Scientific Report of Short Term Scientific Mission

COST STSM Reference Number: COST-STSM-IS1305-34353

Period: 22-09-2016 to 28-09-2016

Duration: 5 working days.

COST Action: IS1305

STSM type: Regular (from Italy to Slovenia)

STSM Title: NLP4CMC & Lexicography

Guest/STSM applicant: Egon Stemle, EURAC, Bolzano/Bozen, Italy

Host: Simon Krek, Centre for Language Resources and Technologies, Ljubljana (SI),

simon.krek@guest.arnes.si

Purpose of the STSM

ENeL's WG3 concerns innovative e-dictionaries with a focus on the development of digitally born dictionaries. The training school 2016 in Ljubljana (SI), May 17-20, introduced participants, among others, to collecting, analysing, and automatically extracting data from web corpora.
Albeit related, the task of processing data from corpora of computer-mediated communication and social media interactions (henceforth referred to as CMC) has been deliberately excluded from the training school's programme. But we know that "new vocabulary is characteristic for CMC discourse, e.g. 'funzen' (an abbreviated variant of the German verb 'funktionieren', en.: 'to function') or 'gruscheln' (verb denoting a function of a German social network platform, most likely a blending of 'grüßen', en.: 'to greet' and 'kuscheln', en.: 'to cuddle')" [1] and therefore relevant to WG3; the goal of this STSM is to apply the methods and tools from the training school to CMC data.

Description of the work carried out during the STSM

Adapting and applying tools and methods for creating innovative e-dictionaries to CMC data

The goal of this STSM was to adapt and apply (part of) the work-flow from the ENeL Training School 2016[1] to data from CMC and social media corpora. In ENeL WG3, "innovative dictionaries are considered to be digitally-born, and thus no longer resemble traditional paper dictionaries but try to fully exploit the new possibilities of the digital medium using methods from computational linguistics." Consequently, participants were introduced to tools and methods that are used to create innovative e-dictionaries. The tools and methods were selected along a general work-flow that consists of three tasks: collecting and cleaning textual data, automatically extracting data, and editing and publishing extracted data in online dictionary writing systems or other publishing platforms. For this STSM, we adapted the work-flow in the following way: Collect data from CMC and social media, perform the second task, and manually check the resulting data as an editing step.

1 http://www.elexicography.eu/events/training-schools/ljubljana-2016/

To avoid technical pitfalls, we used a readily available implementation of this adapted workflow from STyrLogism[2], a project about South Tyrolean German neologisms[3] on news web sites of the region. There, the general idea is to use a list of vetted URLs and regularly crawl data from the corresponding web sites, clean and process the data and compare this latest data set to a reference corpus and the combination of all former data sets.

1. The actual implementation for data collection uses Heritrix[4], the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler. It is configured in such a way as to stay on the initial web sites for the main content but retrieve all relevant data from linked web sites; the configuration also ensures (this is a Heritrix default setting) that all relevant data is saved in the Web ARChive (WARC) archive format[5]. In DH, WARC is used as the standard for web archival.
2. Subsequently, STyrLogism uses the tools from *corpora from the web (COW)*[6] for data cleaning, a software toolkit for web corpus construction of large corpora (over 9 billion tokens) that processes WARC files, and performs basic cleanups as well as boilerplate removal, simple connected text detection as well as shingling to remove duplicates and near-duplicates from corpora [2].
3. The automatic data extraction consists of a straightforward comparison of lists of words where lists of named entities, terminological terms from the region, and specific terms of the regional dialect (altogether ~50k types) are joined with the COW's DECOW14 corpus word list (~60m types), and the data of former crawls; and then, the cleaned data of the current crawl is tokenised - but not lemmatised - and also converted into a word list. The list of candidate words consists of those words in the current crawl that appear less than a predefined number of times in all of the other data. Finally, the candidates are then manually checked.

For data collection in this particular experiment, we used the original STyrLogism list of about 120 region specific web sites pertaining to news-like content and manually selected 5 where user involvement was apparent, i.e. either content was (also) written by users, or users were very actively commenting or discussing, and we generally could expect large(-ish) amounts of user generated content. We configured Heritrix to crawl all 5 sites for a combined duration of 5h, after which around 1.4GB of data had been downloaded. However, this also included non-textual content, e.g. images, fonts, and style sheets, for archiving purposes. The amount of textual data, i.e. file types where text could potentially be extracted (html, plain text, pdf, openoffice / libreoffice / ms office text documents) was ~150MB[7] in ~3500[8] documents, and the amount of data in html and plain text files, the file types we actually extracted text from, was ~100MB in ~3300 documents.

