra-word context, but this has not been done.

Frog is actually a combination of modules performing different tasks. The main module in this study has of course been the Named Entity Recognizer. Before this

[2] Name has been altered due to privacy reasons.

actually ran, it first performed tokenization. Tokenization is in essence splitting words tokens and punctuation marks so that words and word frequency counts are not 'polluted' by punctuation marks. That particular module is called Ucto. Besides tokenization Ucto also performs what we call sentence splitting: it tries to identify what are the sentences that make up the paragraphs of the text. A nice feature of Ucto is further that it labels the word strings it delivers according to a wide range of categories. The future relevance of this for the kind of work we have explored in this study will be immediately clear when I give a few examples of this:

```
      <w xml:id="NL-HaNA_2.09.09_548_049.xml.text.p.Page1_Block1.s.21.w.21"
class="ABBREVIATION">
        <t>S.D.</t>
--
<w xml:id="NL-HaNA_2.09.09_548_050.xml.text.p.Page1_Block5.s.14.w.6"
class="ABBREVIATION">
        <t>N.S.B.</t>
```

We see that already during tokenization the acronyms for two of the probably most mentioned organizations in the CABR are already marked as `abbreviations'.

```
      <w xml:id="NL-HaNA_2.09.09_548_002.xml.text.p.Page1_Block30.s.7.w.11"
class="DATE" space="no">
        <t>26-11-1892</t>
```

Whereas Frog currently does not return any dates, Ucto has a particular rule set to recognize the most common date formats. In our test sample it in fact reported 98 dates, on simple sight correctly. This already represents over 1/5th of the 'date' annotations in our GS.

Now the thing is that Frog's NER-module currently does not even use this valuable information given by it tokenization module. This definitely offers interesting possibilities for future work and further Frog development. Frog used to be called `Tadpole'. Largely in the framework of the CLARIN-NL project is has grown into Frog. We are working on its further development and hope that one day it will evolve into `Prince'.

The effect of `document fitness'

We now turn our attention to the fact that 11 out of 100 test items either held no usable, extractable information at all, or the OCR process failed to extract any sensible information at all.

Over 10% of one's collection delivering nothing of usable value is a sizeable amount. It is to be reckoned with in future work on these archives.

In this study this fact was dealt with manually: the 11 non-documents were set aside and exempted from further processing in the NER pipeline.

In future work this should become part of the automatic workflow.

It is for this reason that Martin Reynaert is developing and intends to incorporate into the PICCL workflow modules for `Automatic Lexical Quality Assessment Reporting', to be known by the acronym `ALQAR'.

Conclusion

There is no doubt to us that applying the available automatic techniques to archival items has more to offer than we thought at the onset. The fact that through the OCR process far more text is recovered from these pieces than at first was even thought possible is good news for fully automatically disclosing archives about which other than the number of archival boxes involved and the kilometers of shelf space required precious little is known about the actual contents. We have further shown that the fully-automatic OCR post-processing offered by TICCL further enhances the overall quality of the text and the amount of valuable metadata in terms of names recognized and categorized according to the classes defined. We have finally shown that providing TICCL with better resources such as domain specific name lists further enhances NE retrieval on these data and that overall results lie not too far off from the upper bound of what can be achieved. We have also looked at the tool we have in this work chosen for linguistically enriching the texts and for the NE extraction. We have pointed out directions in which these tools can be further developed and better set to use in future work.

Martin Reynaert
CLST – Radboud Universiteit Nijmegen
2016-09-21