For data cleaning, the COW toolkit was mostly left with its default configuration, in which it processes html and text documents between 2 and 256 kB of raw content[9], i.e. shorter and longer documents are discarded right away; it removes boilerplate, and performs "markup

2 https://commul.eurac.edu/styrlogism
3 We adopt the interpretation of StyrLogism for neologisms: only consider novel word forms, without semantic analysis, and hence, new meanings of existing words are not detected.
4 https://webarchive.jira.com/wiki/display/Heritrix
5 http://archive-access.sourceforge.net/warc/
6 http://corporafromtheweb.org/
7 This means, if archiving the content is undesired the amount of data that needs to be downloaded and saved can easily be reduced (in our case from 1.4GB to 150MB), and this would also reduce the amount of time needed to download this amount of textual content.
8 Note that this number includes possibly empty documents, and documents that are mere failure notices.
9 In [2] the authors actually suggest to limit the crawling process to only consider documents within these size limits.

and script removal, codepage conversion, etc., deals with faulty markup by favouring cleanliness of output over conservation of text and discards the whole document in case of serious unrecoverable errors. Also, [...] a simple paragraph detection is performed [...]. Perfectly identical subsequent lines are also removed, and excessively repeated punctuation is compressed to reasonable sequences of punctuation tokens." [2] Then, perfect duplicate and near duplicate documents were identified and removed, and we configured COW to detect German documents and discard all documents in other languages. After these cleaning steps the data consisted of 135 documents.
They were tokenised and the resulting word list was reduced with the help of the aforementioned lists of named entities, terminological terms from the region, and specific terms of the regional dialect joined with the COW's DECOW14 corpus word list. Additionally, two successive crawls of the full ~120 STyrLogism web sites[10] were also used to filter possible candidates. The filter criterion was very strict: Namely, a word only stayed on the list if it had not occured at all in the other data. This resulted in a list consisting of 466 words. After manual inspection, 113 were identified as being 'interesting' words, i.e. they were embedded in a sentence or in a context that enabled us to identify a meaning, and they were neither German words with unusual - or wrong - spelling or German words with missing space in between nor Italian words.

Examples of non-interesting words (353 words):

- Studientitel)anerkennung (mis-tokenised: left-over parenthesis)
- Achammersowie (missing space: Achammer sowie)
- Anerkennungzahlreicher (missing space: Anerkennung zahlreicher)
- aufseheheeregendes (misspelled word: aufsehenerregendes)
- beträchtlichhe (misspelled word: beträchliche)
- Europäer*innen (unusual spelling)
- Feigeblättern (wrong inflection/wrong spelling)
- decilino (Italian word)
- fighi (Italian word)

The 113 'interesting' words could be categorized into 4 ad-hoc categories.

Words specific to the region – or broader area (41 words), for example:

- Studientitelanerkennung
- Maturantenbroschüre
- Studientitelfrage
- Supplentenproblem
- Rentennachkauf
- Warenbegleitrechnung
- Wahlbozner

10 These crawls were done one and two months earlier, with a similar overall configuration of the crawler, and lasted for 3 days each.

Uncommon - but not really novel - words (44 words), for example:

- wassermetaphorik
- Universitätslernplattform
- Technokratengebildes
- Studienplatzvergabeportal
- Stauseenbau
- Sportweltmacht
- sicherheitsbedrohung
- Menschenrechtsentscheidungen
- Medienaufmacher
- Lehramtsantwärterinnen
- hydroelektirsche
- Grenzgeschehnissen
- Gehirnflucht
- Drogenlandwirte
- Vorfrankierten
- Genosenschaftsverband

Words used in advertisement - mostly tourism (15 words), for example:

- Wintersonnwendkräuter
- Waldlikör
- ungemostet
- Spitzenstorm
- Schweinsfiletschnitzel
- Lärchenfässer
- Käsegoldschmied
- Käseaffinateur
- Gravitationskelterung

At the end, a list of 13 words remained: these are all very uncommon or novel words with some twist:

Mörteldunst	(dust of plaster)
Dieselbuswelle	(hype of buses driving on diesel)
Weltneugieriger	(someone curios of the world)
Vordadaistische	(pre-dadaistic)
Verteidigungspsychose	(in this particular case: someone who is psychotic about defending their home country from foreigners)
Verschleierungsgesellschaft	(society with strong tendency to disguise and conceal)
Unverwechselbarkeitsfaktor	(distinctiveness marker)
Umweltfeigeblätter	(environmental fig leaf)
Rentnergenre	(the literary genre for retired persons)
Kulturmonolith	(an impenetrable culture)
Euregiobefürworter	(supporter of the Euregio region)
Dieselumkehr	(reversal of the support for diesel)
Leitkulturschalter	(switch for a defining culture)

Discussion

We see that there is something worth while looking for in the crawled data, albeit the ratio of valuable data and data that needs checking could be more favourable. This is particularly true if we consider that the news-like web sites we used for this experiment are still rather well-behaving – blogs, forums, social media sites, among others, usually include more deviations from the syntactic and orthographic norms of the written standard, e.g. colloquial spellings, speedwriting phenomena such as typos, the omission of upper case or the use of acronyms, or even intended, creative spellings (nice2CU, good n8) [3]. This will likely add even more data to sift through to the overall pile, and shift the data ratio even more towards the unfavourable. This emphasises the importance of tools for pre-processing CMC data and for manually sifting through data; optimising these tools for precision, usability and speed is of paramount importance the larger the pile of data grows. Still, the overall number of novel words – once the data has been cleaned – is remarkable considering that only five web sites were used to collect data.

Overall Summary

In the course of this STSM I was able to work more closely with Simon Krek (University of Ljubljana/Jožef Stefan Institute, Slovenia), Darja Fišer (University of Ljubljana, Slovenia), and Tomaž Erjavec (Jožef Stefan Institute, Slovenia). They all are or used to be actively involved in research covering - broadly speaking - methods and tools for the empirical analysis of corpora in the humanities, approaches towards automatic processing and annotation of linguistic data with computational methods, and corpus-linguistic research on collecting, processing, representing and providing corpora on the basis of standards in the field of the digital humanities. They also all have ties to the JANES Project[11] (Linguistic Analysis of Non-standard Slovene: Resources, Tools and Methods for the Research of Non-standard Internet Slovene), i.e. an ongoing project dealing with CMC data, and thereby up-to-date experience with collecting CMC data, and processing, syntactically analysing and normalising CMC data, to name just a few. More concretely, there have already been carried out a few hand full of analyses with the JANES Corpus data, and their experience with data formats and data conversion, with structuring data flows and analyses was very beneficial for this STSM. Our conversations were helpful, inspiring, and invaluable – for this STSM and beyond.

11 http://nl.ijs.si/janes/english/

References

[1] Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer; DeRiK: A German reference corpus of computer-mediated communication; Linguist Computing 2013 28: 531-537.

[2] Schäfer, Roland, and Felix Bildhauer. “Building Large Corpora from the Web Using a New Efficient Tool Chain.” In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12), edited by Nicoletta (Conference Chair) Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA), 2012.

[3] Beißwenger, Michael, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. “EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication, Social Media and Web Corpora.” In Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task, 78–90. Berlin, Germany: Association for Computational Linguistics, 2016.

Appendix

words specific to the region	uncommon words	words from (tourism) ads
Studientitelanerkennung	wassermetaphorik	Wintersonnwendkräuter
Maturantenbroschüre	Universitätslernplattform	Waldlikör
Studientitelfrage	Trubeliger	Wildkräuterblumen
hochschülerzeitung	Technokratengebildes	ungemostet
Supplentenproblem	Studienplatzvergabeportal	Spitzenstorm
Rentennachkauf	Studieninformationsstelle	Schweinsfiletschnitzel
Lehrerinnenrangliste	Stipendienformulars	Rindsgeselchtes
Lehrbefähigungskurse	stipendienbezugsberechtigt	Lärchenfässer
Hochschülervereinigung	Steuerabschreibemöglichkeit	Kräuterschlössl
Warenbegleitrechnung	Stauseenbau	Käsegoldschmied
Wahlbozner	Sportweltmacht	Käseaffinateur
Vizeschulamtsleiter	sicherheitsbedrohung	Gravitationskelterung
Uniheimplätze	Pfeilerbüsten	Grappakultur
Supplenzlehrperson	Paarbeziehungsmodelle	Geniesserseite
südtirolspezifische	Nachbuchstabierung	Wintercalville
Südtirolkoordinator	Menschenrechtsentscheidungen	
Südtirolbezüge	Medienaufmacher	novel words
Südtirolbezug	Lehramtsantwärterinnen	
Steuernummerkarte	Landhausbüros	Mörteldunst
Sonderlehrbefähigungskurse	Kurzzeitnutzers	Dieselbuswelle
Sommerbetreuerin	Klausurgesprächen	Weltneugieriger
Riechungshof	hydroelektirsche	Vordadaistische
Obstbaufelder	Heimplatzvergabe	Verteidigungspsychose
Mutterschaftsersatz	Grenzgeschehnissen	Verschleierungsgesellschaft
Meldeamtswesen	Gehirnflucht	Unverwechselbarkeitsfaktor
Maischeübernahme	gedichteschreibend	Umweltfeigeblätter
Leistungsstipendienwettbewerb	Friedensperformance	Rentnergenre
Landesstellenpläne	Fortschrittspaar	Kulturmonolith
Landesrechtsamt	Festspielarchivs	Euregiobefürworter
Landeskarriere	Extremprojekte	Dieselumkehr
kombinationspflichtiges	Existenzabenteuers	Leitkulturschalter
Kombinationspflichtig	Drogenlandwirte	
Kleinwohungen	Studiengebührenrückerstattung	
Innsbruckspezifische	Wohnheimregelung	
Hochschulfürsorge	Wahlfahrtkostenrückerstattung	
Hochschülerzeitung	Vorfrankierten	
Hochschülerzeitschrift	Mainstreamwellen	
Hochschülerorganisation	Literaturlorbeer	
Fränkschämen	Geographieprüfungen	
Faschistentagen	Genosenschaftsverband	
Englischlehrbefähigung	Gemeindepolitkerinnen	
	Seegebote	
	Wahlabschnitts	
	Wahlfahrtspesenrückerstattung	